Computational prediction of functional RNA structures: A study of the animal miRNAome and cis-acting replication elements in enteroviruses

Gerlach, Daniel

UNIVERSITÉ DE GENÈVE

Département de médecine
génétique et développement

FACULTÉ DE MÉDECINE
Professeur Evgeny M. Zdobnov

Département d'informatique

FACULTÉ DES SCIENCES
Professeur Ron D. Appel

Computational Prediction of Functional RNA Structures: A Study of the Animal miRNAome and *Cis*-Acting Replication Elements in Enteroviruses

THÈSE

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention bioinformatique

par

Daniel GERLACH

de

Nuremberg (Allemagne)

Thèse N° 4208

GENÈVE
Atelier d'impression ReproMail
2010

Daniel Gerlach: *Computational Prediction of Functional RNA Structures: A Study of the Animal miRNAome and Cis-Acting Replication Elements in Enteroviruses,* Diplom-Biologe Univ., © 2010

SUPERVISOR:
Prof. Evgeny M. Zdobnov

LOCATION:
Genève

TIME FRAME:
2006-2010

SHORT CV:

Personal Data
- Birthday: October 10, 1980
- Birthplace: Nuremberg
- Citizenship: German

Education

2006-2010
- Ph.D. Student / Research Assistant
- University of Geneva, Switzerland

2007-2010
- Swiss Institute of Bioinformatics Doctoral School
- Switzerland

2003-2006
- M.Sc. in Biology
- (German Diploma degree: "Diplom-Biologe Univ.")
- University of Wuerzburg, Germany

2001-2003
- B.Sc. in Biology (German Vordiplom)
- University of Erlangen, Germany

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES SCIENCES

Doctorat ès sciences
mention bioinformatique

Thèse de *Monsieur Daniel Johannes GERLACH*

intitulée :

" Computational Prediction of Functional RNA Structures : A Study of the Animal miRNAome and *Cis*-Acting Replication Elements in Enteroviruses "

La Faculté des sciences, sur le préavis de Messieurs E. ZDOBNOV, professeur associé et directeur de thèse (Faculté de médecine, Département de médecine génétique et développement), R. APPEL, professeur ordinaire et codirecteur de thèse (Département d'informatique), et M. ROBINSON-RECHAVI, professeur (Université de Lausanne, Faculté de biologie et de médecine, Département d'écologie et évolution, Lausanne, Suisse), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 4 mai 2010

Thèse - 4208 -

Le Doyen, Jean-Marc TRISCONE

N.B.- La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

PUBLICATIONS

Some of the work presented in this thesis has been published in the following articles:

First author articles:

Gerlach D, Kriventseva EV, Rahman N, Vejnar CE, and Zdobnov EM. miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.* (2009) 37:D111–117.

Cordey S*, Gerlach D*, Junier T, Zdobnov EM, Kaiser L, and Tapparel C. The cis-acting replication elements define human enterovirus and rhinovirus species. *RNA* (2008) 14:1568–1578.

Co-authored articles:

Body Louse Genome Working Group (MicroRNA analysis: Gerlach D, and Zdobnov EM). Genome Sequences of the Human Body Louse and its Primary Endosymbiont: Insights into the Permanent Parasitic Lifestyle. *Proc. Natl. Acad. Sci. U.S.A.* (2010) In Press.

Cordey S, Junier T, Gerlach D, Gobbini F, Farinelli L, Zdobnov EM, Winther B, Tapparel C, Kaiser L . Insights into rhinovirus genome evolution during experimental human and cells infections. *PLoS ONE* (2010) 5:e10588.

Nasonia Genome Working Group (MicroRNAs and tRNAs analysis: Anzola JM, Behura SK, Elsik CG, Gerlach D, Hagen DE, Munoz-Torres MC, and Zdobnov EM). Functional and Evolutionary Insights from the Genomes of Three Parasitoid Nasonia Species. *Science* (2010) 327:343–348.

Gatfield D, Le Martelot G, Vejnar CE, Gerlach D, Schaad O, Fleury-Olela F, Ruskeepää AL, Oresic M, Esau CC, Zdobnov EM, and Schibler U. Integration of microRNA miR-122 in hepatic circadian gene expression. *Genes Dev.* (2009) 23:1313–1326.

Bovine Genome Sequencing Consortium (MicroRNA analysis: Anzola JM, Gerlach D, and Zdobnov EM). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* (2009) 324:522–528.

Tapparel C, Junier T, Gerlach D, Van-Belle S, Turin L, Cordey S, Mühlemann K, Regamey N, Aubert JD, Soccal PM, Eigenmann P, Zdobnov E, and Kaiser L. New respiratory enterovirus and recombinant rhinoviruses among circulating picornaviruses. *Emerging Infect. Dis.* (2009) 15:719–726.

Tapparel C, Junier T, Gerlach D, Cordey S, Van Belle S, Perrin L, Zdobnov EM, and Kaiser L. New complete genome sequences of human rhinoviruses shed light on their phylogeny and genomic features. *BMC Genomics* (2007) 8:224.

* These authors contributed equally to the work

What physics was to the 20th century, biology will be to the 21st — and RNA will be a vital part of it

The Economist, Jun 14th 2007

REMERCIEMENTS

Je tiens à remercier mon directeur de thèse, le Professeur Evgeny M. Zdobnov et mon co-director le Professeur Ron D. Appel qui m'ont encouragé pendant ces derniers trois ans et demi. Je veux surtout remercier Professeur Evgeny M. Zdobnov de m'avoir accueilli et de m'avoir aidé à développer ma carrière scientifique. Grâce à son expérience, j'ai eu la possibilité de publier mes recherches dans plusieurs journaux scientifiques.

Je souhaite aussi exprimer ma reconnaissance au Professeur Laurent Kaiser, Central Virological Laboratory, Hôpitaux Universitaire de Genève, pour sa confiance. Notre collaboration fut vivante et productive. Merci aussi à Caroline Tapparel et Samuel Cordey.

Je tiens à remercier le Professeur Marc Robinson-Rechavi, Université de Lausanne, qui a accepté de juger mon travail et de faire partie de mon jury de thèse.

Je remercie également toutes les personnes dont la collaboration a été essentielle à l'accomplissement de ce travail. Ma reconnaissance va à Charles Vejnar, Stefan Wyder, Nazim Rahman, Evgenia Kriventseva et Robert Waterhouse.

Afin, je remercie aussi tous les membres du laboratoire CEGG du Professeur Evgeny M. Zdobnov pour leur collaboration et pour l'ambiance de travail qu'ils ont su créer.

Je veux remercier les personnes suivantes pour avoir relu ma thèse : Charles Vejnar, Robert Waterhouse, Philipp Berninger, Samuel Cordey, Caroline Tapparel, and Evgeny Zdobnov.

Enfin, le plus grand merci va à ma future femme Marion et à mes parents pour m'avoir soutenu et encouragé durant toute cette période.

ABSTRACT

An estimated 5–7% of the human genome is under selection and thus considered to be of functional importance. Protein-coding genes span only 1–2% of the human genome, leaving a considerable portion of functionally important, but often poorly characterized genomic regions. In addition to numerous, but very short regulatory sites, these regions contain conserved non-coding (CNC) sequences, and probably a large amount of non-protein coding RNA (ncRNA) genes. Many of these ncRNAs fold into stable secondary structures, which are indispensable for their function and therefore evolutionary conserved. Based on sequence and structure conservation, I have developed methods to study conserved and stable RNA secondary structures in animal and virus genomes.

I have implemented a computational approach for the genome-wide identification of microRNA (miRNA) genes. miRNAs are a class of short ncRNAs that regulate gene expression in a post-transcriptional manner and are processed from evolutionary conserved stem-loop sequences. I propose a general miRNA gene prediction method integrating *ab initio*, comparative genomics, and sequence homology based techniques. The pipeline has been used to explore the miRNAomes of over 40 animal genomes. All predictions are available through a web-accessible database named `miROrtho` (`http://cegg.unige.ch/mirortho`).

The `miROrtho` database complements the set of experimentally verified miRNA genes deposited in `miRBase` with consistent and comprehensive miRNA annotations across available genomes and predictions of novel putative miRNAs. Benchmarks with other *ab initio* miRNA gene predictors show a good and scalable performance for the `miROrtho` prediction pipeline. The pipeline was used within the framework of several genome projects including the taurine cattle, the parasitic *Nasonia* wasp, and the body louse. My annotations provided the core for the official miRNA gene sets for these respective genomes.

Applying my experience of conserved secondary RNA structures, I also discovered a novel stem-loop shaped *cis-acting replication element* (*cre*) which is involved in picornavirus replication. The novel *cre* element overlaps the open reading frame (ORF) of the recently described human rhinovirus A2 species (HRV-A2, synonym HRV-C). A common conserved structure for the *cre* is derived together with a loop specific sequence motif. Another *cre* -like element which I discovered in the human rhinovirus B species was computationally shown to be less thermodynamically stable and proved non-functional through forward genetics. Remarkably, the position of the *cre* within the genome is distinct for each of the three rhinovirus species (HRV-A, HRV-B, and HRV-A2), but conserved for the enterovirus species A–D (HEV-A, HEV-B, HEV-C, and HEV-D). Based on these findings and considering the high overall sequence similarity within the enterovirus species A–D, we proposed a reclassification of the four enterovirus species into a single species.

RÉSUMÉ

La part du génome du génome humain soumise à la sélection naturelle, et donc probablement d'importance fonctionnelle, est estimée à entre 5 et 7%. Les gènes de protéines ne couvrent que 1–2% du génome, ce qui laisse une part considérable de régions génomiques importantes au niveau fonctionnel, mais néanmoins mal connues. En plus des sites de régulation, ces régions contiennent des séquences conservées non-codantes (CNC), et probablement de nombreux ARN non-codants (ARNnc). Plusieurs de ces ARNnc adoptent des structures secondaires stables indispensables à leur fonction, qui sont de ce fait conservées pendant l'évolution. Sur la base de la conservation des séquences et structure, j'ai développé des méthodes pour l'étude des structures d'ARN conservées et stables dans des génomes animaux et viraux.

J'ai développé une approche computationnelle pour l'identification de micro-ARNs (miARN) à l'échelle du génome. Les miARN sont une classe de petits ARNnc qui régulent l'expression des gènes de façon post-traductionnelle et qui sont excisés des structures tige-boucle conservées. Je propose une méthode générale de prédiction des miARN intégrant des techniques *ab initio*, la génomique comparative, ainsi que des techniques fondées sur l'homologie. Cet ensemble d'outils a été utilisé pour explorer les miARNomes de plus de 40 génomes animaux. Toutes les prédictions sont disponibles via une base de données nommée miROrtho (`http://cegg.unige.ch/mirortho`).

La base de données `miROrtho` complète l'ensemble des gènes de miARN vérifiés expérimentalement déposés dans `miRBase` par des annotations exhaustives et des prédictions de nouveaux miARN. Des tests comparatifs avec d'autres prédicteurs de gènes de miARN montrent une bonne performance de la procédure de prédiction `miROrtho`, ainsi que sa capacité de mise à l'échelle. La procédure a été utilisée dans le cadre des projets de génome comprenant celui de la vache, d'une guêpe parasitoïde et du pou humain. Mes annotations cohérentes sont la base des listes officielles des miARN pour les génomes respectifs.

Grâce à l'expérience des structures secondaires d'ARN, j'ai aussi pu découvrir un nouvel élément de réplication, *cis-acting replication element* (*cre*), impliqué dans la réplication des picornavirus. Ce nouvel élément *cre* chevauche un cadre de lecture ouvert (ORF) du rhinovirus humain A2 (HRV-A2, synonyme HRV-C) récemment décrit. Une structure conservée commune pour l'élément *cre* est dérivée, en même temps qu'un motif de séquence spécifique à la boucle de la structure structure tige-boucle du *cre*. Un autre élément *cre* prédit chez le rhinovirus B (HRV-B), qui était moins stable au niveau thermodynamique, s'est révélé non-fonctionnel par des expériences de génétique inverse. Il est remarquable que la position de l'élément *cre* est différente chez HRV-A, HRV-B et HRV-A2 mais conservés chez tous les entérovirus (HEV-A, HEV-B, HEV-C et HEV-D). Sur la base de ces découvertes et le fait que les séquences des enterovirus montrent une bonne conservation, nous avons proposé la reclassification des quatre espèces d'entérovirus en une seule espèce.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACC	accuracy
AMFE	adjusted minimum free folding energy
AUC	area under the curve
BLS	branch length score
BRH	best reciprocal hit
cDNA	copy DNA
CEGG	Computational Evolutionary Genomics Group
CFE	centroid free energy
CNC	conserved non-coding
cre	*cis*-acting replication element
DGCR8	DiGeorge syndrome critical region gene 8
dsRNA	double-stranded RNA
EFE	ensemble free energy
EST	expressed sequence tags
FDR	false discovery rate
FMDV	food-and-mouth disease virus
FN	false negative
FP	false positive
HEV	human enterovirus
HMM	Hidden Markov Model
HRV	human rhinovirus
ICTV	International Committee on Taxonomy of Viruses
IRES	internal ribosomal entry side
lRNA	long RNA
MCC	Matthew's correlation coefficient
MFE	minimum free energy
MFEI	minimum free folding energy index
miRNA	microRNA
mRNA	messenger RNA
ncRNA	non-coding RNA
NMR	nuclear magnetic resonance

non-miRNA	non-microRNA
ORF	open reading frame
piRNA	Piwi-interacting RNA
pre-miRNA	precursor microRNA
pri-miRNA	primary microRNA
RBF	radial basis function
RISC	RNA-induced silencing complex
ROC	receiver operating characteristic
rRNA	ribosomal RNA
SCFG	stochastic context free grammar
SCI	structural conservation index
SE	sensitivity
siRNA	small interfering RNA
SMOTE	synthetic minority over-sampling technique
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
SP	specificity
sRNA	small RNA
SRP	signal recognition particle
stRNA	small temporal RNA
SV	support vector
SVM	Support Vector Machine
TN	true negative
TP	true positive
TRBP	TAR RNA-binding protein
tRNA	transfer RNA
UTR	untranslated region
VPg	Viral Protein of the genome
vRNA	vault RNA

Part I

GENERAL INTRODUCTION

1 INTRODUCTION

Apart from serving as the messenger in protein synthesis, RNA has many other important biological roles, including catalyzing RNA splicing as ribozymes (Doudna and Cech, 2002), localizing proteins (Walter and Blobel, 1982), or the guidance of chemical modifications of other RNA (Cavaillé et al., 1996). Based on these multiple functions of RNAs and their ability to store, transmit, and duplicate genetic information, RNAs were postulated to be the origin of all life on earth. This became known as the "RNA world hypothesis" (Gilbert, 1986) which claims that ancient lifeforms, based on RNA rather than DNA, predate the current living organisms.

Within the last years, more and more roles are attributed to RNA. Even more, new classes of novel non-protein-coding RNAs (ncRNA[1]), such as Piwi-interacting RNAs (piRNAs) (reviewed in Klattenhoff and Theurkauf, 2008; Thomson and Lin, 2009) silencing transposons, endogenous small interfering RNAs (siRNAs) (Chung et al., 2008; Czech et al., 2008; Ghildiyal et al., 2008; Okamura et al., 2008), or microRNAs (miRNAs) (reviewed in Kim et al., 2009) involved in post-transcriptional regulation have been discovered. They are all part of the "modern RNA world" (Eddy, 2001), revealing new functions and novel classes of ncRNAs.

1.1 RNA SECONDARY STRUCTURE

Many ncRNAs have regions which adopt a conserved secondary structure as it is important for their function. A primary RNA sequence contains four types of bases or nucleotides[2]: Adenine (A), Cytosine (C), Guanine (G), and Uracil (U). An RNA sequence can adopt a lower energy, and thus more stable confirmation, by base pairing of complementary canonical Watson-Crick base pairs. G forms three hydrogen bonds with C, and A binds to U with two hydrogen bonds. Furthermore, at rarer frequencies, RNA molecules can also form non-canonical base pairs such as the G·U wobble pair, the sheared G·A pair, or the reverse Hoogsteen[3] pair. To date, there are about 29 possible pairs of alternative interactions known, including the standard pairs (Burkard et al., 1999). In addition to the base pairs formed by intermolecular hydrogen bonds, stacking contributes to the formation of an RNA sequences adapting a low free energy which is released during the folding process. Stacking of adjacent nucleotides along the RNA sequence is enabled by interactions between the aromatic rings of the purine (G, A) and pyrimidine (C, U) bases. In an RNA sequences the aromatic rings are almost parallel in position, allowing the overlap of π-bonds (part of double bonds) between adjacent bases. This kind of interaction is a non-covalent chemical bond. RNA stem-loop structures (a lollipop-

1 Basically all RNAs other than mRNAs are called ncRNAs.
2 A nucleotide is composed of a base, a ribose, and a phosphate. For this work we use the terms "base" and "nucleotide" interchangeably
3 After Karst Hoogsteen, born 1923, Dutch-American biochemist, who first described it.

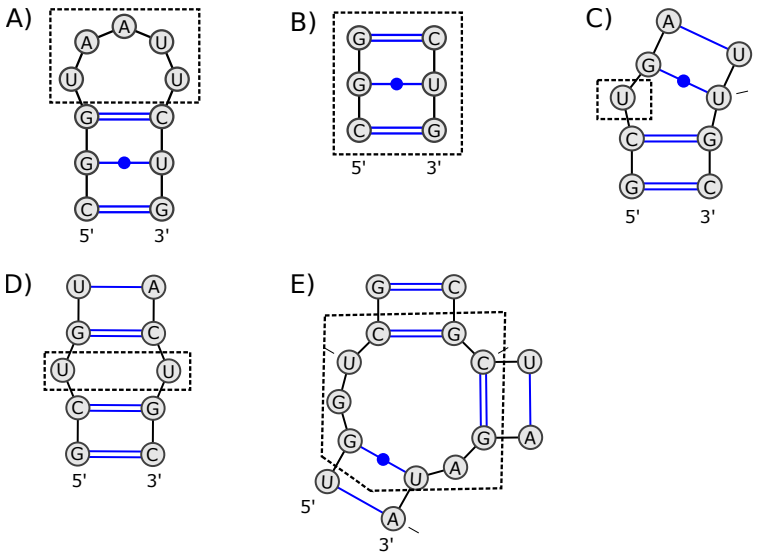


Figure 1: Common secondary structure elements found in folded RNA sequences. The bases in the structure are indicated by their initials, the solid black line shows the sugar-phosphate backbone to which the bases are attached, the blue lines and circles show the pairings. G·C are connected by two lines, A·U by one, and G·U pairs are depicted by one line and a dot. The various elements are marked by dashed boxes. A) Hairpin (terminal) loop B) Stem region C) Bulge loop D) Interior loop E) Multi loop. Visualization: `VARNA` (Darty et al., 2009).

shaped structure with one stem region and one terminal loop) e. g., are stabilized by stacking in the terminal loop region.

1.2 RNA SECONDARY STRUCTURE PREDICTION

Based on a thermodynamic model (Mathews et al., 1999) and a dynamic programming algorithm (Nussinov and Jacobson, 1980), a set of ordered base pairs (i, j) can be computed that minimizes the free energy of an RNA sequence (Zuker and Stiegler, 1981). The energy adopted by a folded RNA is called the minimum free energy (MFE), and the corresponding structure is called the MFE structure. Although the MFE structure is not necessarily the functional fold of an RNA, it provides a single best guess for the real structure. The thermodynamic model for the computational prediction of an RNA structure can be derived from experimental data on small artificial RNAs. Paired complementary bases and stacking of adjacent nucleotides, lower the free energy of an RNA structure, whereas unpaired regions such as loops and bulges (Fig. 1) increase the free energy. The use of a fast and efficient dynamic programming algorithm prohibits the formation of pseudo-knots[4] in the structure. Nested secondary structure elements are known for some experimentally verified structures, but are commonly excluded due to their high computational complexity and the lack of an appropriate thermodynamic model.

Fig. 1 shows some common structural elements that can be part of larger structures. A double-stranded region in an RNA structure, formed by complementary base pairs, is known as a *stem*. Bases that

4 All software used in this work only computes pseudo-knot free structures.

are unpaired and enclosed by at least one closing base pair form a *hairpin loop*. Regions with unpaired nucleotides connecting three or more stem regions are named *multi-loop regions*. *Interior loops* are characterized by a loop with two closing base pairs at each side between unpaired bases. A structural element with unpaired nucleotides on one side and no unpaired nucleotides on the other side of a stem region is known as a *bulge loop*. Some combinations of common elements have specific names, such as a lollipop-shaped stem-loop structure consisting of a stem region with bulge- and interior-loops and one terminal hairpin loop. Free unpaired nucleotides at the ends of folded RNA structures, called dangling ends, have a negative contribution to the stability of an RNA structure. In general, the lower the free energy of an folded RNA sequences, the more likely is the structure to be stable.

A set of base pairs for a pseudo-knot free secondary structure can be uniquely represented by the "dot-parenthesis notation". Unpaired bases are encoded as a simple "`.`", an $i \cdot j$ base pair is shown as a pair of opening and closing parentheses "`()`". Thus, a simple stem-loop structure such as "`GGGAAACCC`" can be written as "`(((...)))`".

Although dynamic programming and a thermodynamically model allow the prediction of a single structure with the lowest free energy for any primary RNA sequence, RNAs *in vivo* are rather represented by populations of structures forming a dynamic equilibrium. Those structures, which are called sub-optimal referred to the MFE structure, can be computed with an algorithm proposed by Zuker (1989). With an extension introduced by Wuchty et al. (1999), all suboptimal secondary structures within a given energy range of the MFE structure can be calculated. McCaskill (1990) introduced a partition function for secondary structure formation to calculate all sub-optimal structures, and their contributions to the ensemble of structures, weighted by their Boltzmann probabilities. The algorithm also computes the base pair probability for every possible base pair that can be formed by the RNA.

Ab initio methods, that rely solely on a thermodynamical model, were shown to have an accuracy of 73% for known canonical base pairs in sequences with fewer than 700 nt (Mathews, 2004). Lower accuracies have been reported for different sets of sequences, mostly for longer sequences (Doshi et al., 2004; Dowell and Eddy, 2004). The accuracy of RNA secondary structure prediction can be improved by computing sub-optimal structures. The single best of 750 sub-optimal structures contains, on average, 87% of known base pairs (Mathews, 2004). Current programs for predicting the secondary structure of proteins score worse, with an average performance of about 75% (Montgomerie et al., 2006).

The accuracy of RNA secondary structure prediction by free energy minimization is limited by several factors. First, the thermodynamic model is incomplete, such that some non-canonical pairs can show non-nearest-neighbor effects which are not included in the standard model. Furthermore, single-stranded nucleotides in multi-branch loops (Fig. 1 on the facing page) are known to influence the stability of RNA structure, but they are not adequately integrated in current dynamic programming algorithms. Second, RNA secondary structures are to some extend determined by folding kinetics such that the MFE structure is not necessarily the most stable one. Third, as already mentioned, most programs exclude pseudo-knots as current heuristics for

predicting these elements do not guarantee an optimal structure. Finally, RNA sequences may fold into more than one structure as exemplified by riboswitches which change their confirmation depending on the environment (Tucker and Breaker, 2005).

Several attempts have been made to increase the performance of secondary structure prediction of RNAs. Mathews (2004) predicted on average 84% of known canonical base pairs on a set of eleven sequences (Mathews et al., 2004) by incorporating chemical modification constraints into the dynamic programming algorithm. By far the most accurate RNA secondary structure predictions, however are based on comparative sequence analysis (Pace et al., 1999). As opposed to *ab initio* methods, they infer base pairs by determining pairs that are common among multiple homologous sequences. Homologous sequences are sequences that are similar in and have arisen from a common ancestor. Those sequences are then aligned such that homologous positions of the individual sequences match. Specific pairs are then proven by the existence of covariations in the respective alignment columns. A simple covariation is given by a G·C pair present in one sequence which is substituted by an A·U pair in another homologous sequence. Both base pairs conserve the base pair at a specific position in the secondary structure. Given enough homologous sequences that are divergent enough to contain covariations, comparative analysis can yield accuracies up to 97% as shown for ribosomal RNAs (rRNAs) and subsequent experiments of crystal structures (Gutell et al., 2002). The high performance of comparative analysis, regardless the need for multiple sequences, compared to *ab initio* methods, makes them the current gold standard for the determination of RNA secondary structure.

Some common software packages used to compute an RNA secondary structure based on a multiple alignment are `COVE` (Eddy and Durbin, 1994), `Construct` (Lück et al., 1999), and `RNAalifold` (Hofacker et al., 2002).

The comparative approach to predict a common secondary structure for a set of homologous sequences can be addressed in several ways. The first rigorous mathematical solution for inferring a common structure to two or more sequences was proposed by Sankoff (1985). Based on thermodynamics and a set of homologous sequences, the algorithms simultaneously computes an alignment and an optimal consensus structure. However, its time complexity is $O(n^6)$, with space requirements of $O(n^4)$ for only two sequences of length n, thus becoming unusable for more sequences. Restricted versions of Sankoff's algorithm have been proposed which limit the search space of possible alignments and structures (Havgaard et al., 2007; Kiryu et al., 2007).

Another approach to predict a common secondary structure is the "fold-then-align" method. First, the structures are predicted using the individual sequences and dynamic programming coupled with a thermodynamic model. Second, the structures are aligned using tree-based metrics (Shapiro and Zhang, 1990). The weakness of this approach is clearly the inaccuracy for single sequence secondary structure prediction.

The most common approach, however, to predict a consensus structure for homologous sequences, is the "align-then-fold" method. This approach yields high performance (Pace et al., 1999) and is feasible in "reasonable" computational time. Its basic principle is to first align a set of sequences and then to detect covarying paired sites in the align-

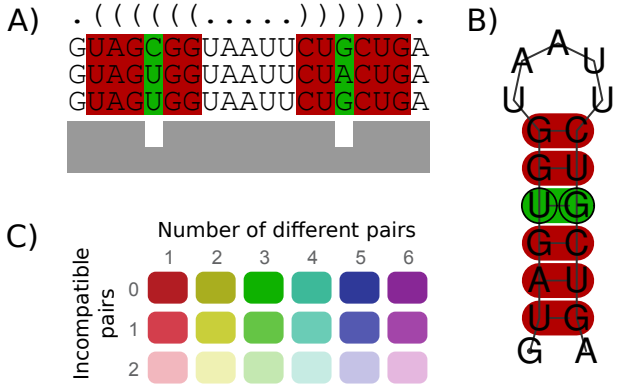


Figure 2: Coloring scheme for highlighting the mutational pattern with respect to the conserved secondary structure. Base pairs that are formed by several different combinations of nucleotides are called consistent or compensatory mutations and are indicated by different colors. Pale colors indicate a base pair that cannot be formed in some sequences of the alignment. A) Alignment with color-coded RNA structure conservation, dot-parenthesis RNA notation on top and primary sequence conservation at bottom. The green columns show three alternative pairs which are consistent with the consensus structure. B) Consensus RNA secondary structure C) Color-coding used for the different number of consistent and compensatory mutations.

ment. Covariations are called *compensatory mutations* with regard to a common secondary structure, if both sites are mutated while preserving the base pair. Fig. 2 shows a consensus structure which is preserved in a set of three homologous sequences. While the first sequence contains a C·G base pair, the homologous position in the second sequences contains a U·A pair. Furthermore, as a G can either pair with an C or an U, another type of mutations preserving the consensus structure can be found in alignments of structural RNA sequences. A C·G pair that is substituted by an U·G pair is called a *consistent mutation* (Fig. 2).

Based on functional constraints of conserved secondary structures, homologous sequences executing those functions tend to preserve a common RNA structure. This creates a specific pattern of covarying sites which can be employed for structure prediction. Nevertheless, the performance of such methods depends heavily on the quality of the input alignment. Sequences have to be divergent enough to contain a substantial number of consistent and compensatory base changes (Fig. 3 on the next page) which can be used for structure prediction. Alignment algorithms which are solely based on primary sequence data information such as `ClustalW` (Thompson et al., 1994), however, have been shown to misalign sequence with less than 65% sequence identity, and thus destroying the underlying secondary structure information (Gardner et al., 2005). Moreover, structural RNAs evolve rapidly on the primary sequence level. The 7SK RNA from the lamprey *Lampetra fluviatilis* e. g., differs in more than 30% of its nucleotide positions from its mammalian counterpart (Gürsoy et al., 2000) and the whole family shows an average sequence identity of 50% (`Rfam`[5] family RF00100). Due to the low sequence conservation, it was only recently that a *Drosophila* 7SK candidate had been proposed (Gruber et al., 2008).

5 A database of ncRNA gene families.

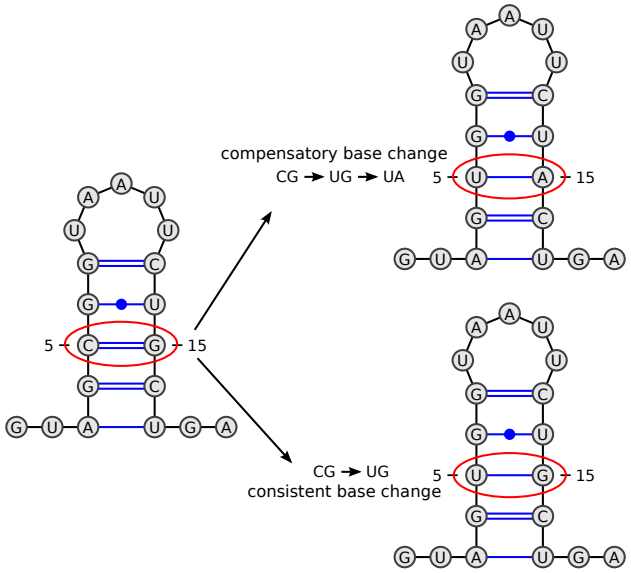


Figure 3: Consistent and compensatory base changes preserving the RNA secondary structure. Changes in the sequence can preserve a common secondary structure. In addition to the usual Watson-Crick base pairs, non-canonical base pairs such as the G·U pair (also called *wobble pair*) can are also found in RNA molecules (Nagaswamy et al., 2000). They allow single base substitutions, leading to consistent mutations preserving the base pair pattern (lower right). A change of both nucleotides which preserves the structure is called a compensatory base change (upper right).

Even though higher conserved sequences are more reliably alignable, they might not contain enough covariations for a good comparative analysis. A solution can be to use sequence-structure based alignment methods such as `R-Coffee` (Wilm et al., 2008), however a big downside to those algorithms is that they rely on single sequence structure prediction in the first place. So for rather distant structural RNAs, the "align-then-fold" method can be viewed as a "chicken or egg" problem. To predict a common secondary structure, an alignment of sufficiently divergent sequences is needed. However, to get an accurate starting alignment of divergent sequences, secondary structure information is needed. Nevertheless, for sequences identity ranging from 70% to 85%, algorithms like `RNAalifold` (Hofacker et al., 2002) can predict a common conserved secondary structure more accurate than using only single sequence information coupled with a thermodynamic model.

The coloring schema used by `RNAalifold` shows the mutational pattern in an alignment of homologous sequences with respect to a conserved RNA structure (Fig. 2 on the preceding page). Consistent and compensatory base changes on the consensus structure are indicated by different color codes.

1.3 NON-CODING RNA GENE FINDER

ncRNAs are RNAs that function directly as RNA molecules without being translated into proteins. Their identification based on a genome sequence is a challenging task. ncRNAs lack strong statistical sequence properties such as promotor regions, start and stop codons, or a codon

bias (Meyer, 2007). However, three key methods have been developed for the discovery of ncRNA genes.

1.3.1 *Homology-based non-coding RNA gene finder*

A straight-forward approach for the discovery of new members of known ncRNA gene families are homology-based approaches. However, as many ncRNAs evolve rapidly on the sequence level, a search for new members using a software like `BLAST` (Altschul et al., 1990) which looks for stretches of conserved DNA, might not always be sensitive enough. As many ncRNAs contain a conserved secondary structure, probabilistic models such as covariance models have been developed, describing the conserved secondary structure of a ncRNA family. A probabilistic covariance model can be trained on a known family and then aligned to a given sequence to find homologs of that family. Covariance models for ncRNA gene discovery were introduced by Eddy and Durbin (1994) and implemented in the software `Infernal` (INFERence of RNA Alignment) (Eddy, 2002) which had been used to create the `Rfam` database (Gardner et al., 2009) of ncRNA families.

1.3.2 *Ab initio-based non-coding RNA gene finder*

Another approach for ncRNA gene prediction are *ab-initio* methods, which rely only on statistical sequence properties to predict genes in genomic sequences. The first attempt for creating a general ncRNA gene-predictor failed, as ncRNAs do not have a significantly lower free energy than random sequences (Rivas and Eddy, 2000; Workman and Krogh, 1999). A specific class of ncRNAs, namely miRNAs, however have been shown to have lower folding energies than random sequences (Bonnet et al., 2004). The success of general *ab initio* ncRNA gene finders has been limited so far, with an exception of identifying ncRNA genes presenting an GC-skew and a high degree of secondary structure in AT-rich hyperthermophile organisms such as *Methanococcus jannaschii* and *Pyrococcus furiosus* (Klein et al., 2002).

1.3.3 *Comparative genomic non-coding RNA gene finder*

By far the most frequently used approach for ncRNA gene detection are methods based on comparative genomic. These methods rely on set of aligned homologous sequences and a pattern of covarying positions in a common conserved secondary structure (Fig. 2 on page 7). The `QRNA` program (Rivas and Eddy, 2001) e. g., models the distribution of covarying positions (a G·C pair mutating to an A·U pair) in an alignment via a stochastic context free grammar (SCFG). The protein-coding potential of a sequences is coded in a Hidden Markov Model (HMM). The final output of `QRNA` consists of a prediction if a region in a non-protein coding region, a protein-coding region or none of both. The program `EvoFold` (Pedersen et al., 2006) extends `QRNA` by using multiple sequences and integrating them in a phylogenetic footprinting approach. The core algorithm is built around an SCFG for finding functional ncRNAs based on an eight-way genome-wide alignment of the human, chimpanzee, mouse, rat, dog, chicken, zebrafish, and pufferfish genomes. Another recently developed approach called `RNAz`

(Gruber et al., 2010; Washietl et al., 2005) incorporates structural conservation with thermodynamic stability. The algorithm is trained on alignments of known ncRNAs. It first computes a normalized measurement of thermodynamic stability by comparing the MFE of a sequence to the ones of random sequences with conserved length and base composition. A z-score is computed as $z = (E - \mu)/\sigma$, where μ and σ are the mean and standard deviations of the MFEs of random sequences. Negative z-scores indicate that the structure is more stable than expected by chance. `RNAz` assigns the structural conservation by computing a structural conservation index (SCI). The consensus structure for an alignment is computed using `RNAalifold` which combines the thermodynamic model with covariance information. Compensatory and consistent mutations with regards to the common secondary structure (Fig. 2 on page 7) give a "bonus" to the free energy. Inconsistent mutations that do not agree with a conserved secondary structure yield a penalty term. The consensus MFE E_A is then compared to the average MFE of the individual sequences as $SCI = E_A/E_{avg}$. A high SCI indicates that the structure is conserved, as the sequences fold individually as well as the alignment. The z-score and the SCI are combined and a classifier is trained for a model allowing the classification of any input alignment as structural ncRNA or "other".

1.3.4 *Family specific non-coding RNA gene finder*

Apart from general ncRNA gene finders, programs have been developed for the annotation of specific ncRNA families. Based on the extremely conserved cloverleaf secondary structure of the transfer RNA (tRNA) molecule, a software named `tRNAscan-SE` (Lowe and Eddy, 1997) has been developed. It used a fast pre-filter to identify candidates, which are then analyzed by a highly selective tRNA covariance model.

A program used for the annotation of rRNAs is `RNAmmer` (Lagesen et al., 2007). It uses an HMM trained on a large set of 16S/18S and 23S/28S rRNA genes. Compared to other ncRNAs, rRNA genes have the advantage of being strongly conserved on the sequence level over all phyla of life.

An algorithm developed for the detection of two classes of small nucleolar RNAs (snoRNAs), namely box C/D and box H/ACA snoRNAs, is `snoReport` (Hertel et al., 2008). The algorithm combines a secondary structure prediction with a machine learning algorithm. Two different models for the two snoRNA classes were developed and integrated in a prediction pipeline that uses a conserved secondary structure information together with position-specific weight matrices. These matrices are used to represent the characteristic sequence motifs of the two snoRNAs. The "ACA" sequence motif e. g., for the H/ACA snoRNAs as well as the secondary structure features are crucial for the function of snoRNAs (Bachellerie et al., 2002; Lafontaine and Tollervey, 1998). Those motifs or boxes have to be unbound in the secondary structure, thus a constrained folding algorithm is used for determining the secondary structure of the snoRNAs.

Specific ncRNA gene finders work mostly better than the general ones, as they can include more constraints and specific information about the properties of a certain ncRNA class.

1.4 CONTRIBUTIONS

1.4.1 *A microRNA gene prediction pipeline & A catalogue of animal microRNA genes*

microRNA (miRNA) genes are small ncRNAs that regulate gene expression at the post-transcriptional level. One of their hallmark features is their processing from stable stem-loop like precursor structures. This work introduces a novel method for the computational prediction of miRNAs in genomic data using homology information of known experimentally verified miRNAs, comparative genomic methods, and state-of-the-art machine learning techniques. Multiple animal genomes are first scanned for stable stem-loop structures which are then scored by a machine learning classifier to assign a miRNA gene probability. The classifier was trained on a set of features from known, experimentally verified miRNAs. Afterwards, putative *ab initio* miRNA predictions are filtered by an orthology assignment step, grouping similar sequences together. Another machine learning classifier, which was trained to distinguish alignments of orthologous miRNA genes from non-miRNA genes, is used to create a final set of miRNA predictions.

All predictions are made available through a web-accessible database named `miROrtho` (`http://cegg.unige.ch/mirortho`), presenting all miRNA predictions from a comparative point of view, and focusing on orthologous groups of miRNAs. `miROrtho` presents information complementary to the `miRBase` database, which mainly focuses on experimentally verified miRNA genes.

1.4.2 *A novel cis-acting replication element in rhinovirus genomes*

Based on the methodology and framework used to discover miRNA genes, further work was conducted on structural RNA elements in enterovirus genomes. This work presents the prediction of a novel structural RNA — a *cis*-acting replication element (*cre*) — in the recently described enterovirus species human rhinovirus C. The discovery of this element was enabled by an alignment of several different virus strains and a modified version of `RNAz`, scanning the alignment for conserved structural elements. Each of the three rhinovirus species contains a single *cre* which overlaps with different genes of the single open reading frame (ORF). The four enterovirus species HEV-A to D also comprise a *cre*, which overlaps yet another gene with regard to the rhinovirus species. However, the genomic position of the *cre* in four enterovirus species is conserved, as opposed to the three rhinovirus species. A reclassification of the four enterovirus species based on a high overall sequence identity and a commonly positioned *cre* element is proposed.

1.5 THESIS OUTLINE

The thesis is divided into two main parts. First, a novel procedure for the genome-wide prediction of miRNA genes in animal genomes will be presented. After an introduction to the biology of miRNA genes, machine learning methods, and current state-of-the-art miRNA gene predictors, I will detail the steps of a novel miRNA gene prediction

pipeline called `miROrtho`. The results of the application of the new method to predict miRNAs in several animal genomes will be presented, finishing with a discussion part which will feature unsolved problems in miRNA gene prediction and possible solutions.

Second, the discovery of a novel structural *cis*-acting replication element (*cre*) in rhinoviruses will be presented. After a review of the field and other known common structural elements in entero- and rhinoviruses, results presenting the *cre* element in the human rhinovirus species C will be compared to known similar elements in other species. The discussion part will detail the specific constraints for *cre* elements in entero- and rhinovirus species and their distinct genomic position in the virus' ORF.

The general discussion part will put the developments and achievements in a global context.

Part II

ANIMAL MICRORNA GENE PREDICTION

2 INTRODUCTION TO MICRORNA GENE PREDICTION

The final[1] version of the human genome estimated about 20,000–25,000 protein-coding genes (International Human Genome Sequencing Consortium, 2004). These numbers and previous estimates however, were focused on protein-coding genes, and mostly neglected non-coding RNA (ncRNA) genes. ncRNA genes rather produce functional RNA molecules than encoding proteins. Some ncRNAs have been known for some time such as ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small nuclear RNAs (snRNAs), or small nucleolar RNAs (snoRNAs). With the dawn of new millennium, several new classes of ncRNAs were described, such as microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), or endogenous small interfering RNAs (siRNAs) (reviewed in Carthew and Sontheimer, 2009; Klattenhoff and Theurkauf, 2008). These RNAs were classified as small RNAs (sRNAs) because they are shorter than 200 bp as opposed to long RNAs (lRNAs) (Kapranov et al., 2007). Many of the recently described ncRNAs seem to have roles in specific nucleic acid recognition, which can lead to post-transcriptional regulation of gene expression or to guiding RNA modifications.

To date, over 6,000[2] ncRNAs are annotated in human, including an increasing number of miRNAs (over 1,000 mature miRNA gene according to `miRBase 14`) (Griffiths-Jones et al., 2008). miRNA genes are short ncRNA genes that regulate gene expression in a post-transcriptional manner. This work introduces a new method for predicting miRNA genes[3] in animal genomes and the development of a comprehensive catalogue of miRNA genes.

After an introduction to the biogenesis and functions of miRNA genes, a series of machine learning methods used in this work will be introduced. A section about current state-of-the-art miRNA gene prediction approaches will conclude the introduction to the novel miRNA gene prediction pipeline presented in this work.

2.1 ANIMAL MICRORNA BIOGENESIS AND FUNCTION

The first miRNA identified was the lin-4 gene in *Caenorhabditis elegans* (Lee et al., 1993; Wightman et al., 1993). The lin-4 locus comprises a short 22 nt RNA, which represses the expression of the nuclear protein lin-14 (Lee et al., 1993; Wightman et al., 1993). The temporal regulation of lin-14 by the small RNA lin-4 is crucial for *C. elegans* larval development. A couple of years later, it was shown that the regulation is mediated by binding of lin-4 to the 3′UTR of lin-14 through partial sequence complementary (Ha et al., 1996; Olsen and Ambros, 1999). It was not until 2000 that a second miRNA, let-7, was discovered (Reinhart et al., 2000), which also controls development timing in *C. elegans*. The deep conservation of let-7 throughout the metazoans (Pasquinelli

1 Over 99% of the euchromatic genome has been sequenced with an error rate of approximately 1 event per 100,000 bases.

2 `Ensembl` release 56, `http://www.ensembl.org/Homo_sapiens/Info/StatsTable`

3 If not stated otherwise, the short term "miRNA" is used for the actual precursor microRNA (pre-miRNA) molecule.

et al., 2000), inspired follow-up studies establishing miRNAs as a new class[4] of post-transcriptional gene regulators (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001).

2.1.1 *Biogenesis*

Although mature miRNAs are only ~22 nt long, their primary genomic transcripts (primary microRNA (pri-miRNA)) are usually several kilobases long (Cai et al., 2004; Lee et al., 2004). pri-miRNAs are capped and polyadenylated, similar to messenger RNA (mRNA) or other polymerase II transcripts[5]. miRNAs are often clustered with other miRNAs, which implies that they could be transcribed as polycistronic transcripts. About 40% of animal miRNAs are found within the introns of protein-coding transcriptional units, another 40% are located in introns of non-protein coding transcriptional units (Kim et al., 2009; Rodriguez et al., 2004). A minor fraction of the known miRNAs overlap exonic regions of coding and non-coding transcriptional units.

The complete miRNA biogenesis pathway is summarized in Fig. 4 on the next page. After transcription of the pri-miRNA by mainly RNA polymerase II (Lee et al., 2004), the nuclear RNase III enzyme Drosha (DRSH-1 in *C. elegans*) excises a stem-loop structure to release the ~65 nt long pre-miRNA (Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Landthaler et al., 2004; Lee et al., 2003). This process requires DiGeorge syndrome critical region gene 8 (DGCR8) in mammals and Pasha in *Drosophila melanogaster* or *C. elegans* as a cofactor. The DGCR8-Drosha complex is also known as the *Microprocessor*. Recently an alternative pathway to generate pre-miRNAs that bypasses Drosha processing has been described. Small introns of the size of a pre-miRNA are spliced out and the lariat structure is opened and refolded as a stem-loop structure (Berezikov et al., 2007; Okamura et al., 2007; Ruby et al., 2007a). The resulting stem-loop structure is called a *mirtron*. The canonical and non-canonical *mirtron* pathway converge on the pre-miRNA molecule. The pre-miRNA molecule contains a 5′ phosphate and 3′ hydroxy terminus, as well as two- or three nucleotide 3′ single-stranded overhanging ends. The pre-miRNA is translocated from the nucleus to the cytoplasm by exportin-5 (Exp-5) (Bohnsack et al., 2002; Lund et al., 2004; Yi et al., 2003; Zeng and Cullen, 2004). Binding of the pre-miRNA to Exp-5 requires the guanine triphosphatase (GTPase) Ran (RanGTP) and is independent from the primary RNA sequence of the pre-miRNA (Zeng and Cullen, 2004). A recent study resolving the structure of the pre-miRNA nuclear export machinery has shown, that Exp-5:RanGTP recognizes the two nt 3′ overhang and the double-stranded stem of the pre-miRNA (Okada et al., 2009). Once in the cytoplasm, a second RNase III enzyme, Dicer (DCR-1 in *C. elegans*), processes the pre-miRNA stem-loop to release a ~22 nt long RNA duplex (Bernstein et al., 2001; Chendrimada et al., 2005; Förstemann et al., 2005; Grishok et al., 2001; Hutvágner et al., 2001; Jiang et al., 2005; Ketting et al., 2001; Saito et al., 2005). This process requires double-stranded RNA (dsRNA)-binding proteins like Loquacious in *D. melanogaster* (Förstemann et al., 2005), or the TAR

4 In the first papers, miRNA genes were named small temporal RNAs (stRNAs) due to their small size and temporal expression pattern.

5 Some miRNA genes associated with Alu repeats, however, can be transcribed by RNA polymerase III (Borchert et al., 2006).

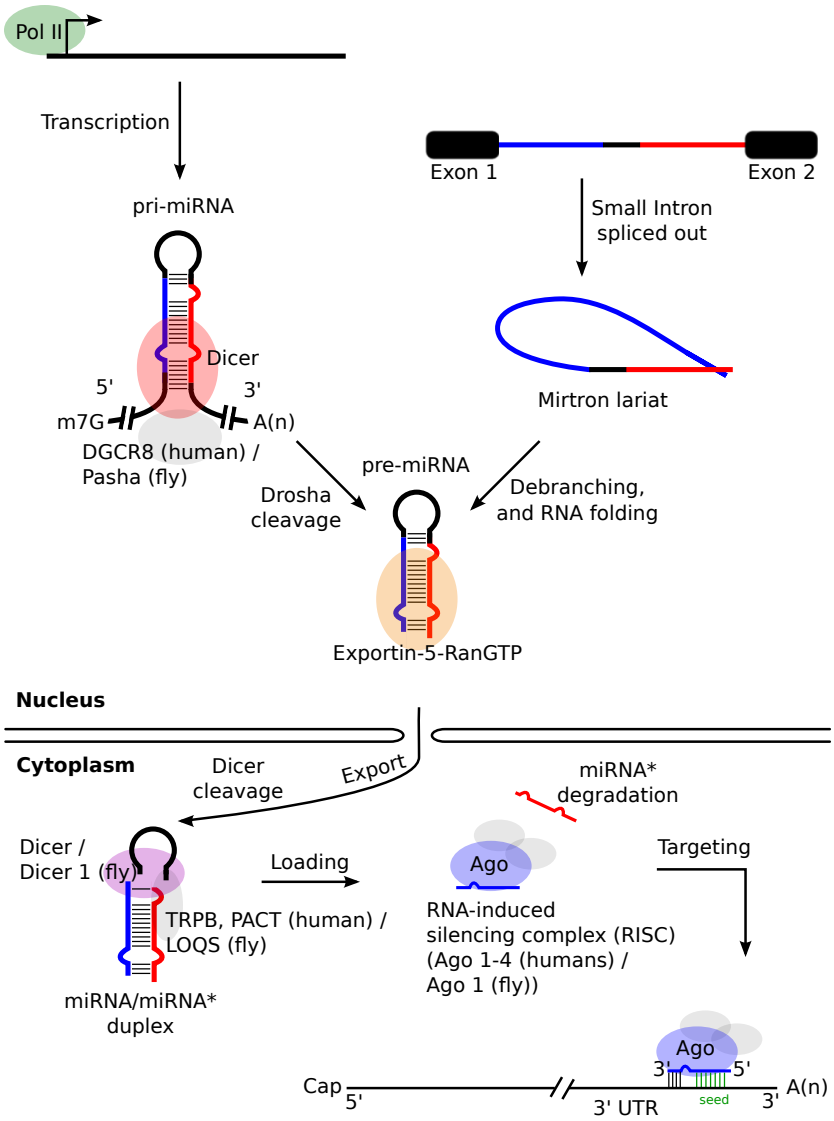


Figure 4: miRNA biogenesis pathway in animals. Genomic encoded long pri-miRNA transcripts are processed by the Drosha enzyme, which cleaves a stable stem-loop structure called the pre-miRNA. After Dicer processing, one strand of the duplex is integrated in the RNA-induced silencing complex (RISC) which contains an Ago protein. The silencing complex then targets genes to be down regulated. An alternative pathway, bypassing the Drosha processing, creates pre-miRNAs originating from short debranched and re-folded short introns which are therefore named *mirtrons*.

RNA-binding protein (TRBP) (Haase et al., 2005) and PACT (Lee et al., 2006) in humans. The duplex containing miRNA/miRNA* is likely to be unwounded by the helicase activity of Dicer and one strand, the mature miRNA, preferentially enters the RNA-induced silencing complex (RISC). The other miRNA* passenger strand is often degraded, although there are cases in which both arms are processed into functional miRNA genes (Stark et al., 2008, 2007a). In general, the strand with relatively unstable base pairs at the 5′ end is typically selected and loaded into the RISC (Khvorova et al., 2003; Schwarz et al., 2003). The RISC contains a member of the Argonaute (Ago) protein family. The complex carries out small RNA-directed gene silencing (Doench et al., 2003; Hammond et al., 2000, 2001; Hutvágner and Zamore, 2002) by binding to the 3′ UTR of target genes.

2.1.2 *Function*

Binding of a miRNA to its target mRNA typically leads to translational repression or exonucleolytic mRNA decay. Many animal miRNAs, are however only imprecisely complementary to their mRNA targets. For animal miRNAs, the binding specificity is mainly determined by the nucleotides 2–8 or 2–7 of the mature miRNA (Bartel, 2009). This 5′ region is also called the *seed region* of the miRNA.

Since 2001 a plethora of new members for this ncRNA class, together with hundreds of experimentally verified targets has been described (Papadopoulos et al., 2009). Computational predictions indicate, that more than half of all human mRNAs might be conserved targets of miRNAs (Friedman et al., 2009b). Table 1 on the facing page lists a selection of some miRNAs together with their experimentally verified target genes.

A recent study e. g., investigating the integration of miR-122 in hepatic circadian gene expression found hundreds of regulated mRNAs (Gatfield et al., 2009). miR-122 has previously been linked to the regulation of cholesterol and lipid metabolism. Another study showed that infection by hepatitis C virus is dependent upon the liver specific miR-122 miRNA (Jopling et al., 2005).

In summary, a number of further fundamental biological processes have been associated with miRNAs, ranging from development (Bejarano et al., 2009; Brennecke et al., 2003; Chen et al., 2006; Friedman et al., 2009a; Giraldez et al., 2005; Wightman et al., 1993; Zeng et al., 2009), to metabolism (Esau et al., 2006; Gatfield et al., 2009; Poy et al., 2004; Xu et al., 2003), cardiac (Zhao et al., 2007) and immune system functions (O'Connell et al., 2007; Perry et al., 2008). Aberrant miRNA expression has also been linked to cancers (Hayashita et al., 2005; Lu et al., 2005; Segura et al., 2009). Dicer null mutants show diverse severe developmental defects (Bernstein et al., 2003; Wienholds et al., 2003).

2.2 MACHINE LEARNING APPROACHES

Machine learning (Alpaydin, 2004; Bishop, 2007) describes the design and development of algorithms which "adapt" their behavior according to given data. Such automatized learning methods have a wide application in biology. A branch of machine learning called supervised learning deals with the construction of functions which map unknown data samples to a specific class, based on previously seen objects. To

Table 1: Selection of animal miRNAs with assigned functions. Note, the first miRNA genes which were described were not assigned "mir-" names (e. g., let-7, bantam).

MICRORNA	FUNCTION	TARGET(S)
cel-lin-4	Development timing	lin-14[a], lin-28[b]
cel-let-7	Development timing	lin-41[c], hbl-1[d]
cel-lsy-6	Neuronal cell fate	cog-1[e]
dme-bantam	Cell death, proliferation	hid[f]
mmu-mir-7	Pathogenesis of Parkinson's disease	SNCA[g]
hsa-mir-21	Cancer	Cdc25A[h]
hsa-mir-145	Cancer	c-Myc[i]
dme-mir-8	Insulin signaling	USH[j]
hsa-mir-200	Insulin signaling	FOG2[j]
mmu-mir-122	Hepatic circadian gene expression	PPAR-beta/delta, Smarcd1/Baf60a[k]

dme, *D. melanogaster*; cel, *C. elegans*; mmu, *M. musculus*; hsa, *H. sapiens*

[a] Lee et al. (1993); Wightman et al. (1993) [b] Moss et al. (1997) [c] Slack et al. (2000) [d] Abrahante et al. (2003); Lin et al. (2003) [e] Johnston and Hobert (2003) [f] Brennecke et al. (2003) [g] Junn et al. (2009) [h] Wang et al. (2009) [i] Sachdeva et al. (2009) [j] Hyun et al. (2009) [k] Gatfield et al. (2009)

date, various classification algorithms have been developed with an increasing performance and robustness to noisy data. Noisiness describes the fact that some of the training examples are mislabeled. A overview of the classical steps in supervised learning is shown in Fig. 5 on the next page.

2.2.1 *Supervised learning algorithms (classifiers)*

Classification is a sub-discipline of supervised learning. In general a classification model (classifier) is trained on data which consist of labeled objects, represented by numerical vectors. Each object is associated with a class-label. The goal of supervised learning is to create a general classifier model, that can be used to find the correct class for new data based on statistical properties, learned from the labeled training data.

In this work, various classifiers were tested for their performance to predict miRNA genes. Based on a set of features (Tables 7 to 8 on pages 40–50) general models are derived, enabling it to assign a class label (miRNA or non-miRNA gene) to new data. Known miRNAs genes and non-miRNA stem-loop like sequences are used as training data. The training and testing data object are labeled and unlabeled n-dimensional numerical vectors, describing characteristic features of miRNA genes. Depending on the classifier, the output is either discrete (e. g., miRNA gene or non-miRNA gene) or continuous, in which case an object is assigned a probability score of being a miRNA gene, ranging from 0 to 1.

The various tested classifiers are introduced in the following sections.

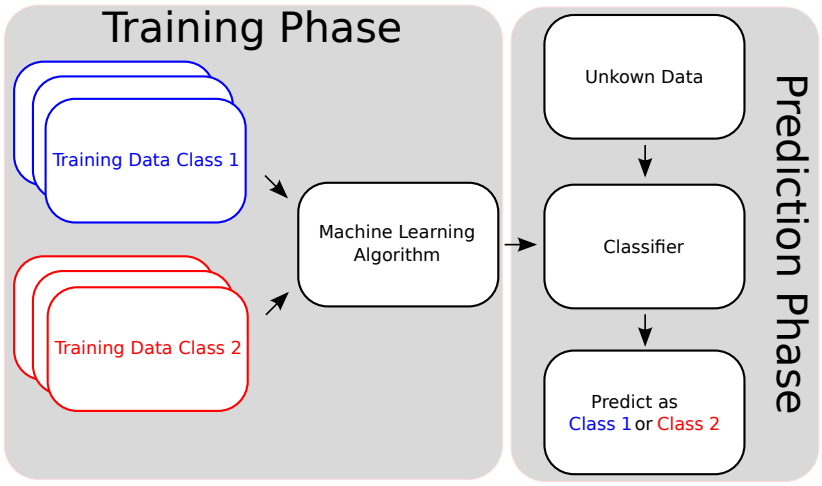


Figure 5: Basic concepts of machine learning. Labeled data samples are used to train a machine learning algorithm (a classifier) — this is called the training phase. In the prediction phase, the classifier is applied to a set of unknown data samples to assign a class label. The example shows a two-class classification problem with training data labeled as either one of the classes. The core component of this method is the machine learning algorithm — here a classifier — which is used to derive a general model based on a common set of features describing characteristics of the two classes.

2.2.1.1 *Naive Bayesian Classifier*

A Naive Bayesian classifier selects the most likely classification $c \in \{-1, 1\}$, given the attributes $a_1, a_2, \ldots, a_n$ and training samples with known class labels. This can be formally written as $P(c|a_1, a_2, \ldots, a_n)$, the probability of assigning a class c given the attributes (features) $a_1, a_2, \ldots, a_n$. To ease the construction of the model, all attributes are assumed to be independent, therefore the probability for a certain class can be calculated from the independent probabilities. Actually the name "naive" comes from the fact that the variables are assumed to be independent. The individual probabilities are computed using Bayes' Theorem which states:

$$P(c|a) = \frac{P(a|c)P(c)}{P(a)} \tag{2.1}$$

The probability of a hypothesis c given observed data a depends on the probability of that data given the hypothesis. The greater the expression in eq. 2.1, the greater the probability, that given a certain attribute value, the sample belongs to the class c (in this case miRNA or non-miRNA). Since $P(a)$, the probability of a certain attribute, does not change over the classes c, it can be ruled out. So given a set of attributes $a_1, a_2, \ldots, a_n \in a$, the class of a new sample can be computed by maximizing the probability for each individual attribute. The advantage of the Naive Bayesian classifier is that it scales easily with increasing number of attributes and is fairly easy to compute. However, mutual information and correlations between different attributes are not taken into account as these are assumed to be independent.

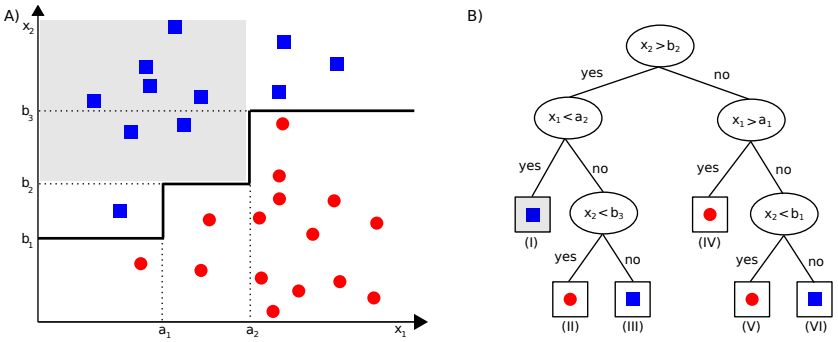


Figure 6: Binary decision tree for a two-class decision problem, with dimensionality $p = 2$. A) Positive (blue squares) and negative (red circles) training data points are plotted. The decision boundary is shown by a thick black line perfectly separating the two classes using two features x_1 and x_2. The gray area reflects the decision area also marked gray in (B). B) A binary decision tree derived from the data set. The different regions are covered by the nodes (I) – (VI). Following the edges and the nodes on the tree allows any new data to be classified in its most likely class. Figure from Tarca et al. (2007).

2.2.1.2 *Decision Tree Classifier*

A decision tree classifier is a supervised machine learning method that uses a tree-like graph or model of decisions and their possible consequences, to describe a set of complex data vectors. Fig. 6 shows a simple binary decision tree used to classify a two-dimensional two-class input set. A basic decision tree consists of nodes, edges and a final prediction outcome (Fig. 6). At each node of the tree, the algorithm chooses one attribute of the training data, that splits best the input samples into the two classes. The criterion for the split is the normalized information gain that results from splitting an attribute at a specific value. The tree is constructed by subsequently adding nodes with descending information gain at the splits, reaching the bottom of the tree. Thus, the attributes that best split the two-classes are at the top of the tree. The advantage of a decision tree over an Support Vector Machine (SVM) or a Neural Network classifier is that it is easy to interpret. The internal model's internal values and variables can be viewed and easily adapted as opposed to an SVM model. A decision tree allows easy interpretation of the data using simple mathematics.

2.2.1.3 *Multilayer Perceptron Neural Network*

A multilayer perceptron is a simple feed-forward neural network. The original ideas for this concept have been developed by Rosenblatt (1958), inspired by the function of the brain. It was not until the 1980s that the interest in neural networks was revived by Rumelhart et al. (1986). The basic principles of a perceptron are depicted in Fig. 7 on the next page. In its simplest form it consists of three layers — an input layer, a hidden layer, and an output layer. Each layer is built from nodes called neurons in allusion to a nervous system. The input layer receives the input variables (e. g., x_1, x_2) which are linearly scaled to a range of $\{-1, 1\}$. The normalized input variables from the input layer are redistributed along the neurons in the hidden layer with a weighting factor. Each neuron in the hidden layer contains a finite-dimensional

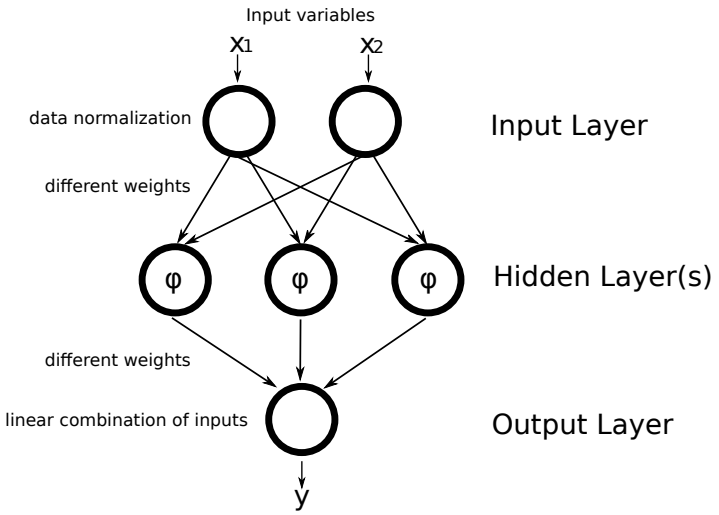


Figure 7: A multilayer perceptron with an activation function φ for classifying two-dimensional data. Values for the features x_1 and x_2 are fed to the input layer which normalizes the data. The output from the input layer is then distributed over the neurons in the hidden layer with an associated function φ. Weighted outputs from the hidden layer are combined in the output layer, leading to a unique prediction outcome y. The depicted three layer feed-forward perceptron is fully connected. The output of each input and hidden neuron is distributed to all the neurons in the following layer without any recursions.

real-valued feature vector coming from the previous layer. A neuron in the output layer is fed by a combination of input vectors that were processed by a transfer function φ. After linear scaling, the output layer creates a prediction value y. The architecture, such as the number of layers, the number of neurons, and the transfer function have to be created individually for each classification task. The model is then trained with labeled data. Learning is done by predicting a label for each input sample, leaving the weights in the model unchanged when the predicted output matches the target, and changing them when it does not.

2.2.1.4 *Support Vector Machine*

Support Vector Machines (SVMs) were first introduced by Vladimir Vapnik and his collaborators (Vapnik and Chervonenkis, 1974) for solving classification problems by finding an optimal hyperplane separating objects. As for all classifiers, the training data for the SVM is a list of p-dimensional data vectors with associated class labels[6] assigned to it. Fig. 8 on the facing page shows an example of a two-dimensional data set with data samples from different classes. Several valid decision boundaries exist, that all separate the training data. However, among all possible boundaries, SVMs find a boundary with a maximum margin between the two classes (Fig. 8 on the next page). Maximizing the margin minimizes the upper bound of the risk to misclassify a new data sample. Thus, SVMs generalize well on unseen data (Vapnik, 1998).

6 This work uses a two-class SVM distinguishing miRNA and non-miRNA genes.

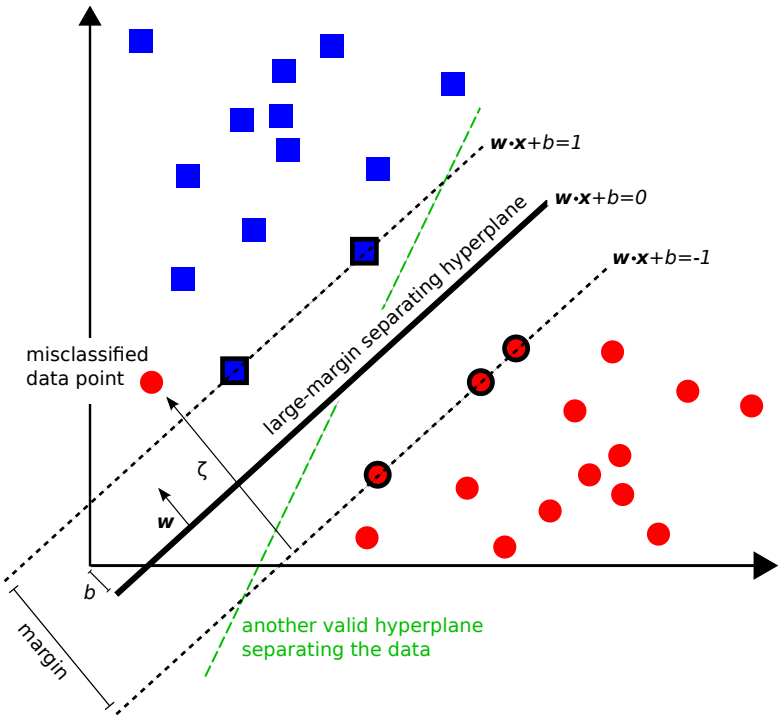


Figure 8: A two-class Support Vector Machine (SVM) showing training data points and a maximum margin hyperplane separating the data. The data points are represented by two-dimensional vectors. Positive training samples are represented as blue squares, negative ones as red circles. The maximum-margin decision boundary (hyperplane) found by the SVM is shown by a thick black line. Another valid hyperplane is depicted by a green dashed line which also separates the data, but does not maximize the margin between the positive and the negative training data. Samples along the dashed black line are called support vectors (SVs). They are marked by a solid black line, encapsulating the respective sample data points. The hyperplane of an SVM is exclusively defined upon those SVs. A red sample data point on the left side of the hyperplane represents a misclassified vector. The SVM deals with those, by introducing non-negative slack variables ξ and a penalty function measuring classification errors.

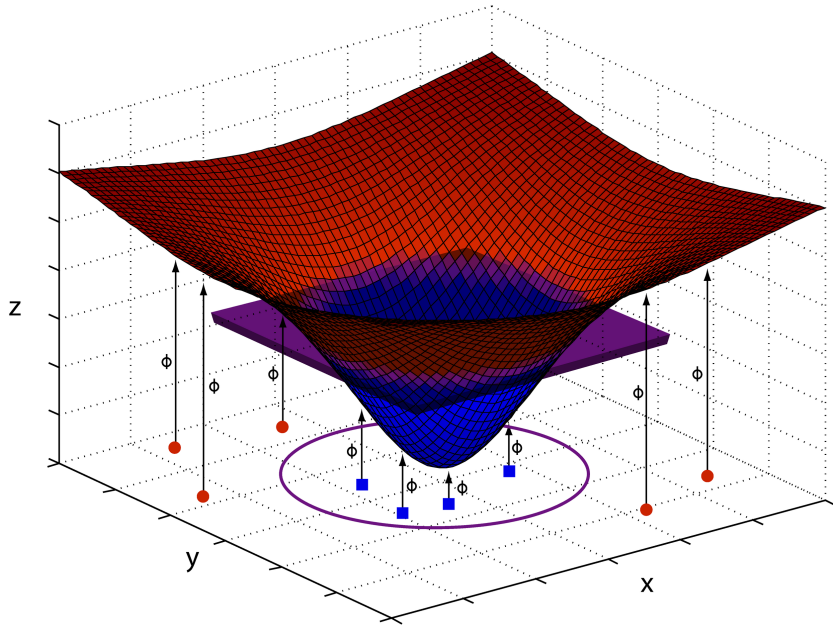


Figure 9: Support Vector Machine: Kernel trick. A data set of blue squares and red dots with attributes x, and y are not easily separable in the two-dimensional input space (purple ellipse in x, y). By applying a kernel mapping function ϕ, every vector is transformed into a higher dimensional feature space (x,y,z) which allows the separation of the data using a linear hyperplane (purple plane). In this example a Gaussian function was used as the radial basis function kernel.

What follows is a formal introduction to the mathematics behind an SVM: Let us define a two-class labeled training data as $\{(x_i, y_i), \ldots, (x_m, y_m)\}$ where $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$. An SVM finds a hyperplane that divides points belonging to either class $y = 1$ or class $y = -1$ with a maximum-margin — therefore SVMs are also called large-margin classifiers. During the training of the SVM an optimal hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ is found. $\mathbf{w}$ is the p-dimensional vector perpendicular to the hyperplane and b is the bias. The $\cdot$ is the dot product of the two vectors $\mathbf{w}$ and $\mathbf{x}$. The objective of the SVM is to find parameters for $\mathbf{w}$ and b, to maximize the margin between the labeled training samples (Fig. 8 on the preceding page). By maximizing the margin of the separating hyperplane, the risk of overfitting the model to the training data can be lowered. To allow for classification errors in the training process due to noisy or mislabeled data, non-negative slack variables ξ, and a penalty function C can be introduced.

Classification problems are not always separable by a linear decision boundary. Therefore, the *kernel-trick* can be used to transform every point from the input space to a higher dimensional feature space by applying a kernel function ϕ (Fig. 9). This work uses the radial basis function (RBF) kernel function to transform the data.

Optimal values for the RBF kernel parameters γ and C were determined using a grid search on the hyperparameter space with 10-fold cross-validation on each set of parameters maximizing the area under the curve (AUC) of the receiver operating characteristic (ROC) curve (Fig. 10 on the facing page). A grid search is a heuristic method which tries to maximize an output value by testing various combinations of

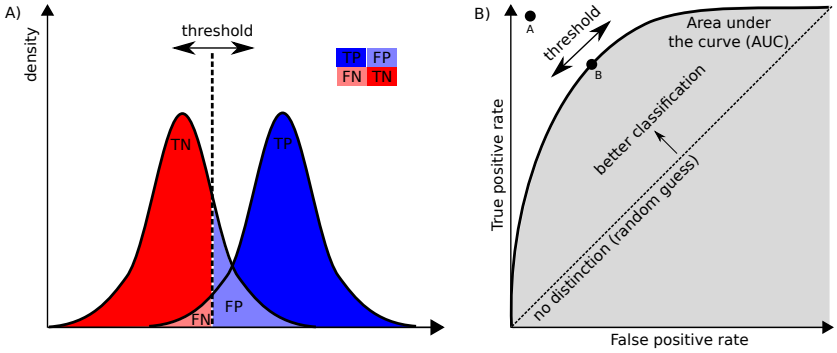


Figure 10: Binary classifier with continuous output and ROC curve interpretation. A) A binary classifier with continuous outcome can produce four possible outcomes for any threshold: true negative (TN), false negative (FN), true positive (TP), and false positive (FP). The proportion of TN, FN, TP, and FP can be changed by adjusting the classifier's threshold, however, for a non-perfect classifier, there is always an overlap between the classes B) ROC curve for a binary classifier with continuous outcome. The true positive rate is plotted versus the false positive rate for a varying threshold. The dot "A" represent a perfect classifier which does not make any errors. The diagonal shows the line of no discrimination — a classifier with such a performance just makes a random guess. Generally the further towards the upper-left a classifier lies, the better it performs. The gray area under the ROC curve is called the area under the curve (AUC) and is used as a single value to describe a classifier's performance. Classifier that perform better in distinguishing several classes have higher values for the AUC. The AUC is equal to the probability that a classifier will rank a randomly chosen instance of one class higher than a randomly chosen sample from the other class (Fawcett, 2006).

input parameters on a function. Cross-validation is a method in which the input training data is partitioned in a new smaller training set and a testing set. A model trained on the new training set is then applied to the test set and the class labels are determined. This repartitioning procedure is repeated several times with randomly chosen subsets. In the end the predicted labels are compared to the true ones, and the performance of the classifier on unseen data can be evaluated. For a 10-fold cross-validation the data is split in $1/10$ th test data and $9/10$ th training data repeatedly. For determining optimal kernel values for γ and C, first a grid space with $\log_2 C \in \{-5, -3, \ldots, 15\}$ and $\log_2 \gamma \in \{-15, -13, \ldots, 3\}$ is constructed. Second, for each set of hyperparameters C and γ, a 10-fold cross-validation procedure is conducted, and the average AUC is reported. Then, a final model is built using the parameters C and γ that lead to the highest AUC value in the ROC plot.

For the performance evaluation of different classifier in this work, the `Weka 3` (Hall et al., 2009) toolbox is used which implements a Java version of the `LIBSVM` package (Chang and Lin, 2001). For the final miRNA gene prediction pipeline, the `C` implementation of `LIBSVM` version 2.89 was used.

Table 2: Confusion matrix for a classifier. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. The purpose of this matrix to check if the system confuses the two classes (i. e. mislabel the classes).

		actual value	
		p	n
prediction outcome	p′	True Positive (TP)	False Positive (FP)
	n′	False Negative (FN)	True Negative (TN)

2.2.2 *Confusion matrix and receiver operating characteristics*

The predictions of a classifier on testing data with known labels can have four distinct outcomes: A positive sample that is scored positive by the classifier is called a TP, if the sample is classified negative the outcome is a FN prediction. Analog, a negative sample which is attributed to the negative class is called a TN, and a FN if it is classified as belonging to the positive class. The actual and predicted classifications done by a classification system can be written down in a confusion matrix (Table 2) from which various performance measurements can be calculated (eqs. 2.2 to 2.6 on this page). These can be directly used to compare the capacity of various classifier to correctly predict class labels for test data.

$$Sn = \frac{TP}{TP + FN} \tag{2.2}$$

$$Sp = \frac{TN}{TN + FP} \tag{2.3}$$

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.4}$$

$$FDR = \frac{FP}{TP + FP} \tag{2.5}$$

$$MCC = \frac{(TP \cdot TN - FP \cdot FN)}{\left(\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}\right)} \tag{2.6}$$

where TP = true positive, TN = true negative, FP = false positive, FN = false negative, Sn = sensitivity, Sp = Specificity, FDR = false discovery rate, ACC = accuracy, and MCC = Matthew's Correlation Coefficient.

For visualizing the performance of a classifier, ROC curves can be plotted using `R` (R Development Core Team, 2009) and the package `ROCR` (Sing et al., 2005). A ROC is a simple plot of the true positive rate versus the false positive rate for a binary classifier system as its discrimination threshold is varied (Fig. 10 on the previous page). A ROC curve allows for testing of the classifier's performance over different thresholds. Furthermore, the area under the curve (AUC) value can be used as a single value to compare various classifiers.

2.3 GENE PREDICTION AND GENE ANNOTATION

Gene prediction describes the endeavor of finding functional regions in genomic sequences. Different approaches exist which can be mainly

classified into sequence similarity, comparative genomic, and *ab initio* methods. Methods based on sequence similarity use extrinsic evidence such as expressed sequence tags (EST) and copy DNA (cDNA) data or homologous sequences from other species, to search the target genome. `BLAST` (Altschul et al., 1990) and similar algorithms are widely used to find related sequences. Basically those methods always compare a sequence of interest with other sequence data. *Ab initio* methods, however, infer a statistical model from sequences which can than be used to find new genes in genomic sequence data.

2.4 PRINCIPLES OF COMPUTATIONAL MICRORNA GENE PREDICTION

Gene prediction programs which are designed for the prediction of miRNA genes face different problems, compared to gene predictors for protein-coding sequences. miRNA genes do not show any strong statistical sequence properties that can be used in a model. Furthermore predictions are also hampered by the short length of the miRNA and the thereby low information content of the molecule.

Nevertheless, some intrinsic features of miRNA genes can be used for gene prediction. Almost all pre-miRNAs fold into non-perfect stem-loop-structures with the mature miRNA located either in the 5′ or 3′ arm. Nevertheless, a stem-loop structure is not limited to miRNA genes; a lot of other coding or ncRNAs, such as rRNAs, tRNAs, and mRNAs can also harbor similar structures (Zhang et al., 2006). Studies have shown, that the free folding energy is generally lower for pre-miRNAs than for random sequences with the same nucleotide composition (Ambros et al., 2003; Bonnet et al., 2004). The minimum free folding energy index (MFEI) introduced by Zhang et al. (2006) normalized the minimum free energy (MFE) by GC content and the length of the precursor. This MFEI tends to be higher for pre-miRNAs than for other non-miRNA stem-loops. The low free folding energy combined with a stem-loop secondary structure (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001) and a strong evolutionary conservation (Berezikov et al., 2005; Grad et al., 2003; Lim et al., 2003a,b) are the most predictive features to distinguish miRNA gene stem-loops from non-miRNA stem-loops found in mRNAs or other ncRNAs.

Some of the miRNAs annotated in `miRBase` (Griffiths-Jones et al., 2008) were identified by homology to previously annotated miRNAs (45%) based on a high sequence similarity, while others were confirmed based on experimental evidence[7] (55%) (Bentwich et al., 2005; Grad et al., 2003; Lai et al., 2003). Recently, other approaches using high-throughput sequencing technologies became more and more important for miRNA discovery (Friedländer et al., 2009, 2008; Lu et al., 2008; Wheeler et al., 2009).

Methods which have been developed for miRNA gene prediction can be grouped into some main categories. Some methods however, also use combined efforts, blurring the boundaries between the approaches.

7 E. g., direct cloning technologies, Northern blotting, PCR, 5′ RACE, qRT-PCR, *454*, Solexa sequencing

Table 3: Selection of homology-based miRNA gene finder

ALGORITHM	PUBLICATION	SOURCE CODE
"based on ERPIN"	Legendre et al. 2005	✓[a]
miralign	Wang et al. 2005	–
miRNAminer	Artzi et al. 2008	on request
Mireval	Ritchie et al. 2008	–

[a] `http://tagc.univ-mrs.fr/erpin/`

HOMOLOGY-BASED APPROACHES

Homology-based methods use external information, e. g., sequences, for miRNA gene discovery. A straight-forward method to annotate homologous miRNAs is to use sequence-based similarity searches. `BLAST` e. g., can use a query set of known, annotated miRNA genes to search for homologous sequences in genomes. If the complete pre-miRNA sequence is used as a query, it is mainly the mature regions that shows extensive sequence conservation between organisms. This method works well for conserved miRNAs such as e. g., mir-122, however `BLAST` parameters have to be adapted, to allow for very sensitive searches (Wang et al., 2005; Weber, 2005). A more sensitive search can be achieved by incorporating sequence and structural information in the search or by building a sequence profile of a known miRNA family. This profile can then be used to find new members of this family in other genomes. Although homology-based methods work well in recovering homologs to known, already annotated miRNA genes in relatively close species, they are obviously limited to the discovery of new members of already known families.

A selection of currently available homology-based miRNA gene finders is listed in Table 3.

FILTER BASED AND COMPARATIVE GENOMIC APPROACHES

The first computational miRNA prediction methods used filter-based approaches. Two relatively closely related genomes like *C. elegans* and *C. briggsae* (Grad et al., 2003; Lim et al., 2003b), or *D. melanogaster* and *D. pseudoobscura* (Lai et al., 2003) were scanned for conserved DNA blocks which fold into conserved stem-loop structures. Different groups used different approaches to filter potential candidates from those conserved stem-loops. Grad et al. (2003) filtered the initial candidates for matches, mismatches, and gap patterns of the stem region as well as GC content and MFE to reflect the characteristics of previously known miRNA genes in *Caenorhabditis*. Lim et al. (2003b) (`miRseeker`) found around 36,000 conserved hairpin sequences in two worms. Based on those and a background set of non-conserved hairpins, the authors developed a log-odds scoring scheme considering several features of the mature miRNA region such as the base pairing pattern to filter out false-positive predictions. `miRseeker` was also used to scan two *Drosophila* genomes and to extract conserved stem-loops. Besides scoring for MFE and a penalty factor for large asymmetric loops and bulges, the authors first used the typical divergence pattern of already known miRNAs (in general few nucleotide divergence in the stem regions, more divergence in the terminal loop) to score new conserved hairpins. These methods, however, rely heavily on the

phylogenetic distance of the studied species and the miRNA families. Filter-based approaches need a set of two input species with an "optimal" divergence to be able to gather enough information from the alignments to get reliable miRNA predictions.

To extend filter-based approaches several additional criteria were used. As many miRNAs are known to appear in cluster of close proximity, studies were done searching for miRNA genes neighboring already annotated miRNAs (Altuvia et al., 2005; Ohler et al., 2004). In most cases, only two or three miRNAs cluster together. Nevertheless, some clusters contain many more members, like the mir-17 cluster in mammals containing six members (Lagos-Quintana et al., 2001; Mourelatos et al., 2002; Tanzer and Stadler, 2004) or a mouse cluster of over 40 miRNAs located within a ~1 Mb DNA stretch (Seitz et al., 2004).

Another way to filter potential candidates is based on the biological function of miRNAs, targeting genes to be down regulated by seed pairing to their 3′ UTR. Mammalian 3′ UTRs were searched for highly conserved short motifs. At least some of these motifs are likely to be real miRNA target sequences. Conserved and stable stem-loops in genomes containing conserved sequences complementary to the previously identified motifs (Xie et al., 2005) can then be scored as putative miRNA candidates.

A filter-based approach which is extended to several genomes is called comparative genomic. Phylogenetic shadowing is a multiple-genome comparative genomic approach, taking the phylogenetic relationships between the analyzed species into account. In Berezikov et al. (2006a), the authors studied human miRNA genes together with sequenced orthologous copies of the same miRNAs in ten different primate species. An orthologous sequence is a special case of a homologous sequence, that is the closest homolog in another species which has arisen via a speciation event. The orthologous genes of several primate species were aligned, and a typical conservation pattern was derived, showing a high conservation for the mature part, a lower conservation for the loop region, and a striking drop in conservation for sequences immediately flanking the pre-miRNA stem-loops. Using this miRNA specific conservation profile, the authors used species versus species sequence comparisons to search for additional miRNA genes. They proposed about 1000 miRNA genes for human, which was much larger than previous estimates of only 255 human miRNA genes (Lim et al., 2003a).

DEEP-SEQUENCING BASED APPROACHES

With the recent development of high-throughput sequencing technologies (Margulies et al., 2005; Shendure et al., 2005), deep sequencing of a pool of size fractionated RNA samples became feasible. After creating a cDNA library from an RNA sample (Lau et al., 2001), the reads contain putative miRNA genes and degradation products of different ncRNAs and mRNAs. Several studies (Berezikov et al., 2010; Lu et al., 2008; Ruby et al., 2007b) aligned short reads from an organism to the genome sequence, and extracted putative new miRNAs by examining the pattern of the reads matching the genome and superimposing the underlying folding of the sequence. Candidates were further filtered

Table 4: Selection of high-throughput sequencing based miRNA gene finder

ALGORITHM	PUBLICATION	SOURCE CODE
miR-Intess ™	based on Berezikov et al. 2006a,b	–
miRDeep	Friedländer et al. 2008	✓[a]
miRanalyzer	Hackenberg et al. 2009	✓[b]

[a] `http://www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/index.html`

[b] `http://web.bioinformatics.cicbiogune.es/microRNA/miRanalyser.php`

by criteria such as a low MFE and evolutionary conservation[8] of stable stem-loop structures.

A more elaborate method called `miRDeep` (Friedländer et al., 2008) uses a probabilistic model to assess the compatibility of the pattern of small RNA reads with properties of miRNA biogenesis. This model accounts for the fact that true pre-miRNAs should have frequent reads corresponding to the mature region and less frequent reads corresponding to the loop region of the hairpin or the miRNA* region.

Most of the animal euchromatin is transcribed (Kapranov et al., 2007; Manak et al., 2006) thus, deep sequencing always discovers many unspecific degradation fragments from several sources. Therefore lower coverage deep sequencing (Lu et al., 2008) may reveal many singletons, that were later shown not to be actual RNase III products after deeper sequencing efforts (Berezikov et al., 2010).

Table 4 shows the main currently available miRNA gene predictors based on deep sequencing of short RNAs.

MACHINE LEARNING BASED APPROACHES

Recently, many different machine learning algorithms for the prediction of miRNA genes have been published (Table 5 on page 32). They formalize the definition and combination of different features that have already been used in filter-based approaches. Although some of the feature such as few asymmetric loops or a low free folding energy have been already been used in filter-based approaches, problems arise when is comes to choose reasonable thresholds for each of the features. Another problem is the weighting of features, as some might have more predictive power than others. To solve these problems, supervised learning methods have been developed. Those methods can create a classification model based on automatic learning from data. Two set of training data are usually used — a positive set drawn from known, annotated miRNA genes (e. g., `miRBase`) and sets of negative training data containing non-miRNA stem-loop like sequences extracted from random genomic loci or other ncRNAs. Many different machine learning algorithms have been used, like Support Vector Machines (SVMs) (Lu et al., 2008), Random-Forest classifiers (Jiang

8 Evolutionary conservation of stem-loops alone, however, is not the *ultima ratio* criteria for miRNA annotation — other functional ncRNAs or mRNAs can also contain conserved stem-loop sequences.

et al., 2007), Naive Bayes classifiers (Yousef et al., 2006), and even genetic programming (Brameier and Wiuf, 2007). Most approaches use a similar set of features describing sequence properties, topological, thermodynamical, and sequence complexity related features.

Table 5: Selection of *ab initio* miRNA gene finder for animal miRNAs

ALGORITHM	PUBLICATION	SOURCE CODE
MiRscan	Lim et al. 2003b	–
"no name"	Grad et al. 2003	–
miRseeker	Lai et al. 2003	–
"no name"	Berezikov et al. 2005	–
mir-abela	Sewer et al. 2005	–
PalGrade	Bentwich et al. 2005	–
ProMiR I & II	Nam et al. 2006, 2005	–
triplet-SVM	Xue et al. 2005	✓[a]
BayesMiRNAFind	Yousef et al. 2006	–
RNAmicro	Hertel and Stadler 2006	✓[b]
miRNA SVM	Helvik et al. 2007	–
miPred	Ng and Mishra 2007	–
MiPred	Jiang et al. 2007	–
MiRFinder	Huang et al. 2007	–
mirCoS	Sheng et al. 2007	–
miRRim	Terai et al. 2007	✓[c]
MiRPred	Brameier and Wiuf 2007	–
One-ClassMirnaFind	Yousef et al. 2008	–
MiRank	Xu et al. 2008b	on request
"no name"	Xu et al. 2008a	–
miR-KDE	Chang et al. 2008	–
CID-miRNA	Tyagi et al. 2008	–
miROrtho	Gerlach et al. 2009	on request
HHMMiR	Kadri et al. 2009	–
microPred	Batuwita and Palade 2009	✓[d]
SSCprofiler	Oulas et al. 2009	–
"L score strategy"	van der Burgt et al. 2009	–

[a] `http://bioinfo.au.tsinghua.edu.cn/mirnasvm/`

[b] For ncRNA alignments.
`http:/www.tbi.univie.ac.at/~jana/software/RNAmicro.html`

[c] No code for calculating the required feature vectors. `http:/mirrim.ncrna.org/`

[d] `http:/web.comlab.ox.ac.uk/people/ManoharaRukshan.Batuwita/microPred.htm`

3 MATERIAL AND METHODS

The miROrtho pipeline was implemented using Perl, Python, and existing C-libraries. All steps in the gene prediction pipeline are independent. The two main cores of the pipeline are two Support Vector Machine (SVM) models. One of which was trained on single miRNA genes and another which was trained on alignments of related miRNA genes. I will refer to the first one as the *sequence SVM* and the second one as the *alignment SVM*. An overview for the complete miROrtho pipeline method is depicted in Fig. 11. A flow chart with all the programs used to produce the data shown in the miROrtho database is depicted in Fig. 12 on the following page.

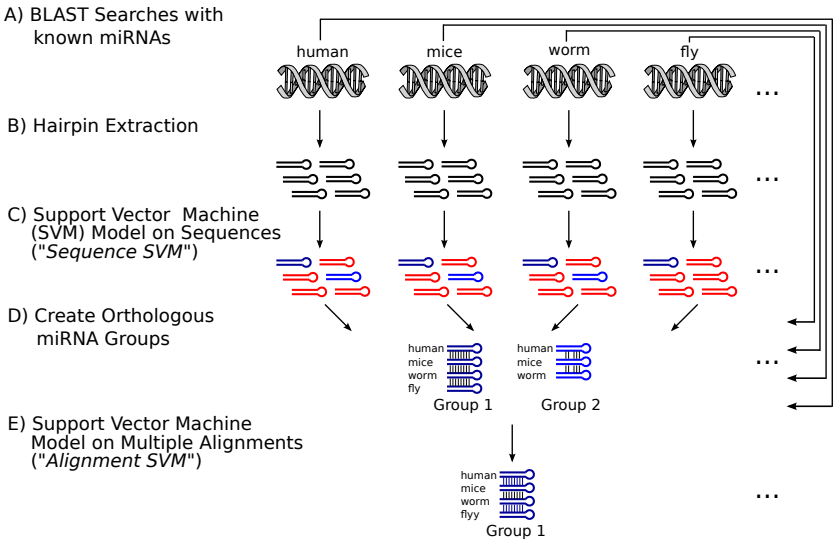


Figure 11: Overview of the miROrtho prediction pipeline. A) Genomes are scanned for homologs of known miRNA genes using sensitive BLAST searches and filters. B) Genomes are also scanned for stable stem-loop like structure. C) Stable stem-loop sequences are then subject to an SVM trained on miRNA and non-miRNA sequences (*sequence SVM*). D) By grouping the putative predictions retried from the SVM prediction and BLAST searches into orthologous groups, alignments of related sequences are created. E) Those alignments are then scored with a second SVM (*alignment SVM*), this time trained on known miRNA family alignments from miRBase and an earlier version of miROrtho.

3.1 GENOME SEQUENCE DATA

Genomic data was retrieved from the following resources: Ensembl[1] (Hubbard et al., 2009), Ensembl Metazoa[2], the UCSC Genome Browser[3], the Human Genome Sequencing Center at the Baylor College of

1 http://www.ensembl.org
2 http://metazoa.ensembl.org
3 http://genome.ucsc.edu/

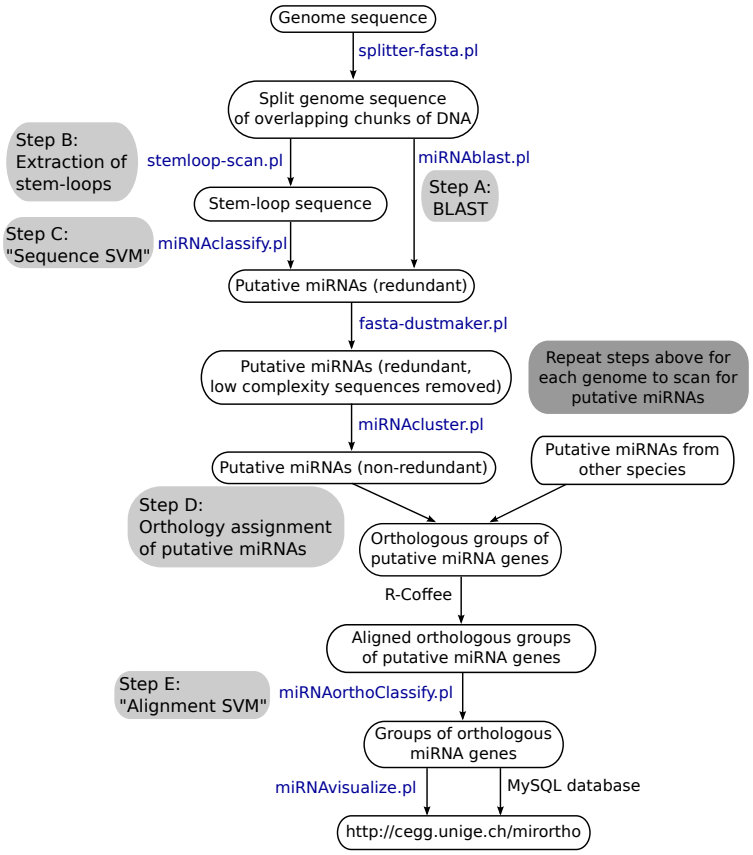


Figure 12: Flow diagram of the `miROrtho` miRNA annotation pipeline steps. The different steps are indicated by black boxes, programs written for the `miROrtho` pipeline are indicated in blue. The basic steps A, B, C, D, and E are marked by gray boxes and refer to the overview Figure 11 on the preceding page. The upper part of the flow chart including steps A, B, and C, is repeated for each genome separately. *Ab initio* miRNA predictions from several genomes are grouped in step D. The final predictions can be accessed via the web at `http://cegg.unige.ch/mirortho`.

Medicine[4], Genoscope[5], the Broad Institute[6], the Washington University Genome Sequencing Center[7], the DOE Joint Genome Institute[8], VectorBase[9], the silkworm genome database[10], Genbank[11], the Sanger Institute[12], and the Chinese National Human Genome Center at Shanghai[13].

Table 10 on page 58 lists all genomes and the number of recovered homologous and non-homologous miRNA candidates as compared to `miRBase 14`.

4 `http://www.hgsc.bcm.tmc.edu`
5 `http://www.genoscope.cns.fr`
6 `http://www.broad.mit.edu`
7 `http://genome.wustl.edu`
8 `http://www.jgi.doe.gov`
9 `http://www.vectorbase.org`
10 `http://silkworm.genomics.org.cn`
11 `http://www.ncbi.nlm.nih.gov`
12 `http://www.sanger.ac.uk`
13 `http://www.chgc.sh.cn/en`

Table 6: Parameters used in `stemloop-scan.pl` for the extraction of stable stem-loops similar to known miRNAs. Refer to the `miROrtho` userguide in the appendix for the `stemloop-scan.pl` manual on page 299.

PARAMETER	VALUE
–min-length	50
–max-length	130
–max-energy	-13
–max-hairpins	2
–max-hairpin-size-multi	8
–max-hairpin-loop	25
–max-multi-loop	15
–max-bulge-loop	5
–max-interior-loop	8
–min-pairs	15
–max-pairs	70
–max-continuous-stack	30
–min-ratio	0.35
–max-ratio	2.3
–T	37

3.2 EXTRACTION OF STEM-LOOP STRUCTURES

To select regions which can potentially fold into stem-loop like structures, the `RNALfold` algorithm (Hofacker et al., 2004) was used. Raw genomic sequences were split into chunks of 5 Mb with an overlap of 200 bp. This step is necessary due to memory requirements of the `RNALfold` algorithm for long sequences. The overlap guarantees, that no candidates are missed as precursor microRNAs (pre-miRNAs) are usually not much longer than 120 bp.

All folded sub-structures were examined according to criteria such as the length of the structure, the minimum free energy (MFE), and the number and sizes of terminal-, bulge-, and interior-loops. All parameters (Table 6) were chosen to retain over 95% of all known Metazoa miRNAs from `miRBase 13.0` while scanning for stable stem-loop like structures.

3.3 SUPPORT VECTOR MACHINE MODEL FOR SEQUENCE CLASSIFICATION

Once all the stem-loop candidates were gathered, an SVM model was built to assign a miRNA gene probability score. Instead of the default cut-off probability of 0.5 for a miRNA, a more restrictive threshold of 0.75 was chosen to classify stem-loops as putative miRNAs.

3.3.1 *Training data*

Training sequences were first scanned for stable stem-loop like structures using the `stemloop-scan.pl` program with the same parameters

used for the hairpin extraction from the genomes. Afterwards, all sequences were clustered on an identity threshold of 90% to prevent biasing the model from very similar sequences in the training data. Cd-hit-est (Li and Godzik, 2006) was used to cluster highly homologous sequences into cluster that meet a sequence identity threshold of 90%.

The final training data were composed of a positive miRNA set and a negative non-miRNA set. Positive training sequences were assembled from true miRNAs gathered from `miRBase 13.0` and deeply conserved candidates from a previous run of the pipeline. The negative training set was constructed from two sources. Part one came from animal noncoding RNAs (ncRNAs) downloaded from `Rfam 9.1` (Griffiths-Jones et al., 2005) that were scanned for stable stem-loop structure resembling miRNAs. miRNA, small nucleolar RNA (snoRNA), and vault RNA (vRNA)[14] sequences were excluded from this set. The later were shown to be potentially processed into functional miRNAs (Ender et al., 2008; Glazov et al., 2009; Persson et al., 2009; Saraiya and Wang, 2008; Taft et al., 2009), even though their biogenesis might differ. Part two of the negative training set was drawn from random genomic hairpins produced in the stem-loop scan step of the pipeline (Fig. 12 on page 34). Only sequences that did not show any sequence homology to known miRNAs were retained.

A random subset of 4,126 negative training samples was combined with an equal number of positive training data, summing up to 8,252 training sequences. For testing purposes some additional training data were used. Negative and positive training sets always contained a balanced number of training sequences, as class imbalance can lead to poor classification results with respect to the minority class (Weiss, 2004).

3.3.2 *Features*

Using the training data, a set of 102 feature attributes was computed for each sequence. The calculations were performed using `miRNAclassify.pl`. The same set of features was calculated for all stem-loop sequences that were produced by the hairpin extraction step. A list of 102 features is summarized in Table 7 on page 40. Absolute numbers rounded to six decimal points were used as the numerical feature vectors to train the SVM.

MINIMUM FREE FOLDING ENERGY (MFE)

The MFE and secondary structure w for a sequence can be computed using the program `RNAfold` (Hofacker et al., 1994) with the options `-d2 -noLP`. The first option assures that the free energy folding and the partition function use the same energy model, the second option excludes lonely isolated base pairs in the structure.

14 Short polymerase III transcripts with a length varying between about 80 and 150 nt. Part of vault particles — large cytoplasmic ribonucleoprotein particles.

ADJUSTED MINIMUM FREE FOLDING ENERGY (AMFE)

As longer sequences tend to have a lower MFE, the raw energy values were normalized with the respective sequence length L (Bonnet et al., 2004; Seffens and Digby, 1999; Zhang et al., 2006).

$$amfe = \frac{MFE}{L * 100} \tag{3.1}$$

MINIMUM FREE FOLDING ENERGY INDEX (MFEI)

Sequences with higher GC content have a lower MFE (Ng Kwang Loong and Mishra, 2007; Zhang et al., 2006), therefore eq. 3.2 is used to normalize for any bias due to GC content.

$$mfei = \frac{amfe}{\%GC} \tag{3.2}$$

MINIMUM FREE FOLDING ENERGY 2, 3, AND 4

Three additional indices based on the MFE were computed, based on similar features first introduced by Batuwita and Palade (2009).

$$mfei\_2 = \frac{MFE}{L * n(S)} \tag{3.3}$$

$$mfei\_3 = \frac{MFE}{L * (n(I) + n(B))} \tag{3.4}$$

$$mfei\_4 = \frac{MFE}{bp(w)} \tag{3.5}$$

where $n(X)$ denotes the number of elements for a certain topological feature (S - stem, I - interior loops, B - bulge loops). $bp(w)$ stands for the number of base pairs in the structure w which was calculated on the sequence.

SEQUENCE ENTROPY

The entropy H for a sequence was calculated using word sizes of one and two (monomers and dimers). The general form of this entropy (also named Shannon entropy after Claude Shannon[15]) is defined as follows:

$$H = -\sum_{i=1}^{n} p(i) \log_2(p(i)) \tag{3.6}$$

where $p(i)$ are either the mono- or dimer-frequencies of the respective sequence. n is four using monomers and 16 using dimers. The entropy is maximized when the probability distribution is uniform. So for a sequence with equal numbers of A, C, G, and U nucleotides we get an entropy of 2 bit, for equally distributed dimers we would expect an entropy of 4 bit.

15 Claude Elwood Shannon (1916–2001), an American electronic engineer and mathematician

STRUCTURE ENSEMBLE

In vivo, a single RNA molecule may have several sub-optimal structures. The space of possible structures is called structure ensemble, with each member having a different energy. Based on a Boltzmann distribution, the partition function Z for the ensemble w of secondary structures can be computed (McCaskill, 1990). Regarding a structure as a set of base pairs, the probability of a structure $w_\alpha \in w$ is given by

$$P(w_\alpha) = \frac{e^{-E_\alpha/RT}}{Z} \tag{3.7}$$

where

$$Z = \sum_{w_\alpha \in w} e^{-E_\alpha/RT} \tag{3.8}$$

E_α is the free energy of w_α, $R = 8.31351\,\mathrm{J mol^{-1} K^{-1}}$ is the molar gas constant, and T is the temperature (310.15 K). Based on eq. 3.7, the base-pair probability p_{ij} for any base pair is given by

$$p_{ij} = \sum_{w_\alpha \in w} P(w_\alpha)\delta_{ij}^{\alpha} \tag{3.9}$$

δ_{ij}^{α} is either 1 or 0, depending if the bases i and j form a base pair. All calculations for both ensemble energies and base pair probabilities, were done using the McCaskill's algorithm implemented in `RNAfold` (Hofacker et al., 1994).

NORMALIZED SHANNON ENTROPY OF THE STRUCTURE ENSEMBLE

Based on the base pair probabilities in the ensemble of structure, the normalized Shannon entropy dQ (Huynen et al., 1997) is defined as follows:

$$dQ = \frac{-\sum_{i<j} p_{ij} \log_2(p_{ij})}{L} \tag{3.10}$$

where p_{ij} is the base pair probability of the bases i and j in the ensemble of structures and L is the length of the sequence. This feature described the "space" of alternative structures in the ensemble.

NORMALIZED BASE PAIR DISTANCE IN THE STRUCTURE ENSEMBLE

The normalized base pair distance dD can also be calculated from the base pair probabilities:

$$dD = \frac{\sum_{i<j}(p_{ij} - p_{ij}^2)}{L} \tag{3.11}$$

It describes the average distance among the structures in the ensemble. The lower this value, the "tighter" the cluster of alternative folds, i. e. the more similar the structures are. This measure has been shown to be strongly correlated with the ensemble entropy dQ (Freyhult et al., 2005). `RNAfold` outputs this measure as "ensemble diversity".

NORMALIZED NUMBERS OF BASE PAIRS

This feature represents the number of $A \cdot U$, $G \cdot C$, or $G \cdot U$ base pairs normalized by the number of stem regions. It is defined as follows:

$$bp\_n = \frac{n(XX)}{n(AU) + n(GC) + n(GU)} * \frac{1}{n(S)} \tag{3.12}$$

where AU, GC, and GU are base pairs and $XX \in \{AU, GC, GU\}$.

Z-SCORE

The z-score (Le and Maizel, 1989) is the number of standard deviations by which the MFE of a sequence deviates from the mean MFE of a set of shuffled versions of the input sequence. As the thermodynamic model (Mathews et al., 1999; Xia et al., 1998) used for RNA folding states different energy contributions for various stacks, Workman and Krogh (1999) postulated that random sequences have to be generated with the same dinucleotide frequencies as the input sequence, in order to draw valid conclusions. For each input sequence 100 random samples were computed using the Altschul-Erikson dinucleotide shuffling algorithm implemented in `Perl` by (Ng Kwang Loong and Mishra, 2007). The z-score is defined as follows:

$$Z = \frac{E - \mu(E_{\text{shuffled}})}{\sigma(E_{\text{shuffled}})} \tag{3.13}$$

where μ is a function to compute the mean MFE of the population of shuffled sequences. σ is the standard deviation of the MFE for the shuffled sequences.

Here the p-value is defined as follows (Bonnet et al., 2004):

$$p = \frac{m}{n+1} \tag{3.14}$$

where m is the number of sequences in E_{shuffled} having a lower MFE than the original sequence. n is the number of shuffled sequences (e. g., 100). It has been shown, that the z-score is more sensitive than p-values (Freyhult et al., 2005).

The z-score feature and its corresponding p-value were not included in any final model, due to its requirement for shuffled sequences which is very CPU intense. Tests showed that the computational time computing the features including the z-score was increased by 2,300%. Nevertheless, the feature only increased the area under the curve (AUC) (Fig. 10 on page 25) by 0.20% in a 10-fold cross-validation setup.

SELF-CONTAINMENT INDEX (SC)

This method was introduced by Lee and Kim (2008) and is a measurement that represents the tendency for an RNA sequence to fold into an optimal secondary structure regardless of its surrounding sequence context. It is defined as follow:

$$sc = \frac{1}{nL} \sum_{i=1}^{n} \text{hamming\_distance}(w, w\prime) \tag{3.15}$$

where w stands for the folded RNA sequence in dot-parenthesis format. $w\prime$ is the structure of the same sequence, when folded in the context of random sequences of length L flanking the original sequence (total length = 3L). The Hamming[16] distance (Hamming, 1950) between these two structures w and $w\prime$ is calculated n times — the number of shuffled sequence contexts. Afterwards, a mean is computed which is used as the final output.

The self-containment index was used not for the final model as it increased the computational time to calculate the features by 196,000% while increasing the AUC by only 0.5%.

16 The number of atomic operations needed, to convert one string of characters into another of equal length.

Table 7: Features for classifying stem-loop sequences split in sequence-, structure-, and energy-based features. The column *Number* shows the number of sub-features. Features that were not used in any final model are marked by ○. Most of them were very CPU-intense, while contributing very little to the accuracy of the model.

FEATURE	ABBREVIATION	NUMBER	MIRORTHO
SEQUENCE BASED FEATURES			
Monomer frequencies for all four bases	`A, G, C, T`	4	•
Dimer frequencies for all 16 dinucleotides	`AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, UU`	16	•
Mono- and di-nucleotide sequence entropies (eq. 3.6 on page 37)	`ws_entropy_1, ws_entropy_2`	2	•
STRUCTURE BASED FEATURES			
Hamming distance MFE structure - centroid structure	`centroid_bp_dist`	1	•
Number of base pairs normalized by sequence length	`bp`	1	•
Shannon entropy of ensemble structures normalized by sequence length (eq. 3.10 on page 38)	`dQ`	1	•
Base pair distance of ensemble structures normalized by sequence length (eq. 3.11 on page 38)	`dD`	1	•
Mean base pair probability for pairs in the MFE structure	`mean_bp_prob_mfe`	1	•

Table 7: Features used for the SVM classifying sequences (continued)

FEATURE	ABBREVIATION	NUMBER	MIRORTHO
Fraction of base pairs in the MFE structure having a base pair probability over 0.9	`over_09`	1	•
Scan the structure with a 18 nt[a] window, and a step size of one. Extract the window for which the average base pair probability for the nucleotides is highest.	`window_bpp`	1	•
Average base pair probability for the base pairs flanking the region containing the highest base pair probabilities (see previous feature)	`window_bp_flanking`	1	•
Ratio of `window_bpp` / `window_bpp_flanking`	`window_bpp_ratio`	1	•
Number of A·U, G·C, and G·U base pairs normalized by sequence length	`AU_bp, GC_bp, GU_bp`	3	•
Fraction of A·U, G·C, and G·U base pairs normalized by number of stem regions (eq. 3.12 on page 38)	`AU_bp_n_stems, GC_bp_n_stems, GU_bp_n_stems`	3	•
Self-containment index (eq. 3.15 on page 39)	`sci`	1	○
Maximum, mean and standard deviation for the distributions of hairpin loops, bulge loops, interior loops, multi loops, and stem sizes	`max_H, mean_H, stddev_H, max_B, mean_B, stddev_B, max_I, mean_I, stddev_I, max_M, mean_M, stddev_M, max_S, mean_S, stddev_S`	15	•

[a] 18 is the mean number of base pairs in the mature region of metazoan `miRBase 14` miRNAs

Table 7: Features used for the SVM classifying sequences (continued)

FEATURE	ABBREVIATION	NUMBER	MIRORTHO
Number of bulges, interior loops, and stems of size 1, size 2, etc.	B1, B2, B3, B4, B5, I2, I3, I4, I5, I6, I7, I8, S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20	29	•
Number of terminal loops	number_hairpins	1	•
Number of base pairs in longest continuous stem region	stacks_longest	1	•
Ratio of paired versus unpaired bases	bp/unpaired	1	•
Average base pairs per stem region	avg_bp_stem	1	•
ENERGY BASED FEATURES			
Minimum free folding energy normalized by sequence length	mfe_n	1	•
Ensemble free folding energy normalized by sequence length	efe_n	1	•
Centroid free folding energy normalized by sequence length	cfe_n	1	•
Difference MFE - EFE normalized by sequence length	mfe_efe_diff_n	1	•
Difference MFE - CFE normalized by sequence length	mfe_cfe_diff_n	1	•

Table 7: Features used for the SVM classifying sequences (continued)

FEATURE	ABBREVIATION	NUMBER	MIRORTHO
Frequency of MFE structure in the ensemble	`mfe_in_ensemble`	1	●
Adjusted minimum free folding energy (eq. 3.1 on page 37)	`amfe`	1	●
Minimum free folding energy index (eq. 3.2 on page 37)	`mfei`	1	●
Minimum free folding energy index 2, 3, and 4 (eqs. 3.3 to 3.5 on page 37)	`mfei_2, mfei_3, mfei_4`	3	●
Z-score and p-value of MFE based on 100 shuffled replicates (eqs. 3.13 to 3.14 on page 39)	`z-score, p-value`	2	○

3.3.3 *F-scores*

To rank the features in order of increasing discriminative power, the F-score criterion from R.A. Fisher was used. The larger the F-score for a certain feature, the more likely it is to be more discriminative. The F-score was computed using the tool `fselect.py`[17] from the `LIBSVM` utility section.

Note that the F-score does not measure dependencies and correlations between different features.

3.3.4 *Training the SVM model*

All features were normalized to a range of $\{0, 1\}$ using `svm-scale` from the `LIBSVM` package (Chang and Lin, 2001). The final SVM for classifying genomic stem-loop sequences was created with `svm-train` using the `-b 1` option to calculate posterior probabilities from the SVM output. A C-SVC[18] SVM with a radial basis function (RBF) kernel function was used. The RBF kernel has been shown to perform well in finding optimal classification solutions in most practical situations (Keerthi and Lin, 2003). A grid search for finding optimal hyperparameter γ and C (Section 3.8.3 on page 53) led to a cost parameter of 2,048 which is likely to overfit the model for a slight performance benefit. Thus, the final model was trained on the default values.

3.3.5 *Applying the SVM model to stem-loop candidate sequences*

Features (Fig. 7 on page 40) were calculated for all genomic stem-loop sequences and a score was assigned using an SVM to classify the sequences as putative miRNA or non-miRNA gene (genomic background hairpins or other ncRNAs).

3.4 LOW-COMPLEXITY SEQUENCES

As neither the positive nor the negative input training set for the SVM classifying stem-loop sequences contained many low-complexity sequences, the model did not perform well in scoring low-complexity sequences. Therefore, the putative miRNA genes predicted by the SVM step on stem-loops produces some low-complexity predictions. To filter those ones out, the tool `fasta-dustmasker.pl` was used, which is a wrapper around `dustmasker` (part of the NCBI's `BLAST` package (Altschul et al., 1990)). Sequences with more than 5% masked nucleotides were excluded from any further analysis[19].

3.5 GENOMIC BLAST SEARCHES FOR DISTANT HOMOLOGOUS SEQUENCES

`WU-BLAST` (Gish, W. (1996–2004) `http://blast.wustl.edu`) was used to find homologs to already annotated `miRBase 14` miRNAs in all studied genomes. Parameters were adjusted for extreme sensitivity as miRNA genes are short and mostly only a sub-portion of the pre-miRNA,

17 `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/fselect/fselect.py`
18 Regularized support vector classification (standard algorithm)
19 Less than 3% of all metazoan `miRBase 14` miRNAs fall into this category.

namely the mature region, is fairly well conserved. Parameters for WU-BLAST were adapted from O'Reilly's Blast book (Korf et al., 2003) and from Ensembl[20]. The final WU-BLAST option were set as follows:

```
-M 1 -N -1 -Q 3 -R 2 -W 7 -T 7 -wordmask seg -hspsepsmax 60 \
-topcomboN 15
```

The option -M is the positive reward score for matching nucleotides, -N is the negative penalty score for mismatching nucleotides. -Q is the gap opening cost, and -R is the gap extension cost for gaps larger or equal than two. The seed word length -W for the ungapped BLAST algorithm is set to seven together with a neighborhood word score threshold of seven (-T). Thus, all target sequences should at least contain a 100% conserved miRNA *seed* region compared to the query sequence. The option -wordmask was used to filter low complexity regions in query sequences; the option -hspsepsmax determines the maximum separation allowed between alignments along the subject. As most miRNAs are highly conserved in the mature region and moderately in the miRNA* region, this option assures that the loop region between them is not longer than 60 nt. Finally the option -topcomboN extracts only the 15 best alignments for a query sequence.

The complete genomic sequence data were formatted as BLAST compatible databases — one per genome. Then, miRNAblast.pl was run on all genomes using all metazoan miRNAs from miRBase 14 as a query. All raw BLAST hits were extended on the 5′ and 3′ ends to match the input query sequence length. Then they were checked for the total length (>40) and the length of the BLAST aligned region (>20). After realigning with Muscle (Edgar, 2004), hits were checked for seed region conservation (100%), for the conservation of the mature region (>90%), and the conservation of the total precursor sequence (>50%). A candidate was rejected, if the alignment of the query and subject sequence, covering the mature region, contained more than two gaps in one of the sequences. Putative new miRNAs were folded and checked for their MFE (<−15) and if more than 30% of the precursors bases are paired. Putative candidates with a mature region containing multi-branching loops were excluded. If the total percent sequence identity of the putative new candidate and the query was less than 95% the sequence had to pass a z-score filter. The probability that one of the 100 shuffled candidate sequence had a lower MFE than the original sequence, had to be lower than 0.05. Only if all of these criteria were met, the original BLAST hit was promoted as a new putative homolog miRNA candidate. The threshold values for this procedure were determined based on all-against-all sequence comparisons of miRNAs within known families from miRBase 14.

3.6 REMOVING OVERLAPPING HITS

As the BLAST query file contained homologous miRNA sequence (e. g., dme-bantam, dps-bantam), multiple candidates from the same locus were clustered into a single representative sequence. The program miRNAcluster.pl was used to created a non-redundant homologous prediction set for all genomes by choosing the sequence with the lowest e-value as the final candidate for a specific locus.

20 Settings "finding distant homologs", from http://www.ensembl.org/Multi/blastview/

As putative miRNAs retrieved via the BLAST homology search, and sequences extracted using the SVM for classifying genomic stem-loop sequences can overlap, a non-redundant set was produced. All predictions were clustered using miRNAcluster.pl, based on their genomic coordinates. For a locus with an SVM- and a BLAST-based predicted miRNA, the pre-miRNA gene boundaries from the BLAST prediction were taken to represent this locus. This was done in order to assign genome coordinates for already annotated miRBase miRNAs in a consistent manner.

3.7 ASSIGNMENT OF ORTHOLOGOUS GROUPS

Putative miRNAs retrieved via the SVM-based and homology-based steps, were grouped into sets of orthologous groups (Fig. 13 on the next page). This was used as an additional filter, to remove false positive candidates from the previously retrieved putative miRNA genes. A strategy previously used for protein-coding genes (Kriventseva et al., 2008) was adopted to group orthologous miRNA genes. Briefly, groups were constructed from all-against-all sequence comparisons using the ParAlign algorithm (Saebø et al., 2005). This was then followed by clustering of the best reciprocal hits (BRHs) (Huynen and Bork, 1998; Tatusov et al., 1997) retrieved from the sequence comparisons. Besides requiring an alignment overlap of 20 nucleotides, an e-value cutoff of $10e^{-6}$ and $10e^{-10}$ was used for triangulating and unsupported BRH respectively. Those seed groups were extended by genes that are more similar to each other within the same genome, than to any other gene in a different species. Furthermore, all sequences within a genome sharing over 97% sequence identity were also grouped into the same orthologous group. These sequences were identified using cd-hit-est (Li and Godzik, 2006).

3.8 SUPPORT VECTOR MACHINE MODEL ON SEQUENCE ALIGNMENTS

Orthologous groups of putative miRNAs were aligned and subject to a second SVM which was trained to distinguish orthologous miRNA gene alignment from alignments of stem-loop sequences which do not show a miRNA-specific conservation pattern.

3.8.1 *Training data*

Positive training data were extracted from known miRNA gene families from miRBase 13.0. The negative training set was composed of alignments of grouped stem-loop sequences, which did not shown homology to known miRNAs retrieved from a previous run of miROrtho. Although this set might contain some true positive candidates, they are expected to be largely underrepresented, and the SVM should be robust enough to deal with noisy, mislabeled data. Sequences within one group were aligned using T-Coffee with the special parameter -mode mrcoffee. This method combines Muscle, Probcons4RNA and MAFFT and uses the secondary structures predicted by RNApfold to create an alignment based on sequence and secondary structure information.

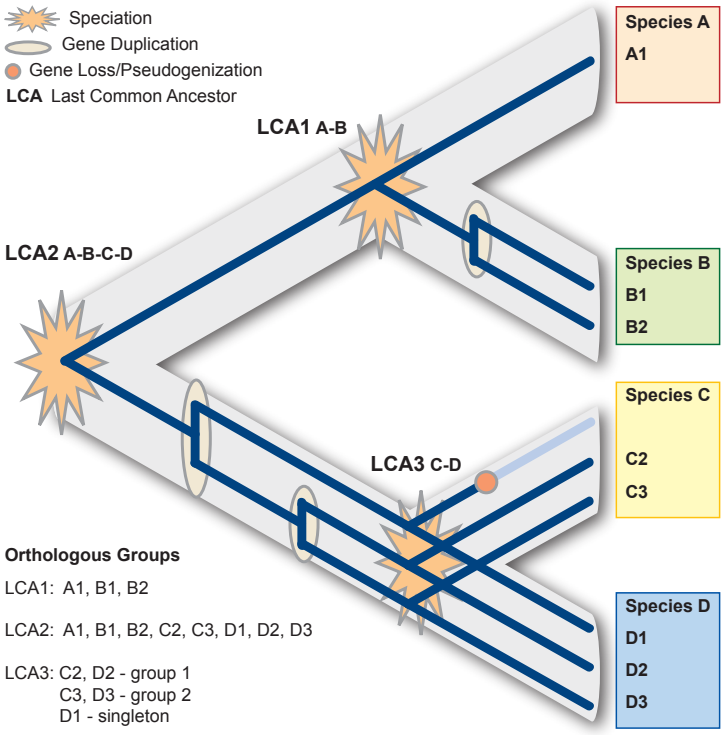


Figure 13: Orthologous and paralogous relationships (Sonnhammer and Koonin, 2002) are defined by gene duplication and speciation events. The two genes `B1` and `B2` in species B are paralogous genes as they appeared after a gene duplication event within the same species. Gene `A1` is an ortholog to gene `B1` and `B2` as they were formed by a speciation event. All genes `A1, B1-2, C2-3, D1-3` are co-orthologs to each other with respect to the last common ancestor `LCA2`. Source: Appel (2009)

3.8.2 *Features*

A set of 34 features was calculated on 792 alignments contained in the training set and used to generate an SVM model. Likewise, the same features were computed for all orthologous groups. The complete list of features is presented in Table 8 on page 50. The following sections present details on how some of the more complex features for classifying miRNA gene alignments were derived.

INFORMATION CONTENT

The information content of an alignment describes the sequence variability within it. It can be used to draw a sequence logo (Schneider and Stephens, 1990). The equation is a minor variation of eq. 3.6 on page 37 and can be written as follows:

$$I_i = \sum_{k \in A} q_{ik} \log_2 \left(\frac{q_{ik}}{p_k} \right) \tag{3.16}$$

were I_i is the information content for a specific alignment position i. $A \in \{A, C, G, U, N, -\}$ is the set of bases, q_{ik} is the fraction of base k at position i. p_k, the probability of finding a base randomly, is computed from the background mononucleotide frequencies of metazoan

miRBase 14 miRNAs. p_- and p_N, the probability of finding a gap or an unspecific nucleotide "N", are set to 1.

The alignment is partitioned in three regions: 5′, loop, and 3′ according to the superimposed consensus secondary structure. The 22 nt long region with the highest information content is put forward as the putative mature region of that miRNA group.

STRUCTURAL CONSERVATION INDEX (SCI)

The consensus structure for an alignment A can be computed using RNAalifold (Hofacker et al., 2002). The algorithm uses the same energy model implemented for single sequence folding. However compensatory mutations (e. g., a C·G pair mutating to a U·A pair) and consistent mutations (e. g., A·U mutates to G·U) give a "bonus energy" while inconsistent mutations (e. g., C·G mutating to C·A) yield a penalty energy term. The result is a consensus MFE called E_A. The consensus energy is then compared to the mean energy $\overline{E}$ of the individual sequences and the sci is calculated as follows:

$$\mathrm{sci}(A) = \frac{E_A}{\overline{E}} \tag{3.17}$$

If a consensus fold can be found, the sci is close to one. If the alignment contains many consistent and compensatory mutations, the sci can even reach values over 1.

BRANCH LENGTH SCORE (BLS)

The branch length score (BLS) measures the conservation level of a miRNA gene across several species. The score was previously successfully applied to motif discovery in flies and mammals (Stark et al., 2007b; Xie et al., 2007). The score automatically only focuses on the relevant subset of species containing a certain miRNA.

For the computation of BLS a species tree with associated branch lengths is needed. The tree was based on 18S ribosomal RNA (rRNA) sequences from all studies organisms. First, rRNA sequences for all organisms used in this work (Table 10 on page 58) were downloaded from Genbank[21] and Silva[22]. Second, the sequences were aligned using the Silva homepage. Afterwards, badly aligned regions were removed using Gblocks (Talavera and Castresana, 2007). The topology of the tree was fixed using "consensus knowledge" about the phylogenetic relationships from Ensembl[23], Flybase[24], Wormbook[25], and from the Fungal Comparative Genomics website[26]. With a fixed topology, the branch length for the final complete tree was estimated using PhyML (Guindon et al., 2005) using the following command line:

```
phyml 18S.silva.fa-gb.phy 0 i 1 0 GTR e e 16 e \
fixed_tree_unrooted.nw n y
```

The BLS can then computed by summing up all the branch lengths of the species-subtree on which the miRNA gene is conserved and dividing this value by the total branch length of the complete tree. The score ranges from $\{0, 1\}$ giving a miRNA family that is conserved over

21 http://www.ncbi.nlm.nih.gov/Genbank/
22 Silva: comprehensive ribosomal RNA database http://www.arb-silva.de/
23 http://www.ensembl.org/info/about/species.html
24 http://flybase.org/
25 http://www.wormbook.org/chapters/www_phylogrhabditids/phylorhab.html
26 http://fungal.genome.duke.edu/

the full tree a BLS of 1. miRNAs conserved in two sister species get a lower BLS, than if they were conserved in two more distantly related species as the branches connecting them is much longer.

Let us denote l_i as the individual branch lengths for the tree, n is the total number branches for the complete tree. Given a miRNA gene is conserved on a sub-part of the tree with m branches connecting this subtree, the BLS can be formally written as:

$$\mathrm{bls}(A) = \frac{\sum_{i=1}^{m} l_i}{\sum_{i=1}^{n} l_i} \tag{3.18}$$

Table 8: Features for classifying alignments. The number of sub-features are shown in the column *Number*.

FEATURE	ABBREVIATION	NUMBER
SEQUENCE BASED FEATURES		
Mean pairwise sequence identity	`mean_id`	1
GC content	`GC_content`	1
Longest region of 100% conserved sequences	`longest_100`	1
Alignment information content for the mature, 5′, loop, and 3′ hairpin region (eq. 3.16 on page 47)	`info_mature, info_five, info_loop, info_three`	4
Seed sequence conservation, conserved structure sequence conservation, and conserved structure sequence conservation normalized by mean pairwise alignment identity	`seed_cons, cs_cons, cs_cons_mean_id`	3
Excess of consistent mutations in the region of the alignment that is paired to the putative mature part [a]	`excess`	1
Information content derived features	`info_mature_div_info_total, info_loop_div_info_mature, info_ratio_arms`	3
Fraction of gap characters in the respective regions of the alignment (5′, loop, 3′, and mature)	`five_gaps, loop_gaps, three_gaps, mature_gaps`	4

Table 8: Features used for the SVM model classifying alignments (continued)

FEATURE	ABBREVIATION	NUMBER
[a] Consistent mutations miRNA - consistent mutation miRNA*		
Consistent and compensatory mutations in the putative mature region and the flanking regions of the alignment	`c_mature, c_flanking`	2
STRUCTURE BASED FEATURES		
Length conserved structure over full alignment length	`ratio_cons`	1
Structural conservation index (eq. 3.17 on page 48)	`sci`	1
Mean base pair distance between individually folded sequences and those folded using constraints from the conserved consensus structure	`mean_bp_dist`	1
Ratio paired versus unpaired alignment columns mapped on the structure	`ratio_paired`	1
Number of stems in the conserved consensus structure	`number_stems`	1
ENERGY BASED FEATURES		
Frequency of MFE structure in the ensemble of structures	`mfe_freq_ens`	1

Table 8: Features used for the SVM model classifying alignments (continued)

FEATURE	ABBREVIATION	NUMBER
Mean MFE difference between unconstrained folded and constrained folded sequences using the consensus fold, and the same measure normalized with the BLS	`mean_mfe_dist`, `normalized_mean_mfe_dist`	2
SPECIES RELATED FEATURES		
Number of sequences, number of taxa, number of taxa normalized by the BLS, number of sequences / number of taxa, maximum number of paralogs	`number_seq`, `number_taxa`, `number_taxa_bls`, `no_seq_div_no_taxa`, `max_no_paralogs`	5
Branch length score (BLS) (eq. 3.18 on page 49)	`bls`	1

3.8.3 *Training the SVM model*

Using a set of 34 features computed for every alignment of the positive and negative training set, an SVM model was derived analogous to the SVM model on single sequences (Section 3.3.4 on page 44). The model was computed including posterior probability outputs for classification. The parameter γ and C for the SVM were estimated using a heuristic search on the parameter space for maximizing the AUC on 10-fold cross-validations sets derived from the training data (See Section 2.2.1.4 on page 24).

The final SVM for classifying alignments was trained using $\gamma = 2.0$ and $C = 8.0$.

3.8.4 *Applying the SVM model to alignments*

Based on the SVM model trained on alignments, the groups produced by the orthology assignment step were classified into miRNA and non-miRNA gene alignments. The program `svm-predict` from the `LIBSVM` package was run with the `-b` option to assign a probability score to the final predictions. All predictions with score over 0.5 were manually checked for obvious false-positive predictions. Groups containing at least four member sequences were promoted as the final miRNA predictions. However, `BLAST` predictions which had an e-value lower than $10^{-}6$ and were thus likely to be true miRNA candidates, were also included in the final set, even if they did not form a group of four or more members. This was certainly the case for species-specific miRNAs such as in *C. elegans*.

3.9 ORGANIZING MICRORNA PREDICTIONS IN A DATABASE

All predicted miRNAs were integrated in a web-accessible database named `miROrtho` (`http://cegg.unige.ch/mirortho`). The main focus was put on the visualization of orthologous miRNA groups. The database is part of the `Drupal`-based[27] Computational Evolutionary Genomics Group (CEGG) website.

3.9.1 *Database content*

The `miROrtho` database presents computationally predicted miRNAs in many animal species. All data is stored in a `MySQL`[28] database. Raw FASTA and alignment files, together with color-coded alignment pictures, are saved as flat files. Alignments were calculated with `T-Coffee` using the `-mode mrcoffee` parameter. Based on the alignments a common secondary structure for each orthologous groups was derived using `RNAalifold` (Hofacker et al., 2002) with a ribosome scoring matrix, designed to align ncRNAs. All alignments and common secondary structures were color-coded (Fig. 2 on page 7) using two script from the `Vienna RNA Utilities`[29] (`coloraln.pl` and `colorrna.pl`).

27 `http://drupal.org` An open source content management system (CMS). GNU General Public License 2.

28 `http://www.mysql.com` A relational database management system (RDBMS) database. GNU General Public License 2.

29 `http://www.tbi.univie.ac.at/~ivo/RNA/utils.html`

In total the miROrtho database contains 7,887 putative miRNA genes that are homologous to annotated miRNAs from miRBase 14 and 1,437 confident predictions that do not yet have experimental support.

3.9.2 *Database web interface*

The miROrtho database is a web-resource for predicted miRNA genes and orthologous groups. For each group the following items are provided: (i) a table of annotated miRNAs together with genomic coordinates, (ii) a multiple alignment of the group displaying RNA structure conservation, (iii) the MFE common structure for that group, (iv) FASTA sequences and multiple alignment files. Furthermore, for each pre-miRNA hairpin additional information is provided, including the MFE structure and color-coded images showing the base pair probabilities and entropies on the precursor structure. The data can be accessed by querying for a specific family or a genomic location for a any species. A species tree allows for querying of all miRNA groups for the individual species. Given a FASTA sequence of interest, the database can also be searched using WU-BLAST (Gish, W. (1996–2004) http://blast.wustl.edu) to find putative miRNA genes.

3.10 PROGRAMS USED IN THIS WORK

Table 9 on the next page summarizes all programs which were used to develop the miROrtho miRNA gene prediction pipeline and the corresponding web resource. Many of the programs developed exclusively for the miROrtho pipeline have names starting with "miRNA". For more details about the usage and the corresponding manual pages, refer to the userguide in the appendix on page 269.

Table 9: List of programs and references, used for the `miROrtho` pipeline.

PROGRAM	REFERENCE
Bioperl	Stajich et al. 2002
BLAST (NCBI)	Altschul et al. 1990
cd-hit-est	Li and Godzik 2006
WU-BLAST	Gish, W. (1996–2004) `http://blast.wustl.edu`
dustmasker	part of NCBI's BLAST package
LIBSVM	Chang and Lin 2001
MAFFT	Katoh et al. 2009
MFOLD	Zuker 2003; Zuker and Stiegler 1981
microPred	Batuwita and Palade 2009
miRNAblast.pl[a]	Gerlach et al. 2009
miRNAclassify.pl[a]	
miRNAclassify.pl[a]	
miRNAorthoClassify.pl[a]	
miRNAvisualize.pl[a]	
splitter-fasta.pl[a]	
stemloop-scan.pl[a]	
Muscle	Edgar 2004
Newick Utilities	Thomas Junier, personal communication
ParAlign	Saebø et al. 2005
PhyML	Guindon et al. 2005
R	R Development Core Team 2009
RNAalifold	Hofacker et al. 2002
RNAfold	Hofacker et al. 1994
RNAmicro	Hertel and Stadler 2006
RNApfold	Hofacker 2007
RNAshapes	Steffen et al. 2006
RNAspectral	Ng Kwang Loong and Mishra 2007
RNAz	Washietl et al. 2005
ROCR	Sing et al. 2005
T-COFFEE	Notredame et al. 2000
triplet-SVM	Xue et al. 2005
VARNA	Darty et al. 2009
WEKA	Hall et al. 2009

[a] See the userguide on page 284 in the appendix for detailed manual pages.

4 RESULTS

4.1 ANNOTATION OF ANIMAL MICRORNAS

The miROrtho pipeline facilitates the systematic analysis of genome sequences to identify and define the animal "miRNAome". The method found a vast repertoire of microRNA (miRNA) genes in 45 genome sequences covering most major clades of the metazoan tree. The method implemented in miROrtho recovered over 7,887 miRNA genes which are homologous to already annotate miRNAs in other species according to miRBase 14 (Fig. 14). Furthermore, 1,437 novel predictions were annotated that do not show any homology to miRBase 14 miRNAs. The predictions were compared to the complete set of animal miRNAs deposited in miRBase 14 using sensitive WU-BLAST parameter (see Section 3.5 on page 44) and an e-value cutoff of 10.

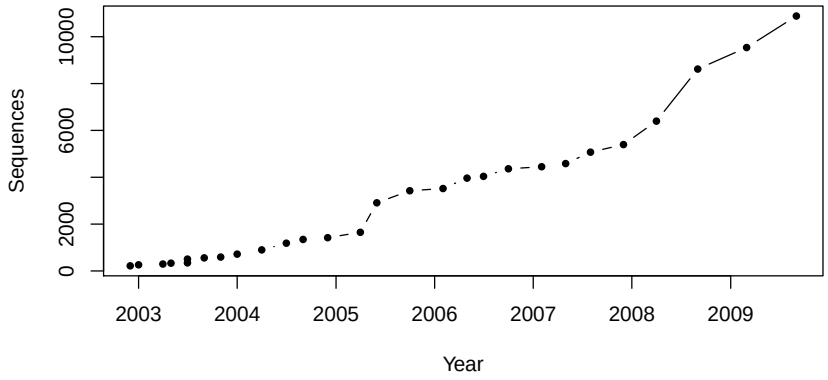


Figure 14: Growth of miRBase in number of sequences from the first release in Dec, 2002 (218 sequences) to the current release of Sept, 2009 (10,883 sequences). Number of sequences are shown for all data in miRBase including animal, plant, and virus miRNAs.

The predictions from the miROrtho pipeline were integrated in the official core of miRNA gene sets for the initial genome analysis of the cow (Elsik et al., 2009) and the parasitoid wasp *Nasonia* (Werren et al., 2010).

An overview of miRNA predictions, both novel and homologous to miRBase, is shown in Table 10 on the next page listing species, assemblies, the total amount of DNA scanned and the total number of predictions. Fig. 15 on page 59 depicts the number of miRNA predictions and maps them to a species tree. The number of predicted miRNAs either homologous or novel, is increased in the vertebrates compared to all other clades.

Table 10: Number of miRNA predictions for each genome. Novel predictions with no homologous sequences in `miRBase 14` are shown in parentheses. The species are ordered according to the species tree (Fig. 15 on the facing page).

SPECIES	ABBR.	ASSEMBLY	MIRNAS
Drosophila erecta	Dere	CAF1	152 (16)
Drosophila yakuba	Dyak	CAF1	151 (16)
Drosophila simulans	Dsim	CAF1	146 (15)
Drosophila sechellia	Dsec	CAF1	155 (16)
Drosophila melanogaster	Dmel	CAF1	168 (15)
Drosophila ananassae	Dana	CAF1	120 (12)
Drosophila pseudoobscura	Dpse	CAF1	121 (15)
Drosophila persimilis	Dper	CAF1	124 (16)
Drosophila willistoni	Dwil	CAF1	124 (12)
Drosophila virilis	Dvir	CAF1	115 (14)
Drosophila mojavensis	Dmoj	CAF1	112 (14)
Drosophila grimshawi	Dgri	CAF1	113 (13)
Anopheles gambiae	Agam	AgamP3	56 (1)
Aedes aegypti	Aaeg	AaegL1	59 (1)
Culex quinquefasciatus	Cqui	CpipJ1	56 (1)
Bombyx mori	Bmor	SW scaffold ge2k	33
Tribolium castaneum	Tcas	Tcas	38 (1)
Nasonia vitripennis	Nvit	Nvit_1.0	47 (1)
Apis mellifera	Amel	Amel_4.0	61 (1)
Pediculus humanus	Phum	PhumU1	25
Daphnia pulex	Dpul	Dappu1	19
Caenorhabditis elegans	Cele	WB170	149
Schmidtea mediterranea	Smed	WUSTL v.3.0	92
Lottia gigantea	Lgig	JGI1	12
Capitella capitella	Ccap	JGI1	14
Helobdella robusta	Hrob	JGI1	5
Tetraodon nigroviridis	Tnig	TETRAODON7	296 (8)
Takifugu rubripes	Trub	FUGU4	263 (7)
Gasterosteus aculeatus	Gacu	BROAD S1	332 (6)
Danio rerio	Drer	ZFISH6	346 (15)
Xenopus tropicalis	Xtro	JGI4.1	374 (21)
Gallus gallus	Ggal	WASHUC2	217 (49)
Anolis carolinensis	Acar	anoCar1	240 (38)
Ornithorhynchus anatinus	Oana	Oana-5.0	264 (57)
Monodelphis domestica	Mdom	monDom5	287 (82)
Canis familiaris	Cfam	CanFam 2.0	521 (138)
Bos taurus	Btau	Btar_3.1	467 (135)
Mus musculus	Mmus	NCBIM36	622 (117)

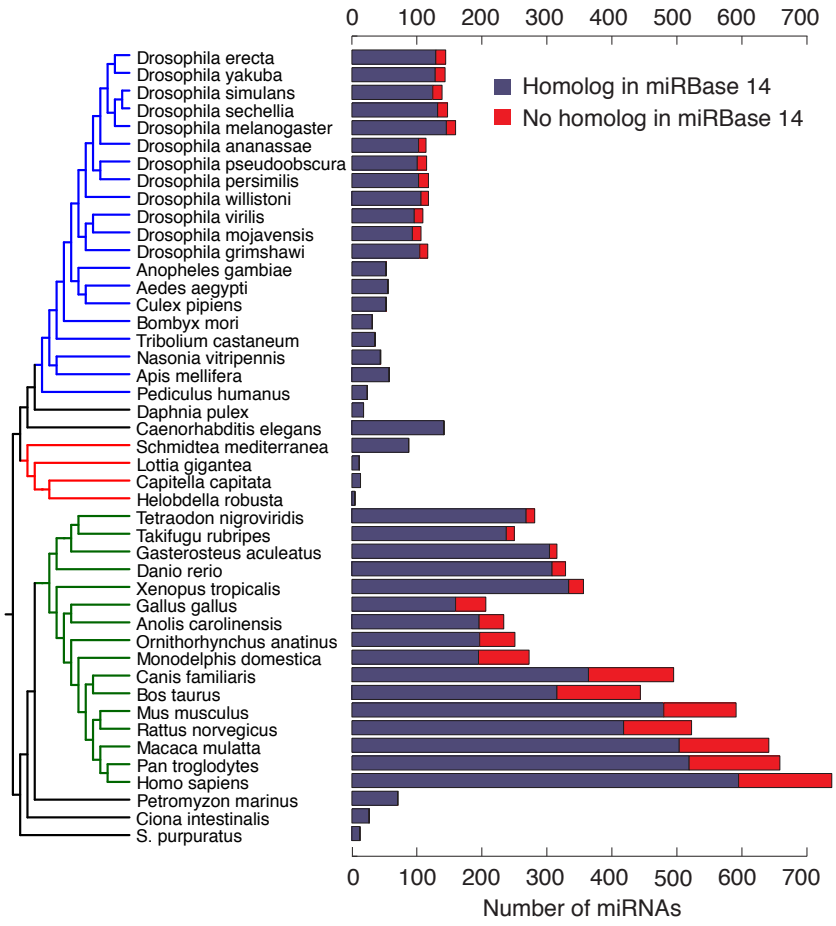


Figure 15: Homologous and non-homologous miRNA predictions as compared to `miRBase 14`. There is an expansion in the number of miRNA genes in the vertebrate lineage.

Table 10: Genomes and number of predicted miRNA genes including novel predictions with no homologous sequences in `miRBase 14` (continued)

SPECIES	ABBR.	ASSEMBLY	MIRNAS
Rattus norvegicus	Rnor	RGSC 3.4	550 (110)
Macaca mulatta	Mmul	MMUL_1	675 (145)
Pan troglodytes	Ptro	PanTro 2.1	692 (146)
Homo sapiens	Hsap	NCBI36	777 (151)
Petromyzon marinus	Pmar	Petromyzon marinus-3.0	75 (1)
Ciona intestinalis	Cint	JGI2	28
Strongylocentrotus purpuratus	Spur	Spur_v2.1	13

The complete set of predictions is accessible via the `miROrtho` web-database (Fig. 16 on the next page): `http://cegg.unige.ch/mirortho` (Gerlach et al., 2009). The database offers an interface that allow to query for any `miRBase` family or orthologous group. Using wild card characters, all miRNAs predicted for a single species can be retrieved (e. g., "Dmel%"). To find miRNA predictions specific to a chromosome location, another interface allows to specify a region and retrieve all miRNAs encoded in that region. Furthermore, any DNA sequence up to 5,000 bp can be used as a `BLAST` query, to find potential miRNAs in the sequence. All miRNAs are presented in the context of multiple alignments and consensus structures for each alignment. Alignment and consensus structure images are color-coded (Fig. 2 on page 7) to visualize the mutational pattern with regard to the conserved structure. FASTA sequence and sequence-structure based alignments can also be downloaded. Furthermore, for each individual miRNA precursor additional information is provided together with color-coded images showing the base pair probability and structure entropy mapped on the minimum free energy (MFE) structure.

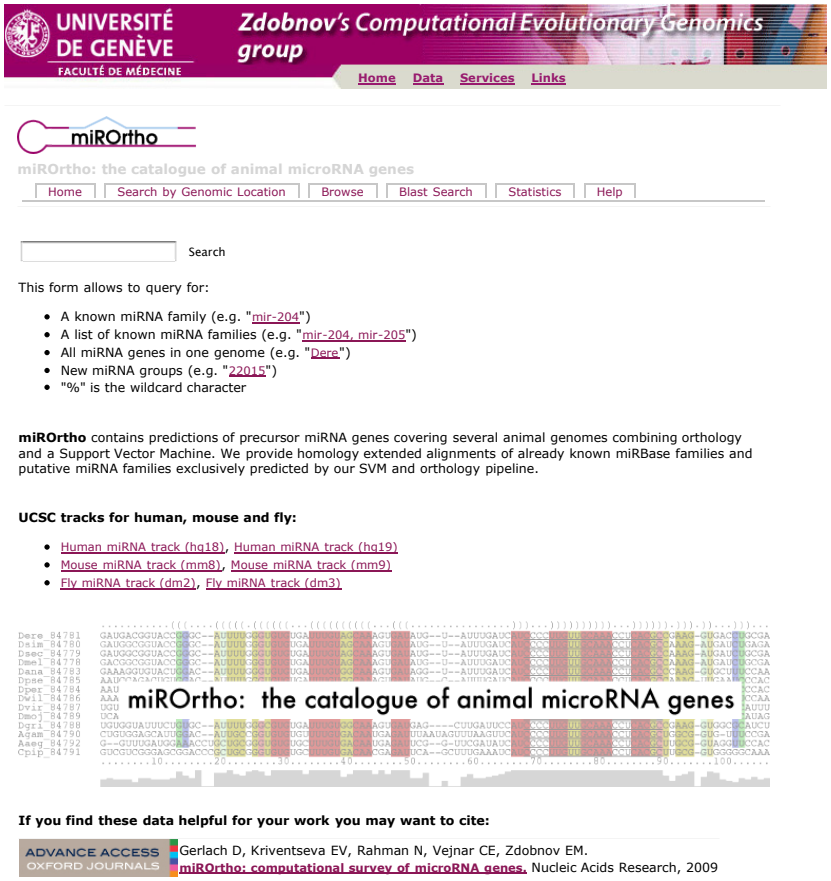


Figure 16: Screenshot of the `miROrtho` web database. The database can be searched with a known miRNA family with a genomic location for a specific species. The "Browse" function shows a species tree and allows access to all miRNA groups for a certain species. A `BLAST` interface allows querying of a sequence of interest for predicted miRNA genes. A "Statistics" and "Help" section provide further background information about how the data were produced and which genomic sequences were scanned. UCSC tracks are provided for three model organism (human, mouse, and fly).

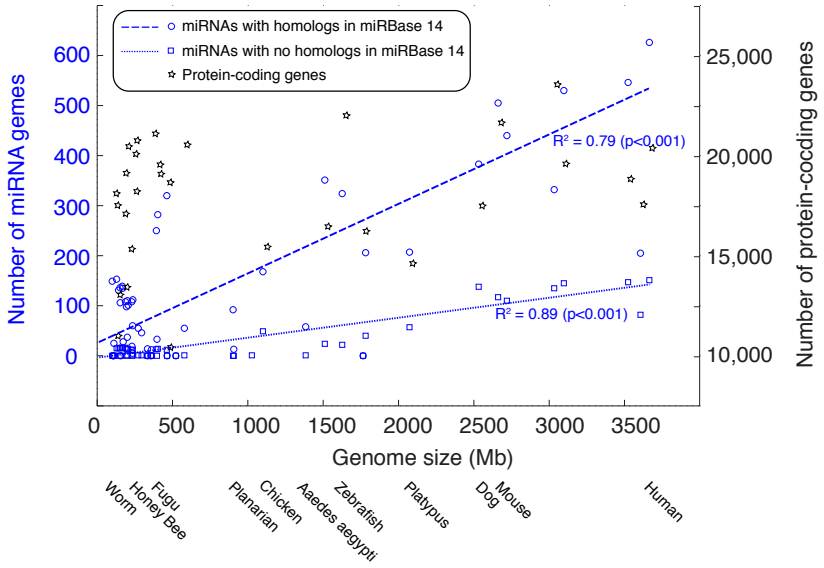


Figure 17: Correlations of gene number and genome sizes. Regression lines together with the Spearman's rank correlation coefficient (R^2) are shown. The number of protein-coding genes are plotted against the genome size (black stars). The number of miRNAs with homology to annotated miRNAs in `miRBase 14` are shown as blue circles and a blue dashed line. Those miRNAs which are exclusively predicted by `miROrtho` (blue squares, blue dotted regression line) show a similar trend as the miRNAs with homologs in `miRBase`. The is no correlation for the protein-coding genes, as opposed to the two sets of miRNA gene predictions. Note that the number of miRNAs are written on the left y-axis, the number of protein-coding genes are shown on the right y-axis. Some genome sizes are larger than expected due to alternative assemblies of some chromosomic regions which were also scanned for miRNAs.

4.1.1 *The expansion of the metazoan microRNAome*

The number of protein-coding genes does not correlate with genome size (Fig. 17). However, our initial annotations of miRNAs show a different trend (Fig. 17) for these non-coding RNAs (ncRNAs). miRNA predictions with known homologs in `miRBase 14` are positively correlated with larger genomes. The same increase in miRNAs for larger genomes holds true for the novel predictions which do not show homology to already annotated `miRBase` miRNAs. An expansion of miRNA gene families in some lineages, especially the vertebrates has already been proposed by Wheeler et al. (2009) and Hertel et al. (2006).

4.1.2 *Putative novel miRNA genes in animal genomes*

Apart from prediction of groups for which homologous candidates were already annotated in `miRBase`, the `miROrtho` pipeline efficiently predicted many novel groups. Some of these groups were experimentally verified after the predictions had been done, while other await verification but show a canonical conservation pattern and are thus likely to be true positive predictions.

4.1.2.1 *Example of an insect-specific group of microRNAs verified by 454 high-throughput sequencing*

Fig. 18 on the next page depicts an insect-specific orthologous group of miRNAs. The predictions were finished before the miRNA family was annotated in `miRBase` as the family mir-1000. The alignment shows the typical saddle-shaped conservation pattern with a 100% conserved mature part. The loop region is more diverse. Many consistent and compensatory mutations support the conserved consensus stem-loop structure. Fig. 19 on page 65 shows the *D. melanogaster* candidate of this group with aligned *454* high-throughput sequence reads. Based on the alignment pattern of the short reads to the precursor sequence, the putative mature part of the precursor microRNA (pre-miRNA) can be derived. In general the *5′* ends of the reads matching to the precursor do not show much heterogeneity as opposed to the *3′* end. Often the mature miRNA is aligned to much more reads than miRNA* region. For this miRNA group the mature part predicted from the alignment's information content, agreed 100% with the mature regions predicted by *454* high-throughput sequence reads from small RNA cloning libraries.

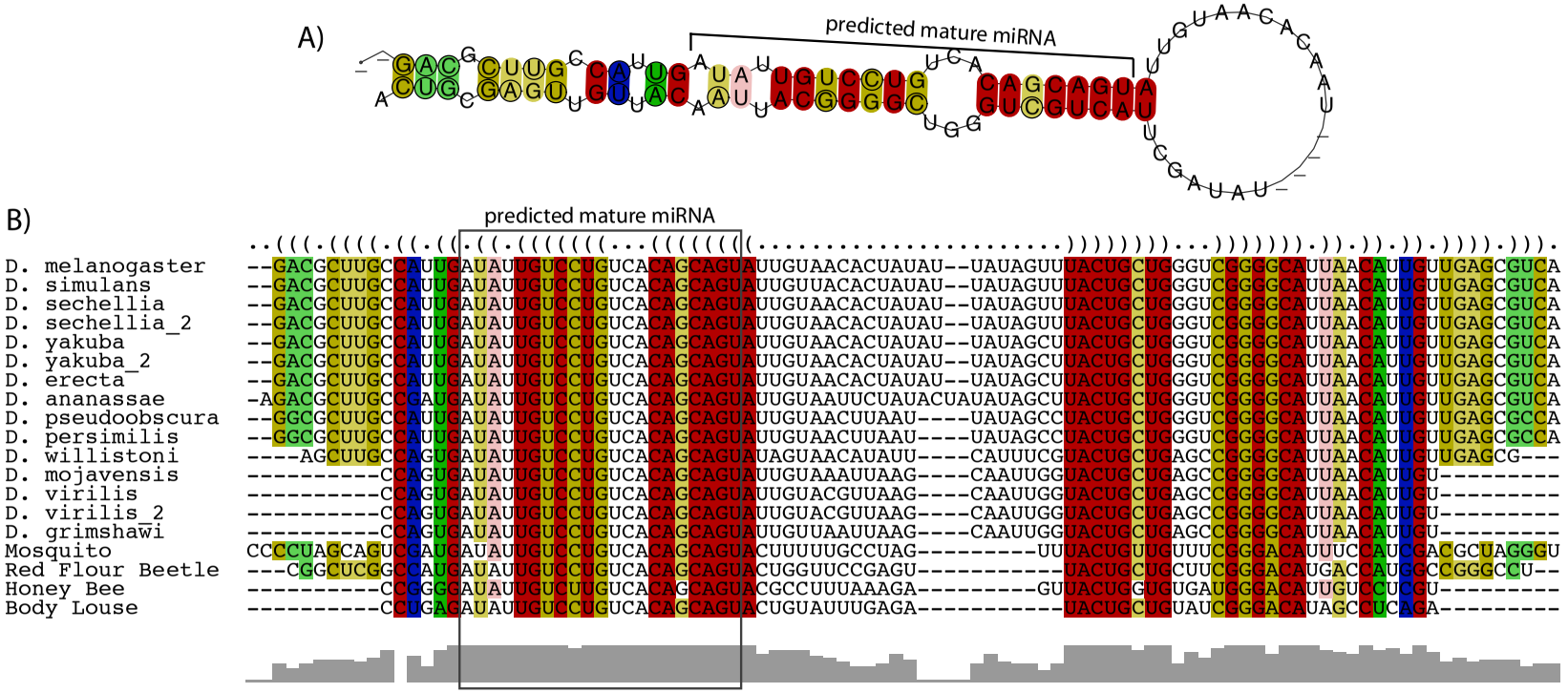


Figure 18: Orthologous group of insect-specific miRNAs. The candidates for *D. melanogaster*, *T. castaneum*, *A. gambiae*, and *A. mellifera* were recently annotated as the mir-1000 family in `miRBase`. A) Consensus secondary structure based on the alignment B) Alignment of the orthologous group. Note the almost perfect conservation of the mature region and the drop in conservation for the loop region. Loops and bulges for the consensus structure are either small or symmetrical. Many consistent and compensatory mutations support a common and conserved stem-loop structure. The color schema reflects the mutational pattern with respect to the common secondary structure. For more details refer to Fig. 2 on page 7.

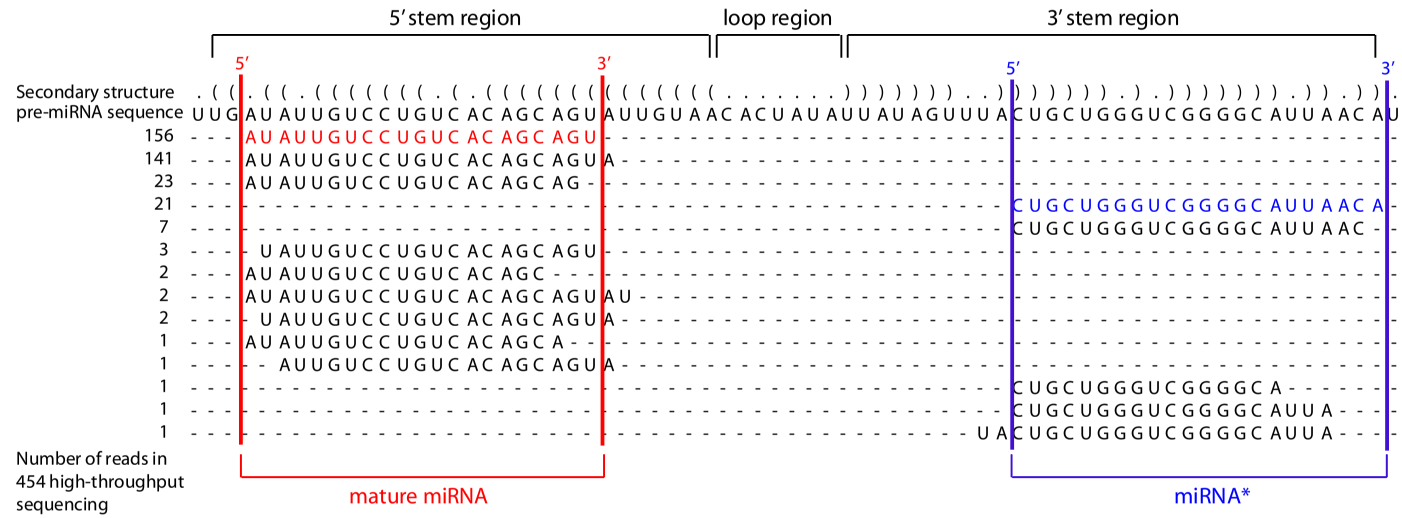


Figure 19: *D. melanogaster* mir-1000 with aligned high-throughput sequencing data reads. Small RNA sequencing data were retrieved from the GEO set GSE7448 (`http://0-www.ncbi.nlm.nih.gov.millennium.unicatt.it/geo/query/acc.cgi?acc=GSE7448`) published by Ruby et al. (2007b). Many more reads align to the 5′ region than to the 3′ region. The mature miRNA is annotated as the region with the most mapped *454* reads (red part). Note that the 5′ end of the 21 nt mature miRNA is well defined whereas the 3′ end shows more variability (Seitz et al., 2008). Much less reads align to the miRNA* region which is color-coded in blue.

4.1.2.2 *Example of a fish/frog-specific group of microRNAs verified by pyrosequencing*

Out of originally 22 novel miRNA predictions for the zebrafish genome which did not show homology to `miRBase` miRNAs, seven were verified after the `miROrtho` predictions had been finished. An example of such a predicted orthologous group for which candidates were later confirmed by pyrosequencing is depicted in Fig.20. Both candidates in *Danio rerio* and *Xenopus tropicalis* have been experimentally verified and deposited in `miRBase 14` under the name mir-2184 (Armisen et al., 2009; Soares et al., 2009).

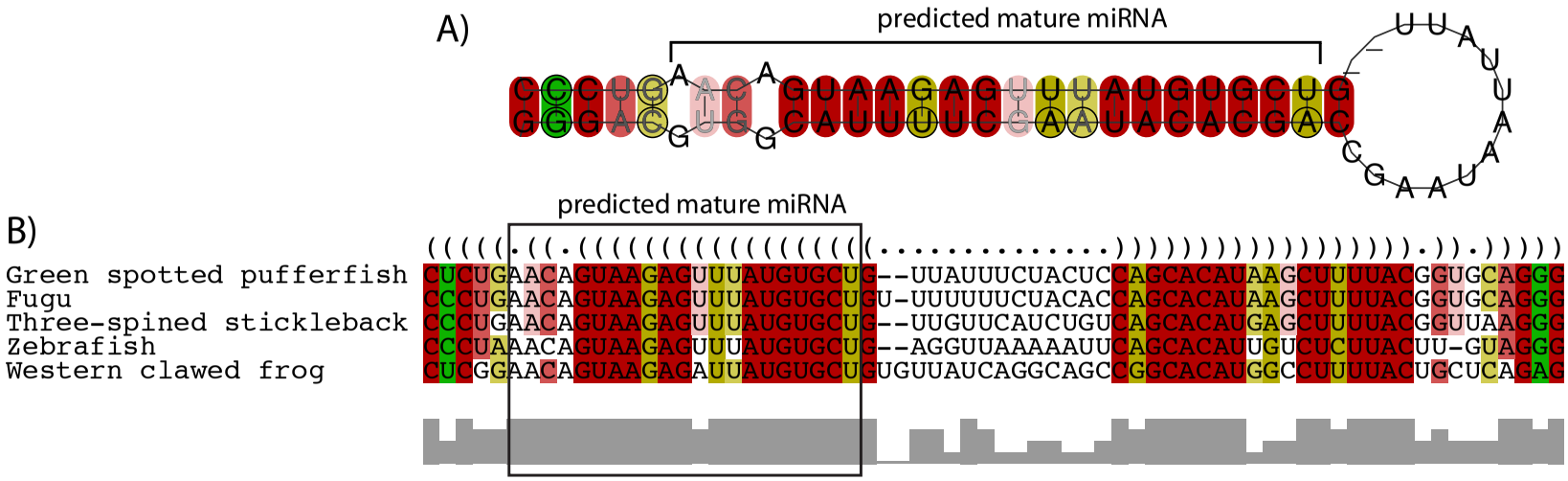


Figure 20: Orthologous group of novel fish- and frog-specific miRNAs. A) Consensus secondary structure based on the alignment B) Alignment of the orthologous group. Two miRNAs of this group were recently included in `miRBase 14`: The candidate in *D. rerio* (`miRBase` name dre-mir-2184, (Soares et al., 2009)) and *X. tropicalis* (`miRBase` name xtr-mir-2184, (Armisen et al., 2009)). Both were discovered using high-throughput sequencing techniques. The conservation and consistent / compensatory mutational pattern are strong evidence for the remaining putative predictions in fishes to also be valid miRNAs. For the color-code refer to Fig. 2 on page 7.

4.1.2.3 *Example of a novel fish-specific group of microRNAs*

Another example of a conserved, but so far unknown miRNA group, is shown in Fig. 21 on the next page. Searching currently available miRNA data (`miRBase 14`) did not find any homologous sequences of the depicted group, however many consistent and compensatory mutations, as well as the highly conserved putative mature region indicate a strong confidence for the prediction. Overall the conservation pattern of this group follows the miRNA specific saddle-shaped pattern with a fairly diverse loop region. Importantly mutations involving the putative mature part are consistent, in such a way, that only the base pair in the miRNA* part changes, while the sequence of the mature part is still conserved. The mutational pattern allows for a conserved stem-loop structure which is maintained during the evolution of the orthologous sequences.

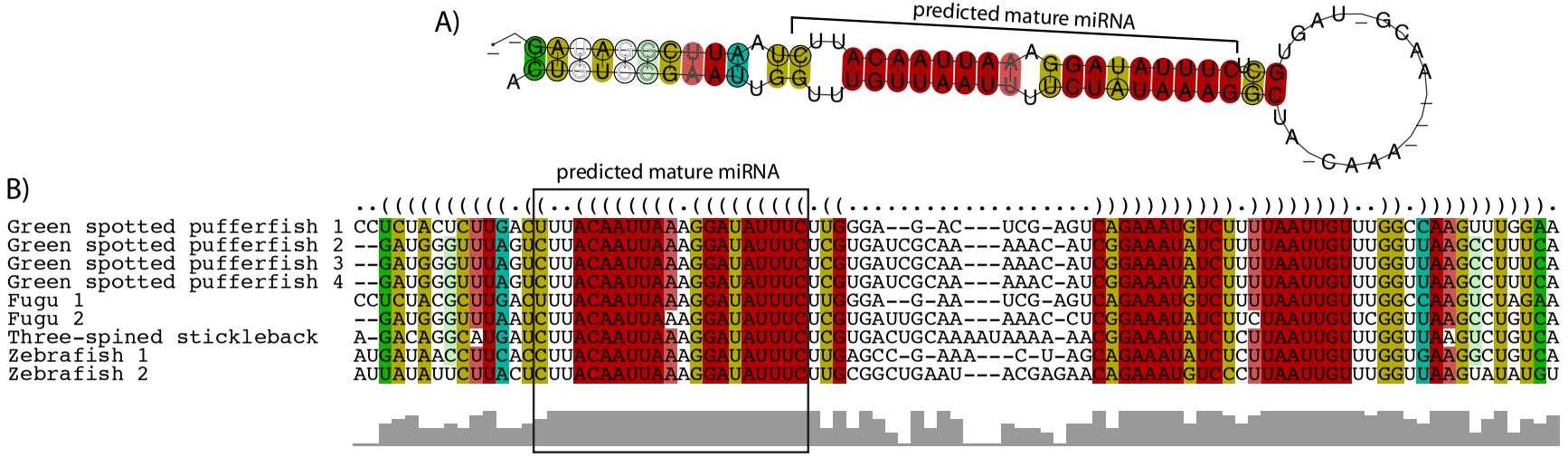


Figure 21: Orthologous group of fish-specific miRNAs. A) Consensus secondary structure based on the alignment B) Alignment of the orthologous group. None of the sequences shows sequence homology to any miRNA in `miRBase 14`. Note the well conserved putative mature region and the consistent and compensatory base changes. For the color-code refer to Fig. 2 on page 7.

4.2 EVALUATION

4.2.1 *Evaluation of training data and various classifier algorithms*

Based on a common set of features (Table 7 on page 40) different training data sets and different classifiers were extensively tested and compared. The performance of the different approaches were compared using metrics defined in eqs. 2.2 to 2.6 on page 26.

4.2.1.1 *Evaluation of different training data sets*

The choice for a specific training data set has an influence on the performance of a classifier. Table 11 on the facing page summarizes the results for different sets of training data used to train the model for stem-loop classification. Different combinations of training sequences were used. The positive training sets were composed of animal `miRBase` or `miROrtho` predictions from a previous run of the pipeline. Stable stem-loop sequences extracted from `Rfam` ncRNAs and random genomic loci were used as negative training sets. Support Vector Machines (SVMs) models were trained on different training set combinations and a 10-fold cross-validation procedure was used to asses the performance of the model on the specific training set.

In general the SVMs trained on different datasets to classify stem-loop sequences showed a comparable performance. The highest area under the curve (AUC) was achieved by a model using only `miRBase` and `Rfam` stem-loop like sequences as training data. This setup, however, is fairly unrealistic. The purpose of the classifier is to distinguish miRNA from non-miRNA stem-loops. A good classifier should be able to classify non-miRNA "genomic background stem-loops" reliably as such. The human genome e. g., encodes for about 11 million hairpins (Bentwich et al., 2005), many of which are just by chance folded into a stem-loop structure. Additionally, the genome contains also other kinds of ncRNAs which form stable stem-loop like substructures (e. g., parts of ribosomal RNAs (rRNAs)) which are however most likely underrepresented in the 11 million "background" hairpins.

Taken together, to train a model that can reliable distinguish miRNA from non-miRNA stem-loops a diverse training set is needed which represents optimally all possible situations. Therefore, the final model was trained using `miRBase` miRNAs and confident predictions from a first round of `miROrtho` as positive training samples, and stem-loops of ncRNAs as well as "background" genomic stem-loop sequences as negative training samples. This training set was thought to be the most realistic one, representing the set of stem-loop sequences that can be found in genomic sequences.

4.2.1.2 *Evaluation of different classifier algorithms*

The classification performance of four different classifiers was tested (Section 2.2.1 on page 19). Most of the tested classifiers use an independent approach and concept for classification, which makes it easier to compare the core performance of each method. Two function-based classifiers (SVM, and multilayer perceptron), a Naive Bayesian classifier, and a decision tree based approach (`J48`[1]) were tested for classi-

1 `Weka 3` uses an open-source Java implementation called `J48` of the `C4.5` decision tree learner first implemented by Quinlan (1993).

Table 11: Different training sets for SVM classifiers on stem-loop sequences. The different combinations of data sets are ordered by decreasing AUC values. `miROrtho` is a training set from a previous run of the pipeline extracting highly conserved predictions.

DATA SET					
POSITIVE	NEGATIVE	AUC[a]	MCC[b]	SE[c]	SP[d]
`miRBase`	`Rfam`	0.948	0.643	0.902	0.859
`miRBase`, `miROrtho`	`Rfam`	0.941	0.642	0.873	0.893
`miRBase`, `miROrtho`	`Rfam`, random genomic hairpins	0.931	0.619	0.877	0.867
`miRBase`, `miROrtho`	random genomic hairpins	0.928	0.599	0.845	0.867

[a] AUC: Area under the curve in the ROC, [b] MCC: Matthew's correlation coefficient, [c] SE: Sensitivity, [d] SP: Specificity (see eqs. 2.2 to 2.6 on page 26)

fying stable stem-loop sequences and gene alignments for their likelihood to encode for a miRNA. Fig. 22 on the next page shows 10-fold cross-validation ROC curves for both training sets — the one on single sequences and the one on alignments.

For both classification problems the tree-based and the Naive Bayes classifier perform worse. Both are "simple" classifiers that allow determining the "base line" of classifier performance on the respective data sets. Methods like the multilayer perceptron or the SVM based approach, outperform the simple classifiers. Notably, the depicted classifiers do not show the same performance over the whole range of false positive rates (Fig. 22 on the following page). For a low false positive rate, the SVM based classifier outperforms all other classifiers. This is particularly important given the low signal to noise ratio for miRNA predictions. Only about one stem-loop out of 10,000 candidates makes it into the final prediction set after the two SVM models and the orthology assignment step.

The performance of the four classifiers was also tested on the complete training data itself. All four tested algorithms have an AUC of less than 1 on the complete training data. An AUC of 1 on a training set, indicates an overfitted classifier that has a weak generalizing ability. So overall the models created in this work are most likely not overfitting the training data.

4.2.2 *Discriminative power of features*

All features were scored for their discriminative power using the F-score (Section 3.3.3 on page 44) criteria. Table 12 on page 73 shows some of best scoring features for the SVMs on sequences and alignments. Features that work best in distinguishing miRNA from non-miRNA genes are clearly structure and energy based features (Fig. 23 on page 73). For the alignments of putative miRNAs, the "information-based" and the "consistent mutations" features score best (Fig. 24 on page 74). In general, sequence-based features, like the GC content,

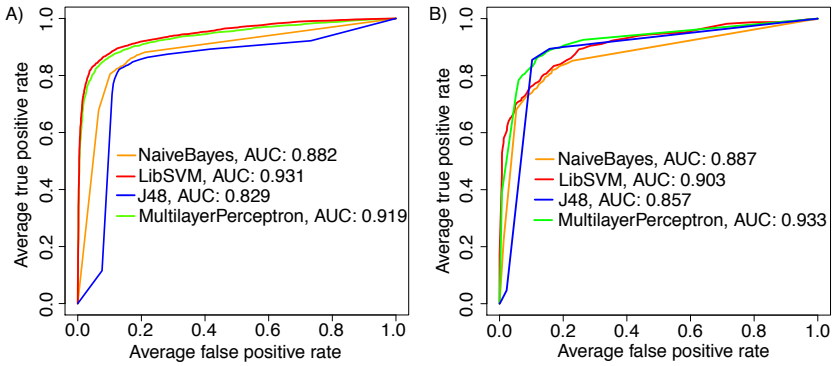


Figure 22: ROC curves comparing different classifiers for two classification tasks. A) A Naive Bayes, SVM, J48 decision tree, and a multilayer perceptron algorithm for classifying stem-loop sequences. The SVM-based, and multilayer perceptron mainly outperform the Naive Bayes, and decision tree based classifiers. B) The same set of algorithms applied on a set of alignments to classify them into miRNA and non-miRNA gene alignments.

do not capture much information to distinguish miRNA from non-miRNA stem-loops (van der Burgt et al., 2009). The better discriminative performance of structure- and energy-based features agrees with recent results from another study (Batuwita and Palade, 2009).

4.2.3 *Feature selection and feature dependency*

Mutual dependencies among features can negatively influence a classifier's performance. Although the F-score measure finds features with likely strong discriminative power, it is not able to detect correlations among features. More elaborate techniques exist that can detect highly correlated features. As some features might influence each other while training the model and could potentially decrease the performance of a classifier, a method called *feature selection* has been developed. The best subset of features should contain the least number of dimensions and contribute most to the accuracy of the classifier. Features were tested for their mutual information content using correlation. Many of the features used for the SVM classifying sequences show little or no correlations (Fig. 25 on page 74). The SVM used to classify alignments on the other hand, shows some slightly correlated features (Fig. 26 on page 75).

To test the influence of different features, subsets of a selection of features were created. Features were first ordered according to their discriminative power using the F-score criteria (Section 3.3.3 on page 44). Then the classifier performance was tested with increasing number of features starting with the most "powerful" ones at first. This procedure is called forward selection (Fig. 27 on page 76). The *sequence SVM* was tested with an increasing number of features ({3, 6, 12, 24, 49, 99}) which led to an accuracy (ACC) of {83.5, 84.4, 87.4, 88.1, 89.3, 89.5} respectively. For the *alignment SVM* results were similar. Starting with a set of features of {2, 4, 8, 17, 34} attributes, the ACC reached {82.2, 88.3, 87.9. 88.3, 89.4}. Expect for a small drop in ACC for the SVM on alignments for a set of eight features, the performance of the classifiers — measured as accuracy — can always be increased using a larger fea-

Table 12: Selection of F-scores for miRNA and non-miRNA classifying features to test their discriminative power. Some high scoring features are listed for both, the SVM on sequences and the one on alignments. In general the higher the F-score value, the better a feature can distinguish between miRNA and non-miRNA genes. For a detailed description of the features see Tables 7 to 8 on pages 40–50. Features with a ⋆ are newly introduced in this work.

MODEL	FEATURE	F-SCORE
Sequence SVM	`stacks_longest`	0.483720
	`cfe_n` ⋆	0.459219
	`amfe`	0.431537
	`bp`	0.405692
	`bp/unpaired` ⋆	0.387099
	`avg_bp_stem`	0.227649
	`mean_bp_prob_mfe` ⋆	0.225925
	`window_bpp` ⋆	0.141097
Alignment SVM	`info_loop_div_info_mature` ⋆	0.669515
	`cs_cons_mean_id` ⋆	0.391865
	`info_mature_info_total` ⋆	0.381696
	`c_flanking` ⋆	0.256956
	`mean_id`	0.214951
	`longest_100` ⋆	0.142497
	`bls` ⋆	0.133174
	`ratio_cons` ⋆	0.128577

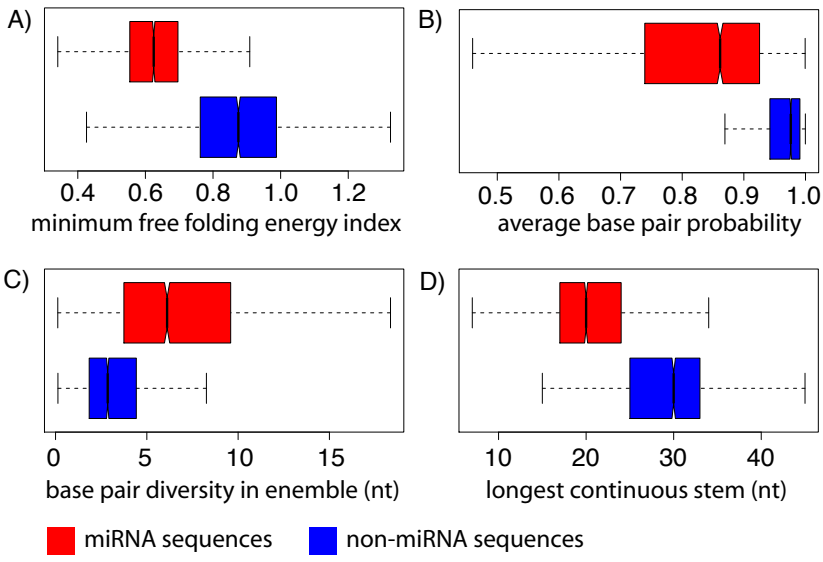


Figure 23: Features with strong discriminative power for stem-loop classification. Note the differences in the distributions for the miRNA (blue) and the non-miRNA (red) class. A) Minimum free folding energy index B) Mean base pair probability for a window of 18 nucleotides C) Mean base pair diversity in ensemble of structures D) Longest continuous stack region

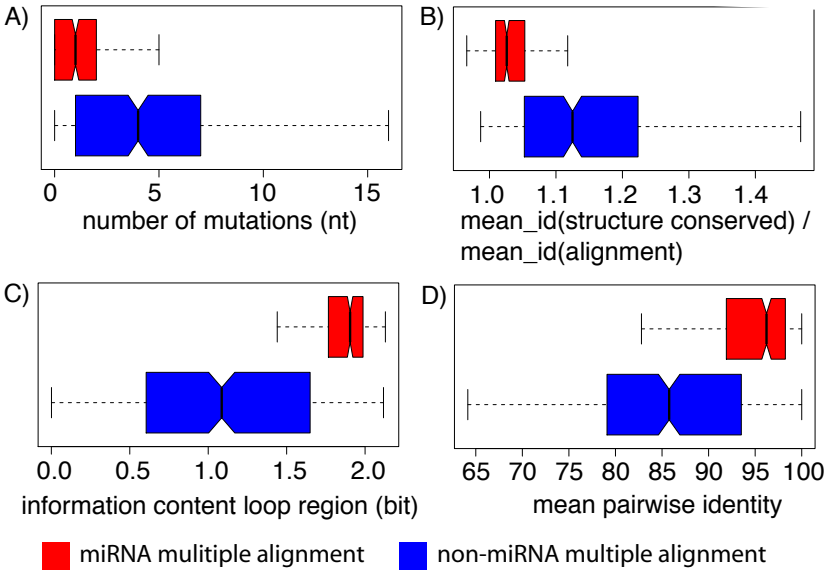


Figure 24: Features with strong discriminative power for alignment classification. Note the differences in the distributions for the miRNA (blue) and the non-miRNA (red) class. A) Structure conserving mutations in regions flanking the putative mature part B) Mean pairwise identity for the conserved structure part / mean pairwise identity for the whole alignment C) Information content loop region D) Mean pairwise sequence identity of the alignment

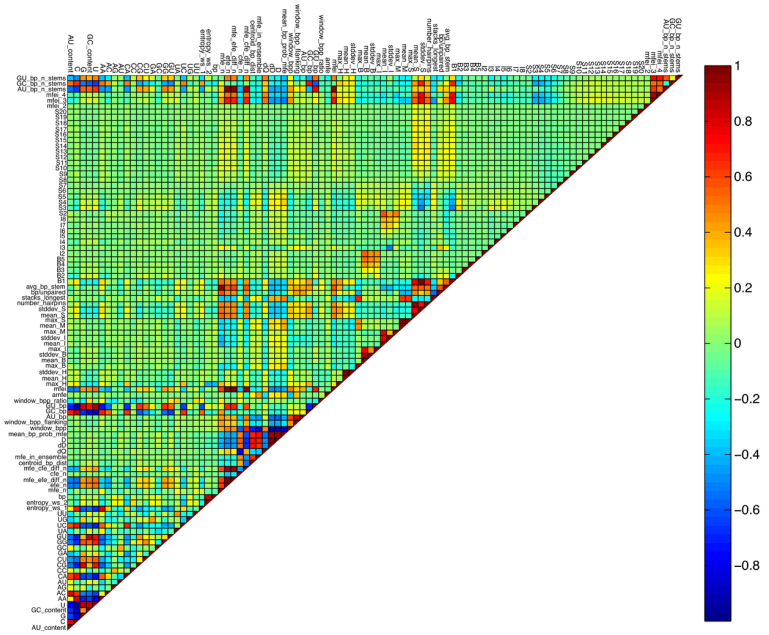


Figure 25: Feature correlations for the SVM classifying stem-loops. A correlation coefficient of 1 (dark red) indicates strongly correlated features, whereas a coefficient of −1 (dark blue) represents anti-correlated features. Example: dQ (Shannon entropy of structure ensemble) correlates strongly with dD (Base pair distance of ensemble structures) (Freyhult et al., 2005) and is anti-correlated with the feature normalized centroid fold free energy (cfe_n). The GC content feature and dQ (green square) do not show any correlation.

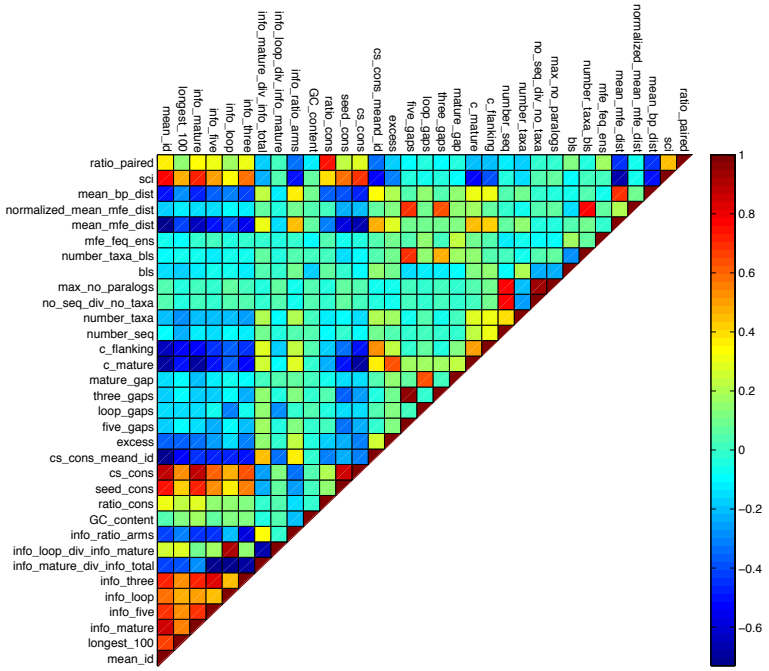


Figure 26: Feature correlations for the SVM classifying alignments. A correlation coefficient of 1 (dark red) indicates strongly correlated features, whereas a coefficient of -1 (dark blue) represents anti-correlated features. Example: ratio_paired (number of paired versus unpaired bases) is correlated with ratio_cons (alignments columns which overlap a conserved common structure), anti-correlated with mean_bp_dist (mean base pair Hamming distance between unconstrained folded and constrained folded individual sequences using the consensus structure), and behaves neutrally compared to the feature longest_100 (longest region of 100% conserved alignment columns).

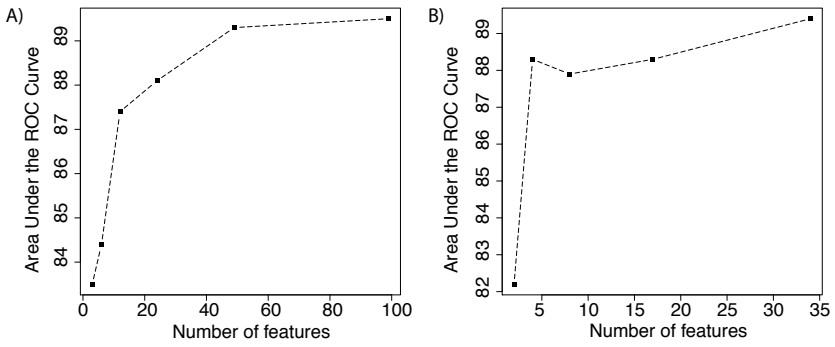


Figure 27: Feature forward selection. An increasing number of features is used to train an SVM. At each step a 10-fold cross-validation procedure is applied and the area under the curve is reported. A) Features used for the SVM to classify sequences B) Features used for the SVM to classify multiple alignments of sequences. Note there is a small drop in AUC for eight features.

tures set. Apparently the slight correlations between features (Figs. 25 to 26 on pages 74–75) can be well handled by the SVM algorithm as they do not lead to an decrease in overall accuracy.

Another approach tested only for the *alignment SVM* was a method called sequential forward floating selection, first proposed by Pudil et al. (1994). A test of various feature selection methods found that the algorithm proposed by Pudil et al. (1994) was most successful (Jain and Zongker, 1997). Briefly the algorithm dynamically changes the number of features included at each step trying to maximize a performance measure value like the AUC on 10-fold cross-validations. Of all the *alignment SVM* features, the best subset showed an AUC improvement of only 1%.

In summary, the ACC can be improved using an increasing number of features. Furthermore, the forward floating selection method on the *alignment SVM* showed that an optimal subset can only marginally increase the performance of the classifier. The total number of features (102 for the SVM classifying stem-loop sequences and 34 for the SVM classifying alignments) did not reach any computational limits and the SVM models seems to deal well with features that are mutually dependent. Therefore, all features were included in the both final SVMs.

4.2.4 *Cross-validations of the final Support Vector Machine models*

A series of 10-fold cross-validations for the SVM models based on sequence and alignment training data was performed. Table 13 on the next page lists averaged performance measures for a series of ten 10-fold cross-validation procedures.

The classifiers' performances were visualized using ROC curves, to assign the quality over a range of false positive rates (Figs. 28 to 29 on pages 77–78). Overall the SVM classifying sequences shows a better performance than the one classifying the alignments. Furthermore, the SVM for classifying alignments shows more variance between the ten series of cross-validation experiments. Both SVMs translate into regular shaped accuracy curves (Figs. 28 to 29 on pages 77–78), showing accuracy peaks at a cutoff of 0.5. This is likely due to a balanced train-

Table 13: 10-fold cross-validation performance evaluation for SVM models on sequences and alignments. The original training sets were partitioned in training and testing sets. 9/10th of the data was used to train a model, the remaining 1/10th of the data were used for testing. This procedure was repeated ten times with different random partitions. Average performance measurement values were reported as final outcomes.

MODEL	AUC	MCC	SE	SP
Sequence SVM	0.931	0.619	0.877	0.867
Alignment SVM	0.903	0.555	0.925	0.800

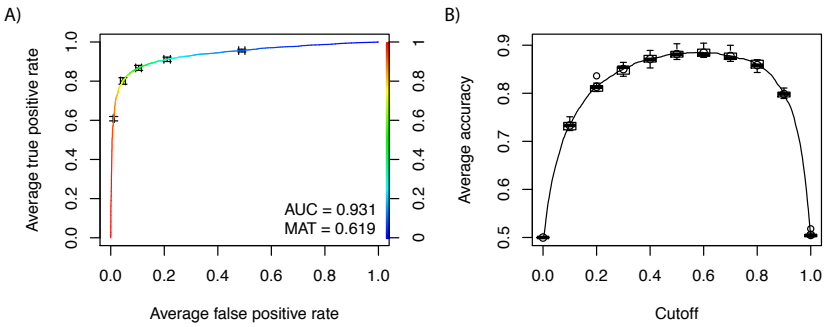


Figure 28: Performance evaluation for the SVM classifying stem-loop sequences. A) ROC curve with cutoff thresholds shown at 0.1, 0.3, 0.5, 0.7, and 0.9. An SVM probability cutoff threshold of 0.9 (red part of the curve) results in an average miRNA recall rate of 60%, with a low average false positive rate of about 1%. *AUC*: Area under the curve, *MCC*: Matthew's correlation coefficient B) Average accuracy over a range of thresholds for the SVM outputting continuous prediction outcomes.

ing data set, containing equal number of positive and negative data samples.

4.2.5 *Comparisons of other ab initio classifiers with miROrtho*

4.2.5.1 *Ab initio classifier on genomic stem-loop candidates*

The performance of the following *ab initio* miRNA gene prediction programs was tested: `miROrtho` (Gerlach et al., 2009), `triplet-SVM` (Xue et al., 2005), and `microPred` (Batuwita and Palade, 2009). Independent test data sets, which have not been used for training in any program were selected. This is important, in order to prevent any bias due an overlap of training and testing data sets. The positive training samples were taken from `miRBase 14`[2]. Two sets were chosen: One containing new[3] `miRBase 14` miRNAs which have annotated homologs in other species, and a second positive set containing new, experimentally verified miRNA, which do not show any homology to previously annotated miRNAs in any version of `miRBase`. The negative training sets were composed of random genomic hairpins and from hairpins of other ncRNA genes (`Rfam`). To assure that the random genomic hair-

2 All SVM models were trained exclusively on `miRBase 13.0` miRNAs.
3 miRNAs that were not annotated in `miRBase 13.0` or any prior version.

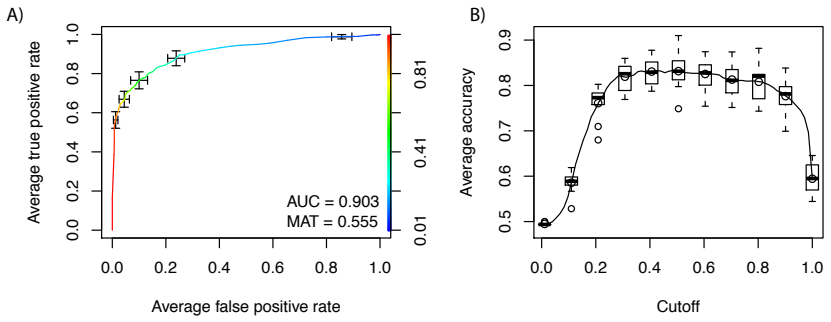


Figure 29: Performance evaluation for the SVM classifying alignments. A) ROC curve with cutoff thresholds shown at 0.1, 0.3, 0.5, 0.7, and 0.9 as in Fig. 28 on the previous page. The curve is slightly skewed to the left. Note the shift of the cutoff threshold at 0.1 (blue part of the curve) which is needed to recall over 95% of the true miRNAs genes in the test set. However with such a low cutoff score, the false positive rate also reaches 90%. Thus at this point the classifier basically makes a random guess about the class of a certain alignment. *AUC*: Area under the curve, *MCC*: Matthew's correlation coefficient B) Average accuracy over a range of thresholds for the SVM. Note the low accuracy for cutoff scores about 0.1.

pins do not contain any annotated miRNAs, all negative training data were queried against `miRBase 14` using `WU-BLAST`. Any sequence with a hit to a known miRNA was excluded from the negative testing data set. The sets are named as follows: Test set 1: `miRBase 14` miRNAs with experimental evidence and no homology plus random genomic hairpins; Test set 2: `miRBase 14` new miRNAs with homology to already known miRNAs plus hairpin sequences from other ncRNAs (`Rfam`).

The run time of the different programs tested, differs significantly, ranging from 7 seconds (`triplet-SVM`) to 22 seconds (`miROrtho`) to over 6.5 hours for the `microPred` implementation on a set of 445 sequences. The results are summarized in Table 14 on the facing page. Over all measures, set 2 shows a better classification performance for all tested programs. The miRNAs in this set most likely contain more of the known features of miRNA genes which are already annotated, and incorporated in the SVM models. `miROrtho` performs best for the ACC, the Matthew's correlation coefficient (MCC), and has a similar false discovery rate (FDR) compared to `Triplet-SVM`. The high specificity of `triplet-SVM` is due to the fact that the algorithm is quite stringent and excludes everything that does not fold into a near perfect stem-loop structure.

4.2.5.2 *Ab initio classifier on alignments*

The performance of the *ab initio* SVM classifier part of `miROrtho` based on alignments was compared with the only existing miRNA *ab initio* predictor for alignments — `RNAmicro` (Hertel and Stadler, 2006). The results of the performance evaluation are summarized in Table 15 on the next page. For this test, a random subset of 400 samples was extracted from the original 792 training alignments and used to train a new SVM model. The 392 independent remaining samples were used as testing data for the `miROrtho` *alignment SVM* classifier and the `RNAmicro` classifier. Overall the `miROrtho` alignment SVM outperforms the currently

Table 14: Performance evaluation for miRNA gene prediction programs. Note that lower values for the false discovery rate (FDR) are better. The best values for each performance measurement are marked in bold. Test set 1 contains new experimentally verified miRNAs with no homology support, test set 2 contains miRNAs with homology to previously annotated miRNAs. For details about the performance evaluations, refer to Section 4.2.5.1 on page 77.

		MIRORTHO	TRIPLET-SVM	MICROPRED
Sensitivity	Set 1	0.62	0.46	0.77
	Set 2	0.92	0.77	**0.94**
Specificity	Set 1	0.82	0.85	0.54
	Set 2	0.86	**0.90**	0.56
Accuracy	Set 1	0.73	0.68	0.64
	Set 2	**0.90**	0.81	0.82
FDR	Set 1	0.28	0.30	0.44
	Set 2	0.07	**0.06**	0.19
MCC	Set 1	0.45	0.34	0.31
	Set 2	**0.78**	0.63	0.57

Table 15: Performance evaluation for miRNA alignment *ab initio* gene prediction programs. Better values are marked in bold. Note that, as opposed to the other measures, a low FDR is better.

	MIRORTHO	RNAMICRO
Sensitivity	0.84	**0.97**
Specificity	**0.91**	0.24
Accuracy	**0.88**	0.60
FDR	**0.10**	0.44
MCC	**0.75**	0.31

existing *ab initio* classifier for ncRNA alignments. `RNAmicro`, however, is more sensitive as it excludes alignments with more than one hairpin, a priori from classifying as miRNA genes, although there are known families which deviate from this consensus feature[4].

4.2.6 *Overlap of miROrtho microRNA predictions with other studies*

In order to evaluate the capacity to detect new *D. melanogaster* miRNAs, predictions from `miROrtho` were compared with two studies that predicted miRNAs in flies (Ruby et al., 2007b; Stark et al., 2007a). None of their newly discovered miRNAs were included in the training data for the first round of `miROrtho` predictions. Stark et al. (2007a) predicted 41 new miRNAs (24 experimentally verified, 4 additional ones highly conserved). `miROrtho` recovers 75% of those predictions — importantly all 24 experimentally verified miRNAs are found. An inde-

4 See e. g.,: `http://cegg.unige.ch/mirortho/results?searchterm=mir-288`

pendent study (Ruby et al., 2007b) published 59 novel verified miRNAs of which `miROrtho` identified 83% prior to the publication.

Ruby et al. (2007a) also revealed new miRNAs in *D. melanogaster* and *C. elegans* which bypass Drosha processing and are part of a new miRNA pathway (Fig. 4 on page 17). Those miRNA genes are processed from spliced introns and were named *mirtrons*. `miROrtho` recovers two *mirtrons* out of 14 in *D. melanogaster* and one out of four in *C. elegans* Ruby et al. (2007a). Many of the *mirtrons* were missed due to their short length, thus they were already filtered out in the stem-loop extraction step. Furthermore none of the *mirtron* sequences were used of any SVM training.

4.2.7 *White-box versus black-box classifier*

SVMs or multilayer perceptrons create *black-box* models. After a classifier has been created, it is relative hard so assess why a certain test objects gets a specific score. Unlike *black-box* models, decision tree classifier e. g., build *white-box* classifiers. The classification results as well as the model itself of these classifiers can be easily translated into simple rule-based assignments. A simple set of rules (Table 16 on the next page) extracted from a J48 decision tree classifier (Section 2.2.1.2 on page 21) which was trained on the same data as the SVM, (Section 3.3.1 on page 35) allows for correct classification of 2,351 (of 4126) sequence from the positive training examples with only 23 misclassified non-miRNA stem-loops. Another set of rules (Table 16 on the facing page) allows for the correct classification of over 1,370 (of 4126) sequence from the non-miRNA stem-loops. Using these rules only 25 miRNA are misclassified as non-miRNAs.

4.3 DISCRIMINATORY FEATURES FOR MICRORNA DISCOVERY

Based on the F-score statistics on the features used for the SVMs and a decision tree to gather some insights into the classification performance finding miRNAs genes in stem-loop sequences, some highly selective features for miRNA can be extracted. The correct classification of single stem-loop sequences as miRNAs mainly depends on structure- and energy-based features. Features measuring the adjusted minimum free folding energy, computing the average base pair probability of pairs in the MFE structure compared to the "basal" stem region, and long continuous stem regions showed a good performance distinguishing miRNA from non-miRNA stem-loops. Many of the experimental verified miRNAs also contain small bulge loops or symmetrical interior-loops. Basically any structural element which leads to a bending of the stem-loop is counter-selected as is might interfere with the miRNA biogenesis pathway.

Features integrated in a previous version of the pipeline measuring specific sequence properties of the bulge and interior loops, did not show any significant differences between miRNA and non-miRNA stem-loops.

The correct classification of multiple alignments of miRNAs, mainly depends on information content-based features and a specific pattern of covariations. Furthermore, the algorithm could well integrate the information encoded by the saddle-shaped conservation profile of a typical divergent miRNA family. The loop region is very weakly conserved

Table 16: Major features distinguishing miRNA from non-miRNA genes based on a decision tree developed in this work. The set of rules used for miRNAs classified over 57% of the positive training samples correctly while only producing 23 misclassifications. The filter set for classifying the non-miRNA genes can classify about 1/3 of the negative data correctly with only a handful false negative predictions. For a detailed explanation of the features listed, see Table 7 on page 40. Each feature has an associated numerical value. All sequences that have feature values in the range of the values indicated in the most right column are classified as miRNA respectively non-miRNA genes.

	FEATURES	VALUE
miRNA	minimum free folding energy index (mfei)	> 0.74
	maximum size of multi-loop (`max_M`)	$= 0$
	minimum free folding energy index 4 (`mfei_4`)	> 0.8
	average base pair probability of pairs in the stem region flanking the putative mature part (`windows_bpp_flanking`)	> 0.25
	number of continuously paired base pairs in the stem (`stacks_longest`)	> 24
	normalized centroid free folding energy (`cfe_n`)	> 0.89
	dimer sequence entropy (`entropy_ws_2`)	> 3.52
	average base pair probability for the putative mature region (`window_bpp`)	> 0.89
non-miRNA	minimum free folding energy index (`mfei`)	$\leqslant 0.74$
	number of continuously paired base pairs in the stem (`stacks_longest`)	$\leqslant 25$
	minimum free folding energy index 2 (`mfei_2`)	$\leqslant 0.05$
	number of bulge-loops with size 2 (`B2`)	$\leqslant 1$
	GU dimer frequency (`GU`)	$\leqslant 7.5$
	mean size of interior loops (`mean_I`)	> 3.25
	AA dimer frequency (`AA`)	> 1.61

as it does not contribute to a specific function. The bases of the stem overlapping with the mature part of the miRNA are more selected to be conserved, as they are crucial for the correct function of the mature miRNA. Based on the miRNA gene biogenesis pathway and the recognition of Drosha and Exp-5, a stem-loop structure "basal" to the mature part is required. As this regions is not selecting to maintain a specific primary sequences, consistent and compensatory mutations can accumulate in homologous sequences. Thus, a conserved sequence part for the mature region, a diverse sequence for the loop region, a stable consensus stem-loop structure with structure preserving mutations, and a high overall base pair probability are highly specific for alignments of homologous miRNA as opposed to other ncRNAs.

5 DISCUSSION

Non-coding RNA (ncRNA) genes produce functional RNAs molecules rather than being translated into proteins. They perform a wide range of functions, including the control of chromosome dynamics, splicing, RNA editing, translation inhibition and mRNA destruction. ncRNAs also play a vital role in gene regulation of eurkaryotes. Apart from some well known ncRNAs such as transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), or small nucleolar RNAs (snoRNAs), more and more other functional RNAs are discovered. The ncRNA 7SL e. g., is a core component of the signal recognition particle (SRP), a ribonucleoprotein complex essential for transportation of nascent proteins containing signal peptides to the endoplasmic reticulum membrane (Doudna and Batey, 2004; Nagai et al., 2003). In 1986 vault complexes, the so far largest ribonucleoprotein complexes were described. They all contain a vault RNA (vRNA) ranging from 86 to 141 nt. The function of this complex is not yet fully understood, although it has been connected to multi-drug resistance (van Zon et al., 2003).

Another example for a ncRNA is the 17 kb human XIST (X(inactive)-specific transcript) RNA which plays a key role in dosage compensation and X-chromosome inactivation (Avner and Heard, 2001; Duret et al., 2006). Several other long ncRNAs have been found in imprinted regions of vertebrate chromosomes, including the IPW (imprinted in Prader-Willi syndrome) and the H19 (imprinted maternally expressed untranslated mRNA) transcripts (Brannan and Bartolomei, 1999; Tilghman, 1999). Recently it was shown that the imprinted H19 ncRNA is a primary microRNA precursor, after years of speculation about its function (Cai and Cullen, 2007).

Indeed the past few years have seen an explosion in the discovery of new classes of other, mainly very short ncRNAs such as microRNAs (miRNAs), endogenous small interfering RNAs (siRNAs), and Piwi-interacting RNAs (piRNAs) involved in germline transposon silencing (reviewed in Carthew and Sontheimer, 2009; Klattenhoff and Theurkauf, 2008). Some functions of those new ncRNAs are not yet fully understood while an increasing number of new members are reported. The current version of `miRBase` e. g., already lists over 10,000 annotated miRNA genes in diverse species such as animals, plants, and viruses (Fig. 14 on page 57). So far it is not clear if animals and plant miRNAs are connected or if they emerged independently in the two lineages. To date, no miRNA gene has been found to be conserved within the two phyla. Furthermore, miRNAs have not yet been reported in fungi.

Many ncRNAs described to date are deposited in a large database called `Rfam` (Griffiths-Jones et al., 2005) which contains over 1,372 families in its current version 9.1. However, many other databases exist which are specialized in certain classes of ncRNAs: miRNAs, `miRBase`[1] (Griffiths-Jones et al., 2008); Orthologous miRNAs: `miROrtho`[2] (Ger-

1 `http://www.mirbase.org/`
2 `http://cegg.unige.ch/mirortho`

lach et al., 2009); tRNAs: GtRNAdb[3] (Chan and Lowe, 2009); snoRNAs: snoRNABase[4] (Lestrade and Weber, 2006); piRNAs: piRNABank[5] (Sai Lakshmi and Agrawal, 2008).

This work presents a new pipeline for detecting miRNAs in animal genomes together with the development of a comprehensive database of such molecules. Computational methods for miRNA discovery are important as experimental approaches are laborious and can miss low-expressed candidates. However, there are some issues to be considered: The hallmark hairpin structure of a precursor microRNA (pre-miRNA) e. g., often used to first filter loci, is not limited to miRNA genes. Hairpin-like secondary structures are common motifs in other types of ncRNAs and can also be found in messenger RNAs (mRNAs) (Griffiths-Jones et al., 2005). State-of-the art machine learning techniques can be used to train models on a known miRNA genes with expert knowledge about their characteristics, like the already mentioned hairpin structure. Additionally features like the minimum free energy (MFE) have been shown to be lower for miRNA genes than for random sequences (Bonnet et al., 2004). However, our knowledge about the miRNA specific features distinguishing them from other ncRNA hairpins or random stem-loops is still limited. Nevertheless, *ab initio* miRNA predictors can perform reasonably well, while coupled with comparative genomic techniques. The following sections give some insights into the current problems of miRNA gene predictors and possible future improvements, to which this work hopefully contributes.

5.1 OPEN PROBLEMS IN MIRNA GENE PREDICTION

5.1.1 *Extraction of stable stem-loop sequences*

The first step in any miRNA gene prediction algorithm involves a stem-loop extraction procedure from genomic sequences. To retain as many of the putative candidates as possible metazoan miRNAs from miRBase were scored for sizes of loops, MFE, and other features. Based on this characteristics, a set of thresholds was derived that were used to filter stem-loop like sequences from all possible substructure within a moving window of 120 nt. About 95% of the known miRNAs were retained using this method, however a primary scan in *D. melanogaster* found over 1.3 million candidates. Nevertheless, in spite of a quiet unrestricted scan for stem-loop sequences, the two Support Vector Machines (SVMs) and the orthology filter were able to exclude most of the stem-loops as non-miRNA genes.

5.1.2 *Lack of training data*

Machine learning algorithms for miRNA gene prediction require a good set of training data. Many methods use a two-class training set containing real miRNAs from miRBase and negative training samples either from non-spliced exonic stem-loops or other ncRNAs that fold into stem-loops. All these training sets have some downsides to consider. First, the positive training samples are limited in number and

3 http://gtrnadb.ucsc.edu
4 http://www-snorna.biotoul.fr/
5 http://pirnabank.ibab.ac.in/

might not capture the complete spectrum of what is currently considered as being a miRNA gene. One of the very first machine learning based miRNA gene prediction approaches (Xue et al., 2005) was trained with little less than 200 human miRNA genes known at that time. The issues related to limited positive training data might be better addressed in the near future, as more and more miRNA genes are annotated.

Second, the training set should be drawn from a broadly sampled set of species. Recent studies have shown distinct properties of pre-miRNAs in Porifera or Cnidaria (Grimson et al., 2008; Wheeler et al., 2009). In Eumetazoa the mature miRNA is usually a few nucleotides from the terminal loop (Kim, 2005) whereas in the poriferan species *Amphimedon queenslandica* it was shown to be mostly about 30 nt away. Furthermore, poriferan pre-miRNAs tend to be longer as usual metazoan pre-miRNAs. Pre-miRNAs from cnidarians however, tend to be even smaller than classical metazoan miRNAs. Some of the miRNAs encoded in the cnidarian *Nematostella vectensis* (starlet sea anemone) have been found via the first SVM of the `miROrtho` pipeline (Fig. 11 on page 33), however due to the lack of orthologous sequences in the other species used in this work, the orthology filter discarded those sequences. This emphasizes the need for a diverse set of pre-miRNA training sequences and a large set o diverse species to scan for miRNA genes.

Regarding the negative training set, two different sources are often used. Either non-spliced exonic stem-loops (Xue et al., 2005) or other hairpin-like ncRNA sequences (Sewer et al., 2005) have been exploited. A good learning algorithm should be able to distinguish pre-miRNA stem-loops from non-pre-miRNA stem-loops and incorporate as much negative training data as possible. In most cases,this is hindered by the limited amount of positive training data as the two classes should be equally represented for good training results. A recent algorithm addresses the class imbalance learning problem by synthetic minority[6] oversampling using the synthetic minority over-sampling technique (SMOTE) technique (Chawla et al., 2002).

Another aspect regarding the training data is mislabeling. Some of the miRNA training sequences might not actually be true miRNA genes (e. g., sme-mir-749, sme-mir-751, and sme-mir-753 in *Schmidtea mediterranea* (Lu et al., 2009)) or even just misannotated fragments of other genomic transcripts (e. g., hsa-mir-1300, fragment of the EEF1A mRNA (Griffiths-Jones et al., 2008)). Also some of the other ncRNA genes in the negative training set contain stem-loop regions that might actually "dual-code". It has been shown that a human snoRNA called ACA45 can be processed into small RNAs, which can down regulate target genes (Ender et al., 2008). Although mislabeling can effect the performance of a classifer, SVMs have been shown to deal well with noisy and mislabeled data.

5.1.3 *Lack of reference data for comparing existing methods*

The lack of good training data covering the whole space of positive and negative samples for training classifiers leads to another issue: The lack of good testing data. To compare the performance of different miRNA gene finder, a test data set is needed that does not include any

6 The minority class are the positive training samples.

samples from the training data sets. This is crucial to prevent any bias in the results. The positive testing samples can be extracted from new miRNAs which were not used for any training. This becomes however more difficult for the non-miRNA negative testing stem-loop as such ones a hard to define. A good approximation can be the use of random genomic stem-loops showing similar characteristics to known miRNA genes. As the amount of random genomic stem-loops is overwhelming, chances of mislabeled real miRNA genes in the negative testing set are quite low. Another problem related to testing stem-loop is that some algorithms (e. g., Hertel and Stadler (2006); Xue et al. (2005)) exclude multi-branching stem-loops a priori, although some known miRNAs like hsa-let-7e fold into such structures.

Another problem comparing different classifiers are the different training data used for each of them. In order to exclude any bias from the underlying training samples[7], the implementations should be compared based on the same set of training data. This is however not always possible to due time constraints and limited access to the source code for some of the published methods for miRNA gene prediction (Table 5 on page 32).

5.2 IMPROVEMENTS OVER EXISTING METHODS

The performance of an *ab initio* miRNA gene finder depends mainly on the training data and the set of features used to described the putative candidates. Most of the current methods only include human pre-miRNAs as training data (Batuwita and Palade, 2009; Sewer et al., 2005; Xue et al., 2005). This somehow limits the general use of such a classifier for a broad set of diverse animal species like in this work. Based on a previous run of the `miROrtho` pipeline, a second run was performed including highly conserved predictions as a new training set. Including putative predictions from a first run can boost the prediction of miRNAs in species which had a low positive training sample coverage in the first run.

Regarding the set of features, `miROrtho` uses many of the features which have already been shown to have strong discriminative power for miRNA discovery (Ng Kwang Loong and Mishra, 2007; Zhang et al., 2006). Additionally many new feature were incorporated that mostly show strong performance in distinguishing miRNA from non-miRNA genes (Table 12 on page 73).

Energy- and structure-based features which were used for the SVM classifying stem-loop sequences had the highest predictive power. In fact the new feature measuring the energy of the centroid structure of a sequence has proven to be even more efficient than the adjusted minimum free folding energy index which is widely used (Bonnet et al., 2004; Seffens and Digby, 1999; Zhang et al., 2006). The centroid fold is a structure containing only base pairs with a base pair probability over 0.5. It is supposed to be a more accurate approximation of the real structure than the MFE structure, as it is based on the whole structure ensemble. Therefore it is less prone to local minimum in the folding space. Another newly introduced feature which measures the average base pair probability of a sub-part of the structure shows a strong discriminative power for classifying miRNA and non-miRNA stem-loop

7 Obviously more recent approaches can profit from more known miRNA genes to be used as training data.

sequences. This is probably due to a evolutionary selection of stable stem-loop structures, which show stable base pairs and that are robust to mutations in the surrounding sequence. Sequence-based features like e. g., dimer frequencies were not very selective for distinguishing miRNA and non-miRNA stem-loops. Another feature implemented in an earlier version of the `miROrtho` pipeline measuring sequence properties mapped to basic sub-structures such as e. g., bulge- and interior-loops have been evaluated to not be distinct between miRNA and non-miRNA genes.

Many features and combinations implemented in the SVM model for classifying alignments were newly developed in this work. A feature that was most selective for miRNA versus non-miRNA gene alignments was the measurement of the information content of the consensus structure loop region against the information content of the stem region. As miRNA alignments are known to adapt a saddle-shaped conservation profile, this feature could well capture the differences in conservation between the two regions. This feature worked best for alignments of sequences which were sufficiently diverse. Another newly presented feature with strong discriminative power was the measurement of the consistent and compensatory mutations in the region flanking the putative mature region of the alignment. This agrees well with the biogenesis of miRNAs, as the mature region is selected for a high sequence conservation, whereas the flanking regions is under selection to preserve a base-paired secondary structure stem region which is important for correct recognition by the Drosha enzyme.

Another improvement of `miROrtho` is the orthology filter. Compared to methods relying on whole genome alignments (Hertel and Stadler, 2006) the `miROrtho` pipeline bypasses this error-prone alignment procedure for weakly conserved regions. Furthermore, such alignments are difficult to obtain for very distantly related species as used in this work. Actually it is already problematic aligning the human with the chicken genome.

Importantly in the final set of prediction we did not exclude — as opposed to other studies — regions overlapping proteins as some known miRNAs share sequences with protein-coding exons (Das, 2009; Kim et al., 2009).

None of the existing methods performs equally well regarding sensitivity and specificity. Some gene finders are able to recall all known miRNAs with a high sensitivity (e. g., `MiRFinder`, sensitivity 99.6%, specificity 70% (Huang et al., 2007)), while others produce few false positive predictions (e. g., `miPred`, sensitivity 84.55%, specificity 97.97% (Ng and Mishra, 2007)). `miROrtho` on the other hand is pretty balanced regarding the sensitivity and specificity. Nevertheless, by adjusting the SVM probability cutoff value, the specificity can be increased. This is certainly of interest when a limited number of top scoring predictions is selected for an experimental verification step.

On an independent test sets of new miRNA genes, `miROrtho` shows competitive performance (Table 14 on page 79) this is certainly the case for the classification of non-coding RNA alignments (Table 15 on page 79).

Taken together `miROrtho` presents an valuable resource for predicting miRNA genes in diverse animal taxa. The miRNA gene predictions produced by the `miROrtho` pipeline have been included in many genome sequencing projects (Elsik et al., 2009; Werren et al., 2010).

5.3 PREDICTION AND VERIFICATION OF MICRORNAS FOR THE RESPECTIVE GENOME PAPERS

5.3.1 *microRNAs in the bovine genome*

The predictions of the `miROrtho` pipeline were integrated in the official miRNA gene set for the taurine cattle *Bos taurus*. 361 miRNA which showed homology to already annotated miRNA from other species, and 135 putative novel miRNAs add up to a total of 496 potential cow miRNAs. About half of the miRNAs occur in 60 genomic miRNA clusters, containing two to seven miRNA genes separated by less than 10 kb. For more details see Elsik et al. (2009).

5.3.2 *microRNAs in the* Nasonia *genome*

The miRNA prediction set for the *Nasonia* genome (Werren et al., 2010) consisted of two computational efforts (including the `miROrtho` predictions) and a high-throughput sequencing based approach. Our computational method identified 52 miRNAs with homology support and 2 novel miRNAs. A total of 30 miRNA from our 54 computational predictions are supported by 454 reads from a high throughput sequencing effort. A combination of all computational and expermintal predictions results in 120 unique miRNA loci for the genome of *Nasonia vitripennis*. For more details see Werren et al. (2010).

5.3.3 *microRNAs in the body louse genome*

We predicted a total of 57 miRNA gene in the genome of the human body louse *Pediculus humanus humanus*. Homologs of the miRNA family mir-iab-4 and mir-46 were not found in the genome assembly but in the trace files. Four miRNA families (mir-315, mir-283, mir-33, and mir-29) which are found in the pancrustacean clade seem to be lost in the body louse lineage. For more details see "Genome Sequences of the Human Body Louse and its Primary Endosymbiont: Insights into the Permanent Parasitic Lifestyle" (Submitted).

5.4 OUTLOOK

Compared to algorithms developed for protein-coding gene prediction, ncRNA gene predictors and especially those tailored for miRNA genes have been proposed quite recently. Although many approach have been published since the first studies (Grad et al., 2003; Lai et al., 2003; Lim et al., 2003b), all of them either have a high sensitivity or a high specificity. Approaches based on homology work relatively well in recovering new homologous miRNA genes, based on a set of already annotated genes. This results in a high specificity and produce very few false positive predictions. The ability to predict new miRNA genes is however limited by the requirement of homology to known miRNA genes. Nevertheless, the few predicted candidates can be easily evaluated in an experimental setup.

Ab initio methods were developed to overcome the limitation of only finding homologs of know miRNA genes. In general those methods are supposed to work well in predicted new and also species specific

miRNAs. In practice, however, they produce an enormous amount of high scoring candidates which have to be filtered, to sample a reasonable number of predictions. Many *ab initio* miRNA finder (including the one presented in this work) use post-processing filter to reduce the number of putative candidates. Some rely on the fact that many miRNAs are found in cluster (Sewer et al., 2005) and limit new predictions to those found in close proximity to already known miRNA loci. In fact over 50% of animal miRNA loci are found cluster. Other methods filter the set of candidates based on evolutionary conservation as many known miRNA are conserved. Although this step reduces the number of false positive predictions enormously, it somehow excludes species-specific miRNA.

The recent breakthrough in high-throughput sequence technologies like Roche's *454*[8], Illumina's Solexa[9], and ABI's SOLID[10] has a huge impact on miRNA discovery. Based on sequence reads of a size-fractionated small RNA library, methods have been developed to predict miRNA genes in these samples and distinguish them from random RNA degradation products (Friedländer et al., 2008; Hackenberg et al., 2009). Those methods work well in recovering species-specific miRNA with low false positive rate. The basic principle is to align the reads to the genome and score the underlying structure and specific read pattern for each locus. A limiting factor might be the number of sequenced species and the problem of finding low copy miRNA genes. Those genes like e. g., lsy-6 in *C. elegans* are not only expressed in a time-dependent manner and low copy number, but also in a very specific tissue (neuronal cell). So in order to capture all the low-copy time and tissue specific expressed miRNAs, the sequencing step has to include as many tissues and development time points as possible. Noticeable different sequencing platforms are known to yield different results (Linsen et al., 2009). This might certainly also be due to the different protocols and library preparation as even the most abundant miRNAs were recovered with large disparity (Berezikov et al., 2010; Lu et al., 2008).

Still the most promising path to pursue for future miRNA gene finding relies on high-throughput sequencing of small RNA samples, including *ab initio* steps to distinguish native miRNA genes from random RNA degradation fragments. The `miROrtho` pipeline can be extended to include high-throughput sequencing data. This would not only reduce the number of false positive predictions, coupled with evolutionary conservation it can provide a strong evidence for a valid miRNA prediction.

8 `http://www.454.com/`
9 `http://www.illumina.com/`
10 `http://solid.appliedbiosystems.com/`

Part III

STRUCTURAL ELEMENTS IN RHINOVIRUSES

6 INTRODUCTION

Viruses are small infectious agents that can only replicate within a host cell of another organism. With the tobacco mosaic virus discovered by Martinus Beijerinck over 100 years ago, the field of virology took off with already 3,575 complete sequenced genomes annotated in `Genbank` (Benson et al., 2010; Sayers et al., 2010). Viruses can be classified into phyla, classes, orders, families, genera, and species based on the Linnaean taxonomy. The system is based on shared properties between viruses and the type of nucleic acid (Lwoff et al., 1962). The universal system for classifying viruses is being developed by the ICTV (International Committee on Taxonomy of Viruses, 2009). Although thousands of viruses have already been classified and described, recent studies examining environment samples collected from soil or seawater, indicate that the vast majority of viruses found, were completely novel species (Delwart, 2007; Monier et al., 2008; Schoenfeld et al., 2008).

6.1 *picornaviridae*

Based on the Balitmore system (Baltimore, 1971), the International Committee on Taxonomy of Viruses (ICTV) classifies *Picornaviridae* (Table 17 on the next page) as group IV, ss(+)RNA (positive-sense single-stranded RNA) viruses. The Baltimore classification groups viruses into families based on their genomic material and mode of replication. Common features of *Picornaviridae* are: (i) a 7–8.5 kb ss(+)RNA genome packed in an icosahedral capsid of about 30 nm in diameter (ii) a genome 5′ end which is linked to the Viral Protein of the genome (VPg) (Ambros and Baltimore, 1978; Lee et al., 1977; Rothberg et al., 1978) and a polyadenylated 3′ end (Dorsch-Häsler et al., 1975; Yogo and Wimmer, 1972) (iii) a single polyprotein precursor that is cleaved into mature proteins by virus-encoded proteinases (iv) sequence similarities between the genes, especially the non-structural ones and (v) a conserved genomic organization (Fig. 30 on page 95). Nevertheless, some other families, such as *Dicistroviridae* or *Marnaviridae* share many of those features. This had led to the now commonly used concept of a "picorna-like viruses" superfamily (Koonin et al., 2008).

Upon infection, the positive, single stranded RNA molecule is directly translated by the host translation machinery via the internal ribosomal entry side (IRES). The IRES is a nucleotide sequence that folds into a conserved structure and allows for translation initiation of an RNA. The translated product from the single open reading frame (ORF) is then processed into smaller precursor and mature polypeptides. For the replication process, the parental plus-strand RNA is converted into a minus-strand copy, which then serves as the template for the synthesis of progeny positive-stranded genomes. The viral life cycle takes place in the host-cell cytoplasm and releases self-assembled newly synthesized virion particles from the host cell.

Table 17: Genera and species of the family *Picornaviridae*. Note that some commonly known species have been reassigned to different genera and species. The three representative serotypes of the previous described poliovirus species e. g., are now classified as human enterovirus type C (HEV-C). The recently described species human rhinovirus C (HRV-C) is also known as HRV-A2 or HRV-QPM in some publications (Cordey et al., 2008; McErlean et al., 2007). Source: International Committee on Taxonomy of Viruses (2009)

GENUS	SPECIES
Aphthovirus	Bovine rhinitis B virus, Equine rhinitis A virus, Foot-and-mouth disease virus
Avihepatovirus	Duck hepatitis A virus
Cardiovirus	Encephalomyocarditis virus, Theilovirus
Enterovirus	Bovine enterovirus, Human enterovirus A-D, Human rhinovirus A-C, Porcine enterovirus B, Simian enterovirus A
Erbovirus	Equine rhinitis B virus
Hepatovirus	Hepatitis A virus
Kobuvirus	Aichi virus, Bovine kobuvirus
Parechovirus	Human parechovirus, Ljungan virus
Sapelovirus	Avian sapelovirus, Porcine sapelovirus, Simian sapelovirus
Senecavirus	Seneca Valley virus
Teschovirus	Porcine teschovirus
Tremovirus	Avian encephalomyelitis virus

6.2 ENTEROVIRUSES AND RHINOVIRUSES

Human enteroviruses (HEV) and human rhinoviruses (HRV)[1] can infect humans and lead to illnesses of various severity. Rhinoviruses are the most frequent source of respiratory infections in humans, causing the common cold and related respiratory illnesses (Denny, 1995). Enteroviruses cause an estimated 10–15 million symptomatic infections in the United States alone, ranging from mild upper respiratory symptoms, to more serious cases of viral meningitis, myocarditis or encephalitis.

Over 200 rhino- and enterovirus serotypes have been described so far (Gwaltney et al., 1967; International Committee on Taxonomy of Viruses, 2009). Since these viruses are closely related (Tapparel et al., 2007), the three rhinovirus species were recently reclassified within the genus *Enterovirus* (International Committee on Taxonomy of Viruses, 2009; Laine et al., 2005).

Recently a new species of human rhinovirus was discovered in Australian children (McErlean et al., 2007), and adults from the United States (Kistler et al., 2007), Hong Kong (Lau et al., 2007), and further countries including Switzerland. The new species was is now anno-

1 If not stated otherwise, the term enterovirus represents the four species of human enterovirus A–D (HEV-A , HEV-B , HEV-C, and HEV-D), whereas rhinovirus represents the three human rhinovirus species HRV-A , HRV-A2 , and HRV-B .

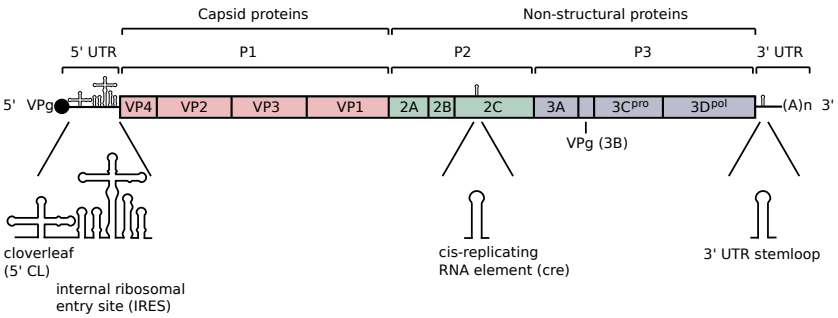


Figure 30: Genomic map of a Picornavirus. The genomic RNA is linked to a virus-encoded peptide (VPg) at the 5′ end and contains a poly(A) tail at the 3′ end. The single open reading frame is divided into one structural domain (P1) and two non-structural domains (P2, P3). The capsid proteins are encoded by the VP4, VP2, VP3, and VP1 genes. Proteinases (2A, $3C^{pro}$), the viral RNA-dependent RNA polymerase ($3D^{pol}$), and other proteins overlap the non-structural domains P2 and P3. The 5′ UTR contains two structural elements: A cloverleaf structure, and separated by a small spacer sequence, an internal ribosomal entry side (IRES). The viral genome also contains a short 3′ UTR which encapsulates a short small stem-loop structure. Furthermore, the ORF encloses one stem-loop structure — the *cis*-acting replication element (*cre*) — which lies in different genes depending on the virus species (e. g., within the 2C gene for HEV A–D).

tated as HRV-C , with known synonyms such as HRV-A2 or HRV-QPM. As those viruses can not be cultivated, old cell culture techniques were not able to dedecte this species. Virus strains from this species are fairly distinct from already known rhinovirus strains. Strains of the HRV-A2 species were used, to study conserved RNA structural elements, and to compare them to elements known in other human rhinoviruses.

Fig. 31 on the next page depicts a phylogenetic tree showing the relationships among all known HRV serotypes.

6.2.1 *Secondary structure elements in the untranslated regions of* Enteroviruses

The *Enterovirus* genome organization follows the general scheme for *Picornaviridae* (Fig. 30). The 5′ non-coding region of enteroviruses contains two functional structural domains. The first one is a cloverleaf (CL) structure (Rivera et al., 1988) that controls both translation and RNA replication (Gamarnik and Andino, 1998). The second structural RNA element in the 5′ UTR is the IRES that enables translation.

6.2.2 *Secondary structure elements in the open reading frame of* Enteroviruses

While RNA elemente in the 5′ (Andino et al., 1993, 1990a,b) and 3′ UTR (Pilipenko et al., 1992; Rohll et al., 1995) have been known since some time, additional important structural RNA elements were found overlapping the ORF of *Picornaviridae* (McKnight and Lemon, 1998). Deletion studies of capsid coding sequences in poliovirus, showed that the capsid regions were dispensable for virus replication (Cole and

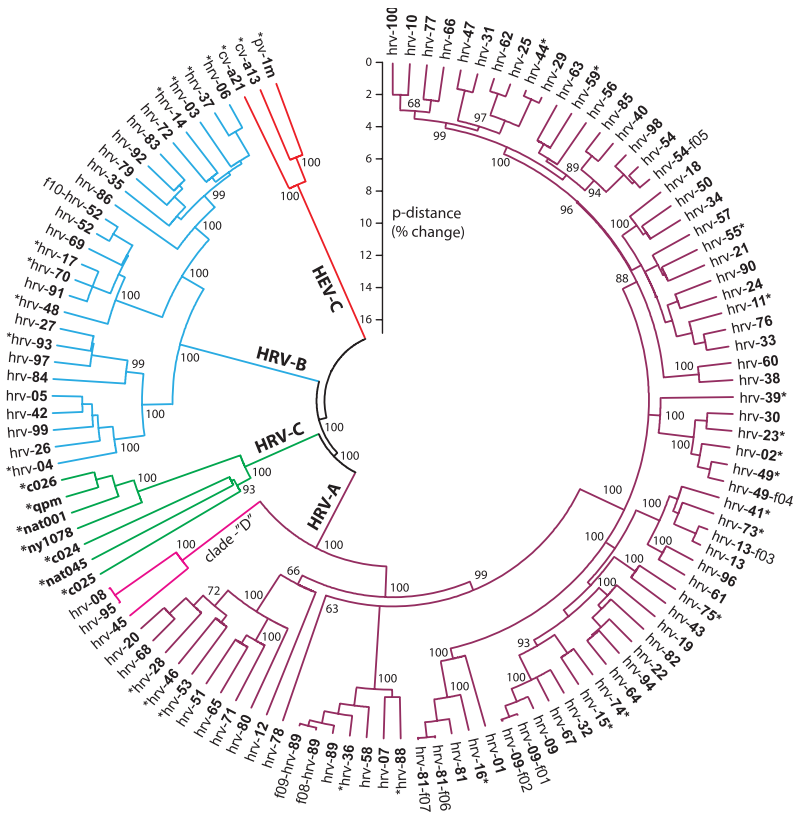


Figure 31: Neighbor-joining tree of human rhinovirus serotypes based on full-length genomic sequences. Source: Palmenberg et al. (2009)

Baltimore, 1973). Indeed, poliovirus replicons, composed of partial virus genomes retaining only the untranslated regions (UTRs) and the two non-structural domains (Hagino-Yamagishi and Nomoto, 1989; Kaplan and Racaniello, 1988), replicate as efficiently as the wild-type virus. The attempt to make RNA replicons from human rhinovirus 14 (HRV-14) with deleted capsid regions failed, as these rhinovirus RNAs were completely defective for RNA replication (McKnight and Lemon, 1996). In a follow-up study McKnight and Lemon (1998) showed, that the capsid VP1 gene of HRV-14 encapsulates a small stable hairpin RNA structure (*cis*-acting replication element (*cre*)), that is indispensable for RNA replication. Since then, similar *cre* structures have been discovered in the following genera: *Aphthovirus* (Mason et al., 2002), *Cardiovirus* (Lobert et al., 1999), *Enterovirus* (Cordey et al., 2008; Gerber et al., 2001; Goodfellow et al., 2000; McKnight and Lemon, 1998; Paul et al., 2000), *Hepatovirus* (Yang et al., 2008), *Parechovirus* (Al-Sunaidi et al., 2007). A *cre* element has not yet been proposed for the remaining picornaviral genomes (Tab. 17 on page 94). The location of the *cre* within the virus genome varies substantially depending on the picornavirus genus or species. Studies in poliovirus have shown, that the *cre* acts as a template for VPg uridylylation through a slide-back mechanism catalyzed by $3D^{pol}$ (RNA-dependent RNA polymerase) (Paul et al., 2000, 2003; Rieder et al., 2000). Uridylated VPg (VPg-pUpU) then serves as the protein primer for new RNA synthesis (Paul et al., 1998).

7 MATERIAL AND METHODS

7.1 MULTIPLE SEQUENCE ALIGNMENTS

Multiple sequence alignments of entero- and rhinovirus strains were constructed using `Muscle` (Edgar, 2004) and `T-Coffee`. First, the translated protein sequences of the open reading frame (ORF) region were aligned using `Muscle`. Second, the UTR were aligned with `T-Coffee` (Notredame et al., 2000). Then a nucleotide-level genome alignments was generated from the nucleotide back-translated (`EMBOSS's TRANSALIGN` (Rice et al., 2000)) amino acid ORF alignment and the alignments of the 5′ and 3′ UTRs.

Protein-coding regions were aligned on the amino-acid level as this is more robust than aligning nucleotide sequences for fairly divergent sequences. RNA/DNA has only four character states, thus two random sequence have just by chance already a sequence identity of 25%. Furthermore, amino-acids have the advantage of distinguishing conservative from non-conservative changes and their sequence is much more constrained. The alignments based on the amino-acid level were certainly indispensable for aligning sequences including divergent outgroup sequences, which were used to compute phylogenetic trees in (Tapparel et al., 2007).

7.2 CONSERVED RNA STRUCTURAL ELEMENTS

Complete genomic alignments of HRV-A, HRV-A2, HRV-B and all four HEV species serotypes were scanned for evolutionary conserved, and thermodynamically stable RNA structures using `RNAz` (Washietl et al., 2005). The structures were evaluated using a sliding window of 120 bp with a step size of 40 bp. A second run with a smaller window size of 60 bp in steps of 20 bp was performed in order to capture smaller conserved RNA structures. Consensus RNA structures were computed using `RNAalifold` from the `Vienna Package` (Hofacker, 2007). Alignments and consensus secondary structures were color-coded (Fig. 2 on page 7) according to the amount of consistent, compensatory, and inconsistent base changes at a certain consensus position using the `Vienna RNA Utilities`[1]. Secondary structures and energies for the individual sequences were calculated using `RNAfold` (Hofacker et al., 1994) and `MFOLD` (Zuker, 2003; Zuker and Stiegler, 1981). The consensus structures were manually evaluated and checked for the *cre* consensus loop sequence motif.

Another approach was used to predict *cre* elements solely based on a single input genome. Genomes were scanned for simple stem-loop structures encompassing a 14 nt large terminal loop. Afterwards, the structures were checked for the conservation of the entero-/rhinovirus loop *cre* sequence motif (*R1NNNA1A2R2NNNNNNR3*). Characteristics of true *cre* elements and such elements which were not supported by comparative evidence, were computed as in Section 3.3.2 on page 36 using `R` (R Development Core Team, 2009).

1 `http://www.tbi.univie.ac.at/~ivo/RNA/utils.html`

8 RESULTS

8.1 5′ and 3′ RNA structural elements

Comparative sequence analysis and consensus folding of entero- and rhinoviruses reveal a well conserved 5′ cloverleaf (5′ CL) RNA element which is highly conserved throughout all picornaviruses. The HRV-A2 5′ CL shows a clear conservation of the structure for all serotypes. This is mainly supported by many compensatory and consistent based changes preserving a common conserved secondary structure (Fig. 32 on the following page). The cloverleaf element was originally discovered in poliviruses (Barton et al., 2001).

Another secondary structure element — the internal ribosomal entry side (IRES) — was discovered in all enteroviruses and rhinoviruses. Fig. 32 on the next page shows the IRES consensus structure for the recently described human rhinovirus A2 species (Kistler et al., 2007; Lamson et al., 2006; McErlean et al., 2008, 2007).

The picornavirus 3′ UTR also encodes a stem-loop structure that is likely to play a role in replication efficiency as well as in polyadenylation of genomic RNA (Brown et al., 2005; van Ooij et al., 2006). As opposed to the 5′ structural elements, the 3′ stem-loop structure element seems to be less constraint regarding its genomic position. Whereas the 5′ elements are located in the same genomic regions for all entero- and rhinovirus species, the 3′ stem-loop position within the untranslated region (UTR) is only conserved for the individual species. The conserved consensus structure of the HRV-A2 3′ stem-loop is shown in Fig. 32 on the following page.

8.2 internal *cis*-acting replication (*cre*) elements

8.2.1 Cre *elements in entero- and rhinoviruses*

Apart from conserved 5′ and 3′ genomic structural RNA elements, *cis*-acting replication elements encoded in the open reading frame (ORF) have been described for many of the major picornavirus genera (reviewed in Steil and Barton, 2009). My comparative secondary structure analysis recovered all known entero- and rhinovirus *cres* with their species-specific positional conservation pattern (Fig. 33 on page 101). The primary RNA sequence of the *cre* element is not well conserved. Nevertheless a specific loop sequence motif and the conserved stable stem-loop structure, are highly conserved among all studies serotypes (Fig. 34 on page 102).

I also predicted a putative second *cre* element in human rhinovirus B at the same location as the already described human enterovirus 2C *cre* (Tapparel et al., 2007). In follow-up study, however, we showed that this putative *cre* is non-functional and is most likely an evolutionary leftover from the species' last common ancestor (Cordey et al., 2008). The *cre* element overlapping the VP1 gene is the only functional one in HRV-B (McKnight and Lemon, 1998).

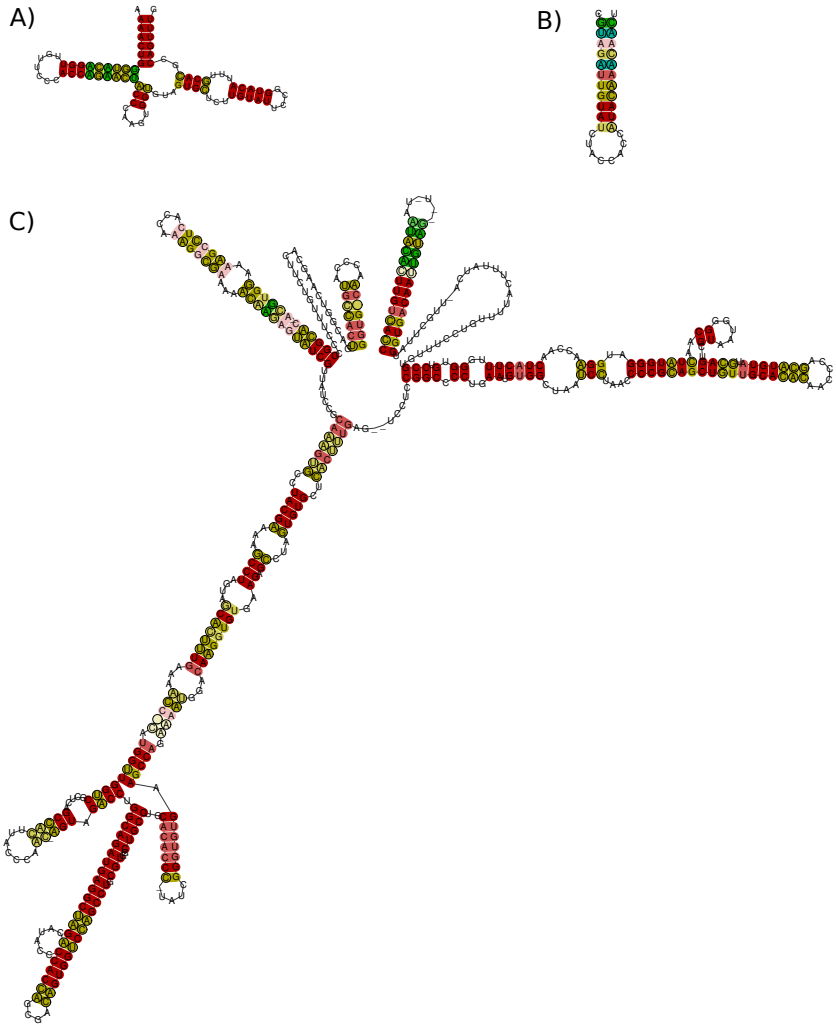


Figure 32: *Cis*-acting RNA elements in HRV-A2. The consensus RNA structures are based on alignments of four different HRV-A2 strains (Genbank accession numbers: EF582385, EF582386, EF582387, EF186077). Note the numerous consistent and compensatory mutations depicted as ocher, green, and blue. A) HRV-A2 consensus cloverleaf in the rhinovirus 5′ UTR B) HRV-A2 consensus stem-loop in the 3′ UTR of the genome C) HRV-A2 consensus structure of the internal ribosomal entry side located in the 5′ UTR just after the cloverleaf element. For the color-code refer to Fig. 2 on page 7.

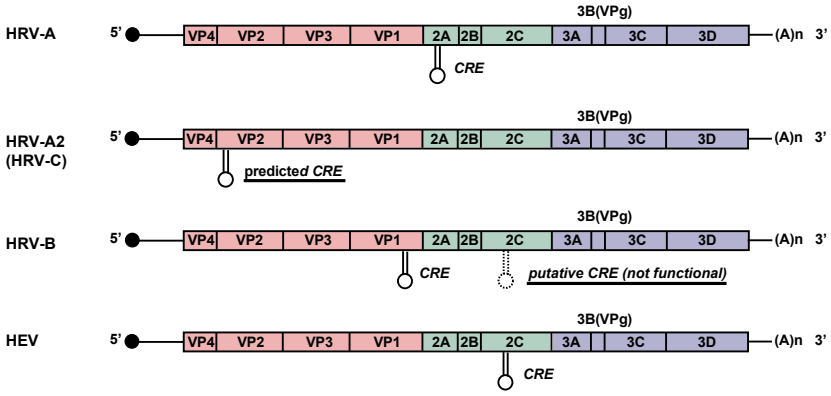


Figure 33: *Cis*-acting replication elements in entero- and rhinoviruses. The single polyprotein consists of one structural (shown in pink) and two non-structural domains (shown in green and purple). The position of the *cre* is conserved within all serotypes of a single species but distinct between the species. A putative *cre* in the 2C gene of HRV-B has been shown to be non-functional (Cordey et al., 2008). Note that for the four enterovirus species A-D the *cre* is always located 2C gene as depicted above.

Furthermore, our analysis (Cordey et al., 2008) also allowed us to propose a new *cre* located in the VP2 gene of the newly described rhinovirus species HRV-A2 (Fig. 35 on page 103). Importantly, this is the first and only predicted *cre* element in this species.

Cre elements have been shown to be functionally exchangeable between different rhinovirus species (Gerber et al., 2001), and several studies have shown that *cre* function is independent from its position within the genome (Goodfellow et al., 2000, 2003; Mason et al., 2002; Yang et al., 2008; Yin et al., 2003). Nevertheless, the location within the genome is highly constrained, at least on the species level.

Cres of enteroviruses and rhinoviruses (Fig. 34 on the following page) are composed of a highly conserved stem-loop structure with a 14 base pair terminal loop. The conserved sequence motif in the loop has the following pattern: *R1NNNA1A2R2NNNNNNR3* (Cordey et al., 2008; Yang et al., 2002). Three R residues (G or A) numbered from 1 to 3 and two Adenosin residues (A1, A2), are conserved at specific positions in the loop. The remaining nucleotides can vary (N) without having an influence on the function. Although the *cres* in the other picornaviruses present more variable loops, two conserved Adenosin residues are always present and indispensable for the *cre* to template for Viral Protein of the genome (VPg) uridylation (Yang et al., 2002).

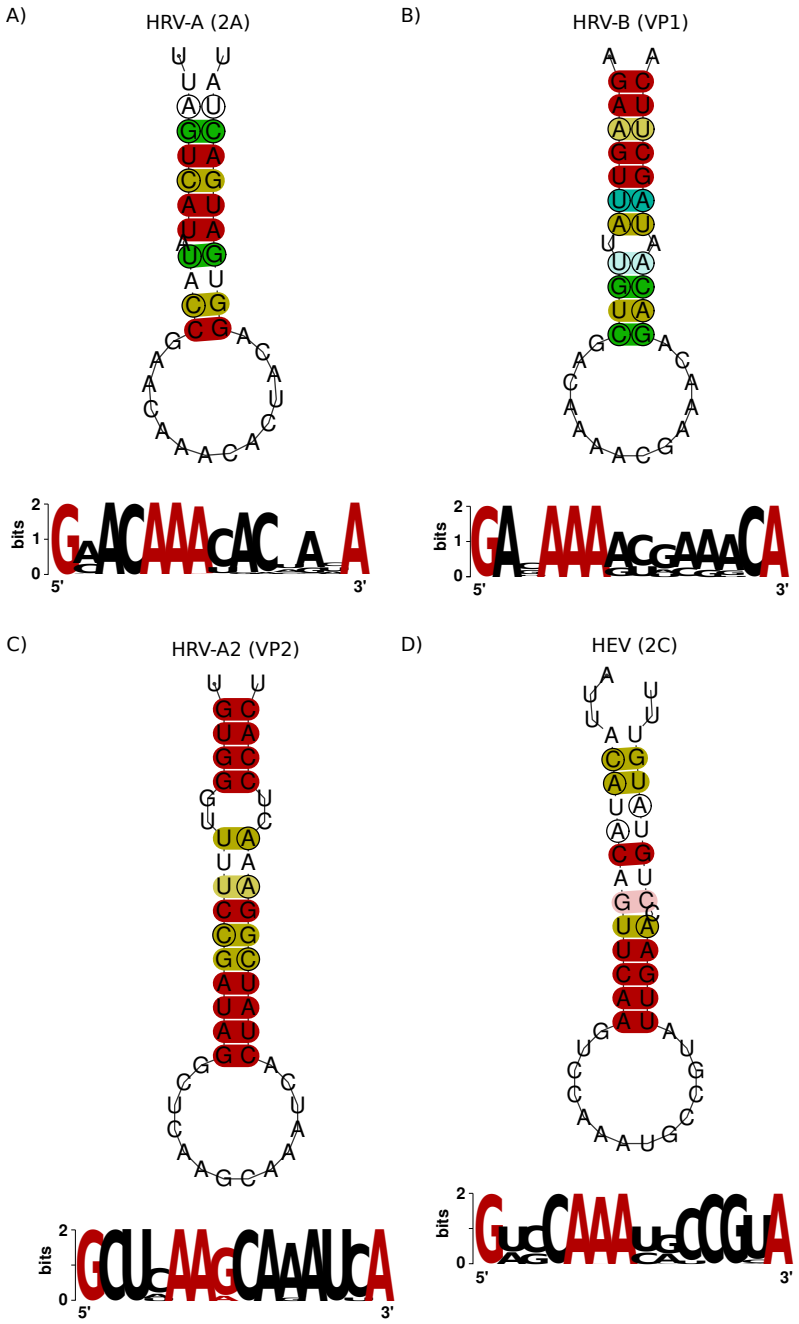


Figure 34: Genomic locations of the *cres* for HRV-A, HRV-A2, and HRV-B and the four other human enterovirus species (HEV-A, HEV-B, HEV-C, and HEV-D). A sequence logo (Crooks et al., 2004; Schneider and Stephens, 1990) of the aligned loop sequences is depicted beneath every structure highlighting the conserved motif in red. The common entero- and rhinovirus loop consensus motif is: *R1NNNA1A2-R2NNNNNNR3* (Yang et al., 2002). A) Human rhinovirus A conserved consensus *cre* derived from 70 strains located in the 2A gene (Gerber et al., 2001) B) Human rhinovirus B conserved consensus *cre* derived from 24 strains located in the VP1 gene (McKnight and Lemon, 1998) C) Human rhinovirus A2 conserved consensus *cre* derived from 28 strains located in the VP2 gene (Cordey et al., 2008) and D) Enterovirus A-D conserved consensus *cre* derived from 86 strains located in the 2C gene (Goodfellow et al., 2000; Paul et al., 2000).

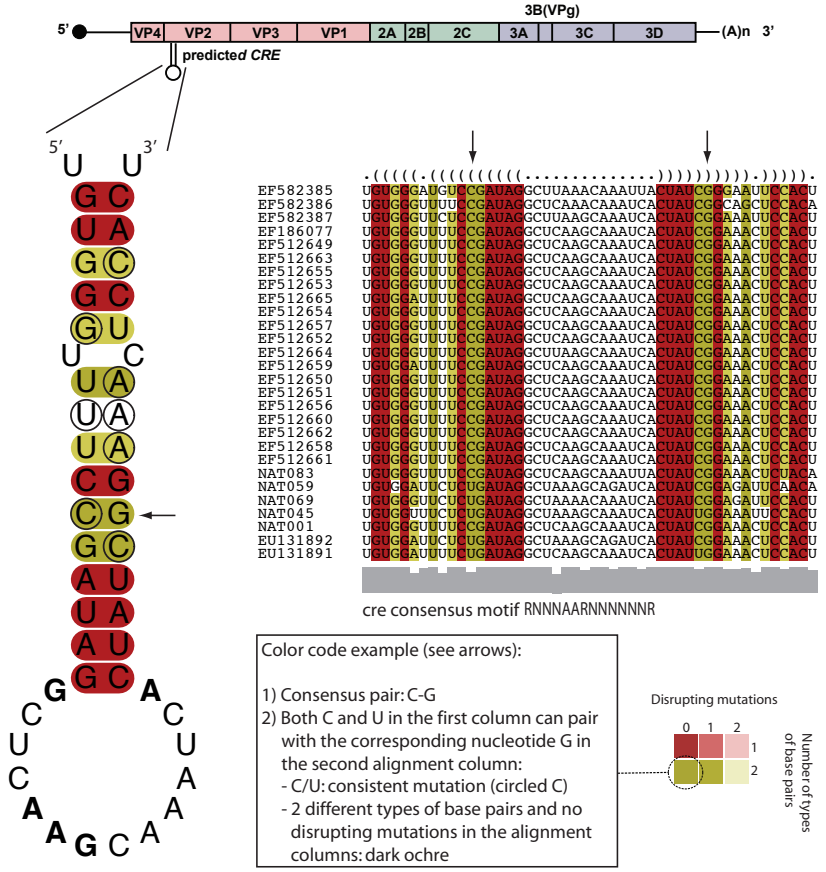


Figure 35: Conservation of a *cre* element in the VP2 gene of HRV-A2. Multiple sequence alignment across all available full-length HRV-A2 and partial VP2 sequences showing a consensus secondary cre structure in VP2. The structure is shown in the dot-parenthesis representation of secondary structure above alignment. Sequences are color-coded according to consistent and compensatory mutations in the aligned sequences regarding the conserved structure (see text box or Fig. 2 on page 7). The sequence conservation profile is shown in gray bars below the alignment. The conserved *cre* motif nucleotides are marked in bold on the consensus structure. Note, that the amino acid sequence corresponding to the loop region is almost 100% conserved in all species (C-G-F-S-D-R-L-K-Q-I-T-I-G/N-S-T). Mutations supporting the structure (consistent mutations) occur almost exclusively at the third codon position, which leads to synonymous codons and the conservation of the amino acid sequence.

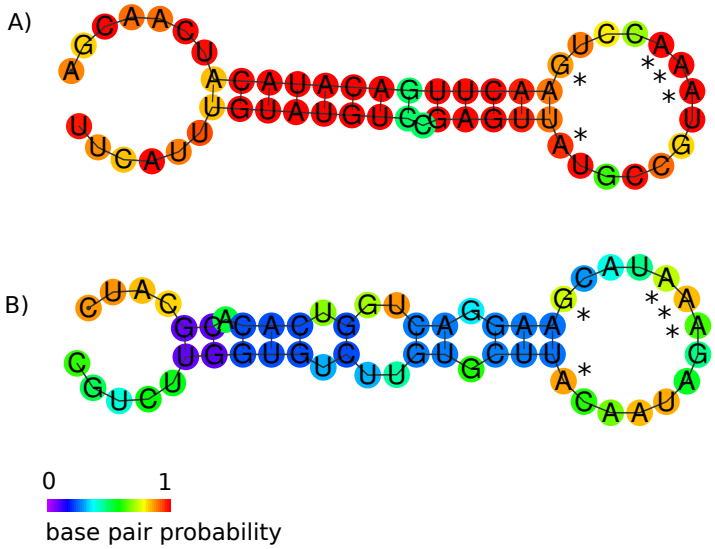


Figure 36: True versus pseudo *cre* elements in enterovirus Echo 1. The base pair probabilty is color-coded — for unpaired regions the color shows the probability of being unpaired. Important nucleotides of the entero- and rhinovirus *cre* loop sequence motif are marked with an asterisk. A) Secondary structure of the functional *cre* element color-coded according to base pair probabilities. B) Another *cre* like structure in the same virus strain presenting a stem-loop structure and a conserved loop sequence motif. Note however, the low overall base pair probability.

8.2.2 *Functional versus "pseudo"* cre *elements*

Based on alignments of related virus strains, comparative genomics allows the reliable prediction of conserved structural *cre* elements. A unique *cre* element can be predicted for each strain using the "back-tranlated" prediction from the `RNAz` scan on the alignments. However, using the minimal sequence and structure requirements deduced from experimentally verified *cre* elements, one can predict several other, similar elements (Fig. 36) for individual entero- and rhinovirus strains. Although the comparative approach retrieves a single prediction, the individual strain is not "aware" of related sequences and has to "choose" the single functional *cre* element. The minimal requirements for a *cre*— a single stem-loop structure with a 14 nt loop and a conserved loop sequence motif (*R1NNNA1A2R2NNNNNNR3*) — were used to search for additional elements in each virus strain used for this study. As shown by Cordey et al. (2008), each rhinovirus species contains only a single functional *cre* element. Therefore, the additional similar elements were named "pseudo" *cres*.

Although the basic requirements for a *cre* are met by both, the true and the "pseudo" *cres*, some characteristics are different. Overall the true *cre* element is the most stable stem-loop structure presenting the loop sequence motif (Fig. 37 on the facing page). Furthermore it is more robust to changes in the flanking sequence context and has a higher overall base pair probability in the ensemble of possible structures (Figs. 36 to 37 on pages 104–105). The functional *cre* is selected for a stable stem-loop structure driven by selective constraints. Finally, comparing true and "pseudo" *cre* elements, the functional elements

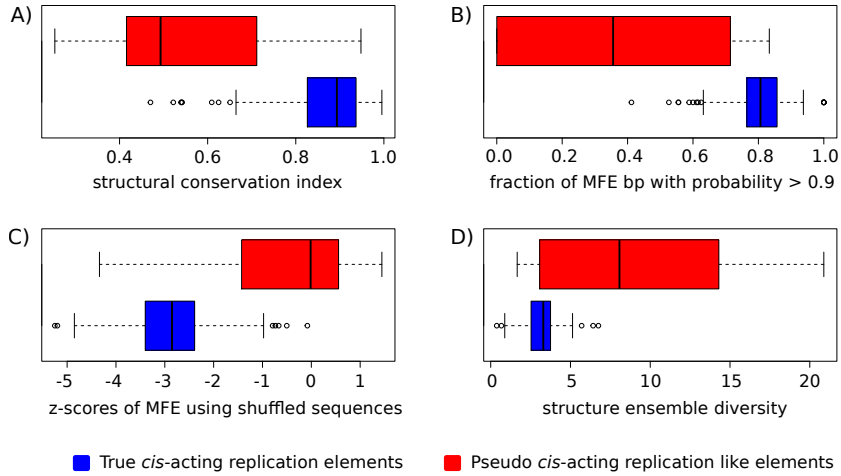


Figure 37: Characteristics of *cre*-like genomic structure elements. Individual entero- and rhinovirus genomes were scanned for stem-loop structures with a 14 nt large loop containing the entero- and rhinovirus *cre* loop sequence motif (*R1NNNA1A2R2NNNNNNR3*). *Cre* elements which have been verified experimentally or show strong evidence from the comparative genomic analysis are named "true" cres, other similar elements (Fig. 36 on the preceding page) are called "pseudo" *cres*. A) Structural conservation index (eq. 3.17 on page 48) B) Fraction of base pairs in the MFE structure with a base pair probability over 0.9 in the ensemble of structures C) Z-scores of MFEs of shuffled sequences (eq. 3.13 on page 39) D) Diversity of structures in the ensemble of structures (eq. 3.11 on page 38).

seem to be depleted of many interior- and bulge-loop structural elements.

9 DISCUSSION

Conserved stem-loop structures that function as a template for the uridylation of VPg have been previously identified in many major picornavirus genera at variable genomic loci. This work presents the identification of a *cre* (*cis*-acting replication element) in the recently described human rhinovirus A2 species (HRV-A2). A functional assay has not been performed yet, due to the lack of infectious clones and the inability of these new viruses to grow in standard cell culture.

The proposed HRV-A2 *cre* has all known features, required for *cre* function: A stable stem-loop structure with mutations in several serotypes supporting a common structure and a throughout all serotypes conserved typical rhinovirus *cre* loop sequences motif (*R1NNNA1A2-R2NNNNNNR3*). Compared to the *cres* in the other two rhinovirus species HRV-A, HRV-B, the HRV-A2 element shows very similar characteristics regarding the stable stem-loop and the structure preserving consistent and compensatory mutations. Nevertheless, as opposed to the other two rhinovirus species that present an A at position R_2, the HRV-A2 *cre* has a G at this position. This observation, however, is still in agreement with both the conserved entero-/rhinovirus *cre* consensus motif (Yang et al., 2002) and the fact that it is the two A residues A_1 and A_2 which are important for the slide-back mechanism of Viral Protein of the genome (VPg) uridylation.

I have identified another *cre* element in the 2C gene of HRV-B (Tapparel et al., 2007), which has been shown to be non-functional (Cordey et al., 2008) in HRV-14 by reverse genetics experiments. This goes in line with my recent studies showing uncharacteristic thermodynamically instability similar to other "pseudo" *cre* elements (see Fig. 36 on page 104 for a "pseudo" *cre* in the enterovirus Echo 1). As the 2C *cre* in HRV-B overlaps the verified 2C *cre* locus of enteroviruses, it may point to a common ancestry of those loci and a "pseudogenization" event in *cre* overlapping the 2C gene of HRV-B. We showed that the sequence conservation in the loop motif is probably due to the requirement for a conserved amino acid in this region (Cordey et al., 2008). Although the respective region folded into a stem-loop structure, the folding was not stable for the individually folded sequences. Noteworthy, a VP1 *cre* had already been described and its function had been experimentally verified in HRV-14 (HRV-B) (McKnight and Lemon, 1998).

This leads to two important findings: First, each species is likely to contain only a single functional *cre* element, although other stem-loop structures with two conserved A residues in the terminal loop exist. Second, the 2C position of the *cre* element is conserved throughout the four human enterovirus species, whereas each rhinovirus species shows a different *cre* location (HRV-A: 2A, HRV-B: VP1, HRV-A2: VP2). Since the sequence identity between each human enterovirus (HEV) species is higher than between all human rhinovirus (HRV) species, a reclassification of all HEV-A – HEV-D species within a unique enterovirus species can be proposed.

In conclusion the extensive study of conserved RNA structural elements in entero- and rhinoviruses shows the evolutionary constraints

on sequence, structure, and position within the genome. The 5′ cloverleaf element is not only conserved in structure but also in its position among all members of the genus. The same holds true for the internal ribosomal entry side (IRES). The stem-loop structure in the 3′ UTR can also be found in all studied enteroviruses, its location within the untranslated region (UTR) is however variable. A *cre* can be found in the open reading frame (ORF) of all enteroviruses, showing many consistent and compensatory mutations among serotypes. The function of the *cre* is determined by a thermodynamically stable stem-loop structure with very few small bulge- and interior-loops. Further, a indispensable loop sequence motif is found in all *cre* elements. Despite studies showing that the *cre* element can work at several different genomic loci (Goodfellow et al., 2000, 2003; Mason et al., 2002; Yang et al., 2008; Yin et al., 2003), the position of the *cre* is conserved for each rhinovirus species and for the four enterovirus species.

Further experimental studies have to be conducted to study the constraints limiting the position of the *cre* and the criteria by which the virus "chooses" a single functional *cre* among similar possible structures along the RNA genome.

Part IV

GENERAL DISCUSSION

10 DISCUSSION

The work presents an study of functional genomic elements that adapt a conserved RNA secondary structure. This is either due to their specific biogenesis or to constraints of the structure pursuing a biological function. The basic concept used for this thesis is built around comparative genomics. This methods uses genome comparison among related organisms to extract biological information. Key achievements in this work are the development of a novel microRNA (miRNA) prediction pipeline and the discovery of a novel functional element required for replication in the human rhinovirus A2 (HRV-A2) folding into a conserved stem-loop structure. Both goals would not have been achieved without modern information technology and the wealth of publically available biological data.

10.1 COMPUTATIONAL MICRORNA DISCOVERY

I have developed methods for the genome-wide consistent annotation of animal miRNA genes. miRNAs are small non-coding RNA (ncRNA) genes that regulate gene expression on a post-transcriptional level. Based on a conserved biogenesis pathway, specific sequence, structural and thermodynamical constraints are conserved for miRNA genes. I used those properties to devise a computational method which is able to learn statistical properties of known miRNA genes to predict novel candidates in complete genome sequences. The method was used to extensively annotate the miRNAome of over 40 animal genomes. Additionally, many of the prediction were included in the official gene sets of large genome sequencing projects (Elsik et al., 2009; Werren et al., 2010). The miRNA data is freely accessible through a web-based database named `miROrtho`: `http://cegg.unige.ch/mirortho` (Gerlach et al., 2009).

The main steps of the `miROrtho` pipeline can be summarized as follows: Stable miRNA-like hairpins are extracted from complete genomic sequences. Afterwards, a specialized Support Vector Machine (SVM) model training on miRNA hairpins and non-miRNA hairpins is used to score those hairpins for their likelihood to be novel putative miRNA genes. High scoring predictions are combined with predictions from searches for homologs of known miRNAs to gather a set of putative miRNA genes. Those are grouped into orthologous groups, which are then subject to a second SVM model — now trained on a set of alignment of known miRNA gene families and unrelated ncRNA gene alignments. Thereby, every orthologous groups gets a score assigned which represents the probability of an RNA alignment to be a new orthologous group of related miRNA genes.

Predictions of miRNAs in a diverse set of species, will help upcoming studies focusing on the evolution of miRNAs in animal genomes and their constraints.

10.2 A *cis*-ACTING REPLICATION ELEMENT IN A NOVEL HUMAN RHINOVIRUS SPECIES

Based on the general concept of evolutionary conserved secondary RNA structures which is helpful to discover the stem-loop like structure of precursor microRNAs (pre-miRNAs), I also studied structural elements in entero- and rhinovirus genomes. These viruses belong to the family *Picornaviridea*, and are characterized by a single stranded, positive RNA genome which has a length of about 7 kb (Fig. 30 on page 95). The genome is transcribed as a single open reading frame (ORF). The product is then processed by proteinases to release the mature proteins and polypepetides. Importantly, the genome also contains a number of conserved structural RNA elements that are crucial for the virus biology. Small stem-loop structures overlapping the ORF of entero- and rhinoviruses have been shown to implied in virus replication (McKnight and Lemon, 1998).

I have developed a method to detect these elements, named *cis*-acting replication element (*cre*), in viruses. Based on an alignment of several strains of a recently described novel rhinovirus species named HRV-A2[1] (McErlean et al., 2008, 2007), I predicted a conserved *cre* element for this species (Cordey et al., 2008). Extensive studies and comparisons of the *cre* elements in the three rhino- and four enterovirus species show specific sequence and structural requirements necessary for the uridylation function of the *cre* acting on Viral Protein of the genome (VPg) to promote replication. The *cre* element has been shown to be independent from the genomic loci were it is located, and can be displaced without disrupting the function (Goodfellow et al., 2000, 2003; Mason et al., 2002; Yang et al., 2008; Yin et al., 2003). Although the genomic location seems not to be constraint implied by function, my work shows that the genomic locus for this element, however, is conserved for all serotypes within each of the three human rhinovirus species (HRV-A, HRV-A2, HRV-B) and all four human enterovirus species (HEV-A, HEV-B, HEV-C, and HEV-D).

So far it is not clear why the position of the *cre* is conserved for each species. The specific *cre* position for each species may point out to additional mechanisms keeping the species "separated". A recent study of ours, however, showed also possible recombination events among rhinovirus species (Tapparel et al., 2009).

Nevertheless, the specific location of the *cre* in each rhino- and the enterovirus species, allows the classification of newly discovered serotypes based on this element. Based on the overall high sequence similarity among the four enterovirus species (HEV-A, HEV-B, HEV-C, and HEV-D) and their common conserved *cre* element overlapping the 2C gene, we propose a reclassification of the four species into one new enterovirus species (Cordey et al., 2008).

Knowledge about RNA structures in the genomes of plus-stranded RNA viruses will not only enhance our understanding of the proliferation of those viruses, but can also lead to the development of new antiviral drugs, targeting structures like the *cre*. This could lead to a complete replication stop of the virus as this element is indispensable for virus replication.

1 HRV-A2, HRV-C, and HRV-QPM are synonyms for the same species.

Abrahante JE, Daul AL, Li M, Volk ML, Tennessen JM, Miller EA, and Rougvie AE (2003) The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. *Dev. Cell*, 4:625–637.

Al-Sunaidi M, Williams CH, Hughes PJ, Schnurr DP, and Stanway G (2007) Analysis of a new human parechovirus allows the definition of parechovirus types and the identification of RNA structural domains. *J. Virol.*, 81:1013–1021.

Alpaydin E (2004) *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. MIT Press, Cambridge, MA.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410.

Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, Aravin A, Brownstein MJ, Tuschl T, and Margalit H (2005) Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res.*, 33:2697–2706.

Ambros V and Baltimore D (1978) Protein is linked to the 5′ end of poliovirus RNA by a phosphodiester linkage to tyrosine. *J. Biol. Chem.*, 253:5263–5266.

Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, and Tuschl T (2003) A uniform system for microRNA annotation. *RNA*, 9:277–279.

Andino R, Rieckhof GE, Achacoso PL, and Baltimore D (1993) Poliovirus RNA synthesis utilizes an RNP complex formed around the 5′-end of viral RNA. *EMBO J.*, 12:3587–3598.

Andino R, Rieckhof GE, and Baltimore D (1990a) A functional ribonucleoprotein complex forms around the 5′ end of poliovirus RNA. *Cell*, 63:369–380.

Andino R, Rieckhof GE, Trono D, and Baltimore D (1990b) Substitutions in the protease (3Cpro) gene of poliovirus can suppress a mutation in the 5′ noncoding region. *J. Virol.*, 64:607–612.

Appel RD (2009) *Bioinformatics: A Swiss Perspective*. World Scientific Publishing Company.

Armisen J, Gilchrist MJ, Wilczynska A, Standart N, and Miska EA (2009) Abundant and dynamically expressed miRNAs, piRNAs, and other small RNAs in the vertebrate Xenopus tropicalis. *Genome Res.*, 19:1766–1775.

Artzi S, Kiezun A, and Shomron N (2008) miRNAminer: a tool for homologous microRNA gene search. *BMC Bioinformatics*, 9:39.

Avner P and Heard E (2001) X-chromosome inactivation: counting, choice and initiation. *Nat. Rev. Genet.*, 2:59–67.

Bachellerie JP, Cavaillé J, and Hüttenhofer A (2002) The expanding snoRNA world. *Biochimie*, 84:775–790.

Baltimore D (1971) Expression of animal virus genomes. *Bacteriol Rev*, 35:235–241.

Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, 136:215–233.

Barton DJ, O'Donnell BJ, and Flanegan JB (2001) 5′ cloverleaf in poliovirus RNA is a cis-acting replication element required for negative-strand synthesis. *EMBO J.*, 20:1439–1448.

Batuwita R and Palade V (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, 25:989–995.

Bejarano F, Smibert P, and Lai EC (2009) miR-9a prevents apoptosis during wing development by repressing Drosophila LIM-only. *Dev. Biol.*

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Sayers EW (2010) GenBank. *Nucleic Acids Res.*, 38:46–51.

Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, and Bentwich Z (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.*, 37:766–770.

Berezikov E, Chung WJ, Willis J, Cuppen E, and Lai EC (2007) Mammalian mirtron genes. *Mol. Cell*, 28:328–336.

Berezikov E, Cuppen E, and Plasterk RH (2006a) Approaches to microRNA discovery. *Nat. Genet.*, 38 Suppl:2–7.

Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, and Cuppen E (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 120:21–24.

Berezikov E, Liu N, Flynt AS, Hodges E, Rooks M, Hannon GJ, and Lai EC (2010) Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. *Nat. Genet.*, 42:6–9.

Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, and Plasterk RH (2006b) Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.*, 38:1375–1377.

Bernstein E, Caudy AA, Hammond SM, and Hannon GJ (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409:363–366.

Bernstein E, Kim SY, Carmell MA, Murchison EP, Alcorn H, Li MZ, Mills AA, Elledge SJ, Anderson KV, and Hannon GJ (2003) Dicer is essential for mouse development. *Nat. Genet.*, 35:215–217.

Bishop CM (2007) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, Luxemburg, Berlin.

Bohnsack MT, Regener K, Schwappach B, Saffrich R, Paraskeva E, Hartmann E, and Görlich D (2002) Exp5 exports eEF1A via tRNA from nuclei and synergizes with other transport pathways to confine translation to the cytoplasm. *EMBO J.*, 21:6205–6215.

Bonnet E, Wuyts J, Rouzé P, and Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20:2911–2917.

Borchert GM, Lanier W, and Davidson BL (2006) RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol.*, 13:1097–1101.

Brameier M and Wiuf C (2007) Ab initio identification of human microRNAs based on structure motifs. *BMC Bioinformatics*, 8:478.

Brannan CI and Bartolomei MS (1999) Mechanisms of genomic imprinting. *Curr. Opin. Genet. Dev.*, 9:164–170.

Brennecke J, Hipfner DR, Stark A, Russell RB, and Cohen SM (2003) bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell*, 113:25–36.

Bringhurst R (2002) *The Elements of Typographic Style*. Hartley and Marks Publishers.

Brown DM, Cornell CT, Tran GP, Nguyen JH, and Semler BL (2005) An authentic 3′ noncoding region is necessary for efficient poliovirus replication. *J. Virol.*, 79:11962–11973.

Burkard ME, Turner DH, and Tinoco Jr I (1999) Structures of base pairs involving at least two hydrogen bonds. In: Gesteland RF, Cech TR, and Atkins JF (Editors), *The RNA World*, pp. 675–680. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Cai X and Cullen BR (2007) The imprinted H19 noncoding RNA is a primary microRNA precursor. *RNA*, 13:313–316.

Cai X, Hagedorn CH, and Cullen BR (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10:1957–1966.

Carthew RW and Sontheimer EJ (2009) Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136:642–655.

Cavaillé J, Nicoloso M, and Bachellerie JP (1996) Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides. *Nature*, 383:732–735.

Chan PP and Lowe TM (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, 37:D93–97.

Chang CC and Lin CJ (2001) *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chang DT, Wang CC, and Chen JW (2008) Using a kernel density estimation based classifier to predict species-specific microRNA precursors. *BMC Bioinformatics*, 9 Suppl 12:S2.

Chawla NV, Bowyer KW, Hall LO, and Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res. (JAIR)*, 16.

Chen JF, Mandel EM, Thomson JM, Wu Q, Callis TE, Hammond SM, Conlon FL, and Wang DZ (2006) The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat. Genet.*, 38:228–233.

Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, Nishikura K, and Shiekhattar R (2005) TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*, 436:740–744.

Chung WJ, Okamura K, Martin R, and Lai EC (2008) Endogenous RNA interference provides a somatic defense against Drosophila transposons. *Curr. Biol.*, 18:795–802.

Cole CN and Baltimore D (1973) Defective interfering particles of poliovirus. II. Nature of the defect. *J. Mol. Biol.*, 76:325–343.

Cordey S, Gerlach D, Junier T, Zdobnov EM, Kaiser L, and Tapparel C (2008) The cis-acting replication elements define human enterovirus and rhinovirus species. *RNA*, 14:1568–1578.

Cordey S, Junier T, Gerlach D, Gobbini F, Farinelli L, Zdobnov EM, Winther B, Tapparel C, and Kaiser L (2010) Rhinovirus Genome Evolution during Experimental Human Infection. *PLoS ONE*, 5:e10588.

Crooks GE, Hon G, Chandonia JM, and Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res.*, 14:1188–1190.

Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, Hannon GJ, and Brennecke J (2008) An endogenous small interfering RNA pathway in Drosophila. *Nature*, 453:798–802.

Darty K, Denise A, and Ponty Y (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25:1974–1975.

Das S (2009) Evolutionary origin and genomic organization of micro-RNA genes in immunoglobulin lambda variable region gene family. *Mol. Biol. Evol.*, 26:1179–1189.

Delwart EL (2007) Viral metagenomics. *Rev. Med. Virol.*, 17:115–131.

Denli AM, Tops BB, Plasterk RH, Ketting RF, and Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*, 432:231–235.

Denny FW (1995) The clinical impact of human respiratory virus infections. *Am. J. Respir. Crit. Care Med.*, 152:4–12.

Doench JG, Petersen CP, and Sharp PA (2003) siRNAs can function as miRNAs. *Genes Dev.*, 17:438–442.

Dorsch-Häsler K, Yogo Y, and Wimmer E (1975) Replication of picornaviruses. I. Evidence from in vitro RNA synthesis that poly(A) of the poliovirus genome is genetically coded. *J. Virol.*, 16:1512–1517.

Doshi KJ, Cannone JJ, Cobaugh CW, and Gutell RR (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5:105.

Doudna JA and Batey RT (2004) Structural insights into the signal recognition particle. *Annu. Rev. Biochem.*, 73:539–557.

Doudna JA and Cech TR (2002) The chemical repertoire of natural ribozymes. *Nature*, 418:222–228.

Dowell RD and Eddy SR (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5:71.

Duret L, Chureau C, Samain S, Weissenbach J, and Avner P (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science*, 312:1653–1655.

Eddy SR (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, 2:919–929.

Eddy SR (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, 3:18.

Eddy SR and Durbin R (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, 22:2079–2088.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32:1792–1797.

Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE, Elnitski L, Guigó R, Hamernik DL, Kappes SM, Lewin HA, Lynn DJ, Nicholas FW, et al. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, 324:522–528.

Ender C, Krek A, Friedländer MR, Beitzinger M, Weinmann L, Chen W, Pfeffer S, Rajewsky N, and Meister G (2008) A human snoRNA with microRNA-like functions. *Mol. Cell*, 32:519–528.

Esau C, Davis S, Murray SF, Yu XX, Pandey SK, Pear M, Watts L, Booten SL, Graham M, McKay R, Subramaniam A, Propp S, Lollo BA, Freier S, Bennett CF, et al. (2006) miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. *Cell Metab.*, 3:87–98.

Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27:861–874.

Freyhult E, Gardner PP, and Moulton V (2005) A comparison of RNA folding measures. *BMC Bioinformatics*, 6:241.

Friedländer MR, Adamidi C, Han T, Lebedeva S, Isenbarger TA, Hirst M, Marra M, Nusbaum C, Lee WL, Jenkin JC, Sánchez Alvarado A, Kim JK, and Rajewsky N (2009) High-resolution profiling and discovery of planarian small RNAs. *Proc. Natl. Acad. Sci. U.S.A.*, 106:11546–11551.

Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, and Rajewsky N (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, 26:407–415.

Friedman LM, Dror AA, Mor E, Tenne T, Toren G, Satoh T, Biesemeier DJ, Shomron N, Fekete DM, Hornstein E, and Avraham KB (2009a) MicroRNAs are essential for development and function of inner ear hair cells in vertebrates. *Proc. Natl. Acad. Sci. U.S.A.*, 106:7915–7920.

Friedman RC, Farh KK, Burge CB, and Bartel DP (2009b) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19:92–105.

Förstemann K, Tomari Y, Du T, Vagin VV, Denli AM, Bratu DP, Klattenhoff C, Theurkauf WE, and Zamore PD (2005) Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol.*, 3:e236.

Gamarnik AV and Andino R (1998) Switch from translation to RNA replication in a positive-stranded RNA virus. *Genes Dev.*, 12:2293–2304.

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, and Bateman A (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, 37:D136–140.

Gardner PP, Wilm A, and Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, 33:2433–2439.

Gatfield D, Le Martelot G, Vejnar CE, Gerlach D, Schaad O, Fleury-Olela F, Ruskeepää AL, Oresic M, Esau CC, Zdobnov EM, and Schibler U (2009) Integration of microRNA miR-122 in hepatic circadian gene expression. *Genes Dev.*, 23:1313–1326.

Gerber K, Wimmer E, and Paul AV (2001) Biochemical and genetic studies of the initiation of human rhinovirus 2 RNA replication: identification of a cis-replicating element in the coding sequence of 2A(pro). *J. Virol.*, 75:10979–10990.

Gerlach D, Kriventseva EV, Rahman N, Vejnar CE, and Zdobnov EM (2009) miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.*, 37:D111–117.

Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler EL, Zapp ML, Weng Z, and Zamore PD (2008) Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science*, 320:1077–1081.

Gilbert W (1986) Origin of life: The RNA world. *Nature*, 319:618.

Giraldez AJ, Cinalli RM, Glasner ME, Enright AJ, Thomson JM, Baskerville S, Hammond SM, Bartel DP, and Schier AF (2005) MicroRNAs regulate brain morphogenesis in zebrafish. *Science*, 308:833–838.

Glazov EA, Kongsuwan K, Assavalapsakul W, Horwood PF, Mitter N, and Mahony TJ (2009) Repertoire of bovine miRNA and miRNA-like small regulatory RNAs expressed upon viral infection. *PLoS ONE*, 4:e6349.

Goodfellow I, Chaudhry Y, Richardson A, Meredith J, Almond JW, Barclay W, and Evans DJ (2000) Identification of a cis-acting replication element within the poliovirus coding region. *J. Virol.*, 74:4590–4600.

Goodfellow IG, Kerrigan D, and Evans DJ (2003) Structure and function analysis of the poliovirus cis-acting replication element (CRE). *RNA*, 9:124–137.

Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, and Kim J (2003) Computational and experimental identification of C. elegans microRNAs. *Mol. Cell*, 11:1253–1263.

Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, and Shiekhattar R (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature*, 432:235–240.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, and Bateman A (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, 33:D121–124.

Griffiths-Jones S, Saini HK, van Dongen S, and Enright AJ (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, 36:D154–158.

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, and Bartel DP (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455:1193–1197.

Grishok A, Pasquinelli AE, Conte D, Li N, Parrish S, Ha I, Baillie DL, Fire A, Ruvkun G, and Mello CC (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106:23–34.

Gruber AR, Findeiß S, Washietl S, Hofacker IL, and Stadler PF (2010) RNAZ 2.0: IMPROVED NONCODING RNA DETECTION. *Pac Symp Biocomput*, 15:69–79.

Gruber AR, Kilgus C, Mosig A, Hofacker IL, Hennig W, and Stadler PF (2008) Arthropod 7SK RNA. *Mol. Biol. Evol.*, 25:1923–1930.

Guindon S, Lethiec F, Duroux P, and Gascuel O (2005) PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.*, 33:W557–559.

Gutell RR, Lee JC, and Cannone JJ (2002) The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.*, 12:301–310.

Gwaltney JM, Hendley JO, Simon G, and Jordan WS (1967) Rhinovirus infections in an industrial population. II. Characteristics of illness and antibody response. *JAMA*, 202:494–500.

Gürsoy HC, Koper D, and Benecke BJ (2000) The vertebrate 7S K RNA separates hagfish (Myxine glutinosa) and lamprey (Lampetra fluviatilis). *J. Mol. Evol.*, 50:456–464.

Ha I, Wightman B, and Ruvkun G (1996) A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. *Genes Dev.*, 10:3041–3050.

Haase AD, Jaskiewicz L, Zhang H, Lainé S, Sack R, Gatignol A, and Filipowicz W (2005) TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. *EMBO Rep.*, 6:961–967.

Hackenberg M, Sturm M, Langenberger D, Falcón-Pérez JM, and Aransay AM (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.*, 37:68–76.

Hagino-Yamagishi K and Nomoto A (1989) In vitro construction of poliovirus defective interfering particles. *J. Virol.*, 63:5386–5392.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, and Witten IH (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11.

Hamming RW (1950) Error detecting and error correcting codes. *Bell System Tech. J.*, 29:147–160.

Hammond SM, Bernstein E, Beach D, and Hannon GJ (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, 404:293–296.

Hammond SM, Boettcher S, Caudy AA, Kobayashi R, and Hannon GJ (2001) Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science*, 293:1146–1150.

Han J, Lee Y, Yeom KH, Kim YK, Jin H, and Kim VN (2004) The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.*, 18:3016–3027.

Havgaard JH, Torarinsson E, and Gorodkin J (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.*, 3:1896–1908.

Hayashita Y, Osada H, Tatematsu Y, Yamada H, Yanagisawa K, Tomida S, Yatabe Y, Kawahara K, Sekido Y, and Takahashi T (2005) A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer Res.*, 65:9628–9632.

Helvik SA, Snøve O, and Saetrom P (2007) Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics*, 23:142–149.

Hertel J, Hofacker IL, and Stadler PF (2008) SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24:158–164.

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, and Stadler PF (2006) The expansion of the metazoan microRNA repertoire. *BMC Genomics*, 7:25.

Hertel J and Stadler PF (2006) Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 22:197–202.

Hofacker IL (2007) RNA consensus structure prediction with RNAalifold. *Methods Mol. Biol.*, 395:527–544.

Hofacker IL, Fekete M, and Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066.

Hofacker IL, Fontana W, Stadler PF, Bonhöffer S, Tacker M, and Schuster P (1994) Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie*, 125:167–188.

Hofacker IL, Priwitzer B, and Stadler PF (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20:186–190.

Huang TH, Fan B, Rothschild MF, Hu ZL, Li K, and Zhao SH (2007) MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics*, 8:341.

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, et al. (2009) Ensembl 2009. *Nucleic Acids Res.*, 37:D690–697.

Hutvágner G, McLachlan J, Pasquinelli AE, Bálint E, Tuschl T, and Zamore PD (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293:834–838.

Hutvágner G and Zamore PD (2002) A microRNA in a multiple-turnover RNAi enzyme complex. *Science*, 297:2056–2060.

Huynen M, Gutell R, and Konings D (1997) Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.*, 267:1104–1112.

Huynen MA and Bork P (1998) Measuring genome evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 95:5849–5856.

Hyun S, Lee JH, Jin H, Nam J, Namkoong B, Lee G, Chung J, and Kim VN (2009) Conserved MicroRNA miR-8/miR-200 and its target USH/FOG2 control growth by regulating PI3K. *Cell*, 139:1096–1108.

International Committee on Taxonomy of Viruses (2009) Ictv master species list 2009 - version 3. `http://www.ictvonline.org`.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945.

Jain A and Zongker D (1997) Feature selection: evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19:153–158.

Jiang F, Ye X, Liu X, Fincher L, McKearin D, and Liu Q (2005) Dicer-1 and R3D1-L catalyze microRNA maturation in Drosophila. *Genes Dev.*, 19:1674–1679.

Jiang P, Wu H, Wang W, Ma W, Sun X, and Lu Z (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, 35:W339–344.

Johnston RJ and Hobert O (2003) A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature*, 426:845–849.

Jopling CL, Yi M, Lancaster AM, Lemon SM, and Sarnow P (2005) Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science*, 309:1577–1581.

Junn E, Lee KW, Jeong BS, Chan TW, Im JY, and Mouradian MM (2009) Repression of alpha-synuclein expression and toxicity by microRNA-7. *Proc. Natl. Acad. Sci. U.S.A.*, 106:13052–13057.

Kadri S, Hinman V, and Benos PV (2009) HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics*, 10 Suppl 1:S35.

Kaplan G and Racaniello VR (1988) Construction and characterization of poliovirus subgenomic replicons. *J. Virol.*, 62:1687–1696.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316:1484–1488.

Katoh K, Asimenos G, and Toh H (2009) Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.*, 537:39–64.

Keerthi SS and Lin CJ (2003) Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput*, 15:1667–1689.

Ketting RF, Fischer SE, Bernstein E, Sijen T, Hannon GJ, and Plasterk RH (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev.*, 15:2654–2659.

Khvorova A, Reynolds A, and Jayasena SD (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115:209–216.

Kim VN (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell Biol.*, 6:376–385.

Kim VN, Han J, and Siomi MC (2009) Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, 10:126–139.

Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, Gerlach D, Kriventseva EV, Elsik CG, Graur D, Hill CA, et al. (2010) Genome Sequences of the Human Body Louse and its Primary Endosymbiont: Insights into the Permanent Parasitic Lifestyle. *Proc. Natl. Acad. Sci. U.S.A.* In Press.

Kiryu H, Tabei Y, Kin T, and Asai K (2007) Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, 23:1588–1598.

Kistler A, Avila PC, Rouskin S, Wang D, Ward T, Yagi S, Schnurr D, Ganem D, DeRisi JL, and Boushey HA (2007) Pan-viral screening of respiratory tract infections in adults with and without asthma reveals unexpected human coronavirus and human rhinovirus diversity. *J. Infect. Dis.*, 196:817–825.

Klattenhoff C and Theurkauf W (2008) Biogenesis and germline functions of piRNAs. *Development*, 135:3–9.

Klein RJ, Misulovin Z, and Eddy SR (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci. U.S.A.*, 99:7542–7547.

Koonin EV, Wolf YI, Nagasaki K, and Dolja VV (2008) The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nat. Rev. Microbiol.*, 6:925–939.

Korf I, Yandell M, and Bedell J (2003) *Blast*. O'Reilly Media, Sebastopol, CA.

Kriventseva EV, Rahman N, Espinosa O, and Zdobnov EM (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.*, 36:D271–275.

Lafontaine DL and Tollervey D (1998) Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. *Trends Biochem. Sci.*, 23:383–388.

Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, and Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, 35:3100–3108.

Lagos-Quintana M, Rauhut R, Lendeckel W, and Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. *Science*, 294:853–858.

Lai EC, Tomancak P, Williams RW, and Rubin GM (2003) Computational identification of Drosophila microRNA genes. *Genome Biol.*, 4:R42.

Laine P, Savolainen C, Blomqvist S, and Hovi T (2005) Phylogenetic analysis of human rhinovirus capsid protein VP1 and 2A protease coding sequences confirms shared genus-like relationships with human enteroviruses. *J. Gen. Virol.*, 86:697–706.

Lamson D, Renwick N, Kapoor V, Liu Z, Palacios G, Ju J, Dean A, St George K, Briese T, and Lipkin WI (2006) MassTag polymerase-chain-reaction detection of respiratory pathogens, including a new rhinovirus genotype, that caused influenza-like illness in New York State during 2004-2005. *J. Infect. Dis.*, 194:1398–1402.

Landthaler M, Yalcin A, and Tuschl T (2004) The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. *Curr. Biol.*, 14:2162–2167.

Lau NC, Lim LP, Weinstein EG, and Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294:858–862.

Lau SK, Yip CC, Que TL, Lee RA, Au-Yeung RK, Zhou B, So LY, Lau YL, Chan KH, Woo PC, and Yuen KY (2007) Clinical and molecular epidemiology of human bocavirus in respiratory and fecal samples from children in Hong Kong. *J. Infect. Dis.*, 196:986–993.

Le SY and Maizel JV (1989) A method for assessing the statistical significance of RNA folding. *J. Theor. Biol.*, 138:495–510.

Lee MT and Kim J (2008) Self containment, a property of modular RNA structures, distinguishes microRNAs. *PLoS Comput. Biol.*, 4:e1000150.

Lee RC and Ambros V (2001) An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294:862–864.

Lee RC, Feinbaum RL, and Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843–854.

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, and Kim VN (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425:415–419.

Lee Y, Hur I, Park SY, Kim YK, Suh MR, and Kim VN (2006) The role of PACT in the RNA silencing pathway. *EMBO J.*, 25:522–532.

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, and Kim VN (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, 23:4051–4060.

Lee YF, Nomoto A, Detjen BM, and Wimmer E (1977) A protein covalently linked to poliovirus genome RNA. *Proc. Natl. Acad. Sci. U.S.A.*, 74:59–63.

Legendre M, Lambert A, and Gautheret D (2005) Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, 21:841–845.

Lestrade L and Weber MJ (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, 34:D158–162.

Li W and Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22:1658–1659.

Lim LP, Glasner ME, Yekta S, Burge CB, and Bartel DP (2003a) Vertebrate microRNA genes. *Science*, 299:1540.

Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, and Bartel DP (2003b) The microRNAs of Caenorhabditis elegans. *Genes Dev.*, 17:991–1008.

Lin SY, Johnson SM, Abraham M, Vella MC, Pasquinelli A, Gamberi C, Gottlieb E, and Slack FJ (2003) The C elegans hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. *Dev. Cell*, 4:639–650.

Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, Kuersten S, Tewari M, and Cuppen E (2009) Limitations and possibilities of small RNA digital gene expression profiling. *Nat. Methods*, 6:474–476.

Lobert PE, Escriou N, Ruelle J, and Michiels T (1999) A coding RNA sequence acts as a replication signal in cardioviruses. *Proc. Natl. Acad. Sci. U.S.A.*, 96:11560–11565.

Lowe TM and Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, 25:955–964.

Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, and Golub TR (2005) MicroRNA expression profiles classify human cancers. *Nature*, 435:834–838.

Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, and Wu CI (2008) The birth and death of microRNA genes in Drosophila. *Nat. Genet.*, 40:351–355.

Lu YC, Smielewska M, Palakodeti D, Lovci MT, Aigner S, Yeo GW, and Graveley BR (2009) Deep sequencing identifies new and regulated microRNAs in Schmidtea mediterranea. *RNA*, 15:1483–1491.

Lund E, Güttinger S, Calado A, Dahlberg JE, and Kutay U (2004) Nuclear export of microRNA precursors. *Science*, 303:95–98.

Lwoff A, Horne R, and Tournier P (1962) A system of viruses. *Cold Spring Harb. Symp. Quant. Biol.*, 27:51–55.

Lück R, Gräf S, and Steger G (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res.*, 27:4208–4217.

Manak JR, Dike S, Sementchenko V, Kapranov P, Biemar F, Long J, Cheng J, Bell I, Ghosh S, Piccolboni A, and Gingeras TR (2006) Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat. Genet.*, 38:1151–1158.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380.

Mason PW, Bezborodova SV, and Henry TM (2002) Identification and characterization of a cis-acting replication element (cre) adjacent to the internal ribosome entry site of foot-and-mouth disease virus. *J. Virol.*, 76:9686–9694.

Mathews DH (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10:1178–1190.

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, and Turner DH (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, 101:7287–7292.

Mathews DH, Sabina J, Zuker M, and Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940.

McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119.

McErlean P, Shackelton LA, Andrews E, Webster DR, Lambert SB, Nissen MD, Sloots TP, and Mackay IM (2008) Distinguishing molecular features and clinical characteristics of a putative new rhinovirus species, human rhinovirus C (HRV C). *PLoS ONE*, 3:e1847.

McErlean P, Shackelton LA, Lambert SB, Nissen MD, Sloots TP, and Mackay IM (2007) Characterisation of a newly identified human rhinovirus, HRV-QPM, discovered in infants with bronchiolitis. *J. Clin. Virol.*, 39:67–75.

McKnight KL and Lemon SM (1996) Capsid coding sequence is required for efficient replication of human rhinovirus 14 RNA. *J. Virol.*, 70:1941–1952.

McKnight KL and Lemon SM (1998) The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *RNA*, 4:1569–1584.

Meyer IM (2007) A practical guide to the art of RNA gene prediction. *Brief. Bioinformatics*, 8:396–414.

Monier A, Claverie JM, and Ogata H (2008) Taxonomic distribution of large DNA viruses in the sea. *Genome Biol.*, 9:R106.

Montgomerie S, Sundararaj S, Gallin WJ, and Wishart DS (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics*, 7:301.

Moss EG, Lee RC, and Ambros V (1997) The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. *Cell*, 88:637–646.

Mourelatos Z, Dostie J, Paushkin S, Sharma A, Charroux B, Abel L, Rappsilber J, Mann M, and Dreyfuss G (2002) miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.*, 16:720–728.

Nagai K, Oubridge C, Kuglstatter A, Menichelli E, Isel C, and Jovine L (2003) Structure, function and evolution of the signal recognition particle. *EMBO J.*, 22:3479–3485.

Nagaswamy U, Voss N, Zhang Z, Fox GE, and Fox GE (2000) Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Res.*, 28:375–376.

Nam JW, Kim J, Kim SK, and Zhang BT (2006) ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Res.*, 34:W455–458.

Nam JW, Shin KR, Han J, Lee Y, Kim VN, and Zhang BT (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.*, 33:3570–3581.

Ng KL and Mishra SK (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 23:1321–1330.

Ng Kwang Loong S and Mishra SK (2007) Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification. *RNA*, 13:170–187.

Notredame C, Higgins DG, and Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302:205–217.

Nussinov R and Jacobson AB (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. U.S.A.*, 77:6309–6313.

O'Connell RM, Taganov KD, Boldin MP, Cheng G, and Baltimore D (2007) MicroRNA-155 is induced during the macrophage inflammatory response. *Proc. Natl. Acad. Sci. U.S.A.*, 104:1604–1609.

Ohler U, Yekta S, Lim LP, Bartel DP, and Burge CB (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA*, 10:1309–1322.

Okada C, Yamashita E, Lee SJ, Shibata S, Katahira J, Nakagawa A, Yoneda Y, and Tsukihara T (2009) A high-resolution structure of the pre-microRNA nuclear export machinery. *Science*, 326:1275–1279.

Okamura K, Chung WJ, Ruby JG, Guo H, Bartel DP, and Lai EC (2008) The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature*, 453:803–806.

Okamura K, Hagen JW, Duan H, Tyler DM, and Lai EC (2007) The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell*, 130:89–100.

Olsen PH and Ambros V (1999) The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.*, 216:671–680.

Oulas A, Boutla A, Gkirtzou K, Reczko M, Kalantidis K, and Poirazi P (2009) Prediction of novel microRNA genes in cancer-associated genomic regions–a combined computational and experimental approach. *Nucleic Acids Res.*, 37:3276–3287.

Pace NR, Thomas BC, and Woese CR (1999) Probing RNA structure, function, and history by comparative anaylsis. In: Gesteland RF, Cech TR, and Atkins JF (Editors), *The RNA World*, pp. 113–141. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Palmenberg AC, Spiro D, Kuzmickas R, Wang S, Djikeng A, Rathe JA, Fraser-Liggett CM, and Liggett SB (2009) Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. *Science*, 324:55–59.

Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, and Hatzigeorgiou AG (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, 37:D155–158.

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Müller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, et al. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408:86–89.

Paul AV, Rieder E, Kim DW, van Boom JH, and Wimmer E (2000) Identification of an RNA hairpin in poliovirus RNA that serves as the primary template in the in vitro uridylylation of VPg. *J. Virol.*, 74:10359–10370.

Paul AV, van Boom JH, Filippov D, and Wimmer E (1998) Protein-primed RNA synthesis by purified poliovirus RNA polymerase. *Nature*, 393:280–284.

Paul AV, Yin J, Mugavero J, Rieder E, Liu Y, and Wimmer E (2003) A "slide-back" mechanism for the initiation of protein-primed RNA synthesis by the RNA polymerase of poliovirus. *J. Biol. Chem.*, 278:43951–43960.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, and Haussler D (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, 2:e33.

Perry MM, Moschos SA, Williams AE, Shepherd NJ, Larner-Svensson HM, and Lindsay MA (2008) Rapid changes in microRNA-146a expression negatively regulate the IL-1beta-induced inflammatory response in human lung alveolar epithelial cells. *J. Immunol.*, 180:5689–5698.

Persson H, Kvist A, Vallon-Christersson J, Medstrand P, Borg A, and Rovira C (2009) The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nat. Cell Biol.*, 11:1268–1271.

Pilipenko EV, Maslova SV, Sinyakov AN, and Agol VI (1992) Towards identification of cis-acting elements involved in the replication of enterovirus and rhinovirus RNAs: a proposal for the existence of tRNA-like terminal structures. *Nucleic Acids Res.*, 20:1739–1745.

Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, Macdonald PE, Pfeffer S, Tuschl T, Rajewsky N, Rorsman P, and Stoffel M (2004) A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, 432:226–230.

Pudil P, Novovičová J, and Kittler J (1994) Floating search methods in feature selection. *Pattern Recognit. Lett.*, 15:1119–1125.

Quinlan R (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco , CA.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Software available at: `http://www.R-project.org`.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, and Ruvkun G (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403:901–906.

Rice P, Longden I, and Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, 16:276–277.

Rieder E, Paul AV, Kim DW, van Boom JH, and Wimmer E (2000) Genetic and biochemical studies of poliovirus cis-acting replication element cre in relation to VPg uridylylation. *J. Virol.*, 74:10371–10380.

Ritchie W, Théodule FX, and Gautheret D (2008) Mireval: a web tool for simple microRNA prediction in genome sequences. *Bioinformatics*, 24:1394–1396.

Rivas E and Eddy SR (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16:583–605.

Rivas E and Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8.

Rivera VM, Welsh JD, and Maizel JV (1988) Comparative sequence analysis of the 5' noncoding region of the enteroviruses and rhinoviruses. *Virology*, 165:42–50.

Rodriguez A, Griffiths-Jones S, Ashurst JL, and Bradley A (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, 14:1902–1910.

Rohll JB, Moon DH, Evans DJ, and Almond JW (1995) The 3' untranslated region of picornavirus RNA: features required for efficient genome replication. *J. Virol.*, 69:7835–7844.

Rosenblatt F (1958) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychol Rev*, 6:386–408.

Rothberg PG, Harris TJ, Nomoto A, and Wimmer E (1978) O4-(5'-uridylyl)tyrosine is the bond between the genome-linked protein and the RNA of poliovirus. *Proc. Natl. Acad. Sci. U.S.A.*, 75:4868–4872.

Ruby JG, Jan CH, and Bartel DP (2007a) Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448:83–86.

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, and Lai EC (2007b) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res.*, 17:1850–1864.

Rumelhart DE, Hinton GE, and Williams RJ (1986) Learning Internal Representations by Error-Propagation. In: Rumelhart DE and McClelland RJ (Editors), *Parallel Distributed Processing*, chap. 8. MIT Press, Cambridge, MA.

Sachdeva M, Zhu S, Wu F, Wu H, Walia V, Kumar S, Elble R, Watabe K, and Mo YY (2009) p53 represses c-Myc through induction of the tumor suppressor miR-145. *Proc. Natl. Acad. Sci. U.S.A.*, 106:3207–3212.

Saebø PE, Andersen SM, Myrseth J, Laerdahl JK, and Rognes T (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.*, 33:W535–539.

Sai Lakshmi S and Agrawal S (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.*, 36:D173–177.

Saito K, Ishizuka A, Siomi H, and Siomi MC (2005) Processing of pre-microRNAs by the Dicer-1-Loquacious complex in Drosophila cells. *PLoS Biol.*, 3:e235.

Sankoff D (1985) Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM Journal on Applied Mathematics*, 45.

Saraiya AA and Wang CC (2008) snoRNA, a novel precursor of microRNA in Giardia lamblia. *PLoS Pathog.*, 4:e1000224.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, et al. (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 38:D5–D16.

Schneider TD and Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18:6097–6100.

Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, and Mead D (2008) Assembly of viral metagenomes from yellowstone hot springs. *Appl. Environ. Microbiol.*, 74:4164–4174.

Schwarz DS, Hutvágner G, Du T, Xu Z, Aronin N, and Zamore PD (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115:199–208.

Seffens W and Digby D (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.*, 27:1578–1584.

Segura MF, Hanniford D, Menendez S, Reavie L, Zou X, Alvarez-Diaz S, Zakrzewski J, Blochin E, Rose A, Bogunovic D, Polsky D, Wei J, Lee P, Belitskaya-Levy I, Bhardwaj N, et al. (2009) Aberrant miR-182 expression promotes melanoma metastasis by repressing FOXO3 and microphthalmia-associated transcription factor. *Proc. Natl. Acad. Sci. U.S.A.*, 106:1814–1819.

Seitz H, Ghildiyal M, and Zamore PD (2008) Argonaute loading improves the 5′ precision of both MicroRNAs and their miRNA strands in flies. *Curr. Biol.*, 18:147–151.

Seitz H, Royo H, Bortolin ML, Lin SP, Ferguson-Smith AC, and Cavaillé J (2004) A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. *Genome Res.*, 14:1741–1748.

Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein MJ, Tuschl T, van Nimwegen E, and Zavolan M (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, 6:267.

Shapiro BA and Zhang K (1990) Comparing multiple RNA secondary structures using tree comparisons. *Computer Applications in the Biosciences*, 6:309–318.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, and Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309:1728–1732.

Sheng Y, Engström PG, and Lenhard B (2007) Mammalian microRNA prediction through a support vector machine model of sequence and structure. *PLoS ONE*, 2:e946.

Sing T, Sander O, Beerenwinkel N, and Lengauer T (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, 21:3940–3941.

Slack FJ, Basson M, Liu Z, Ambros V, Horvitz HR, and Ruvkun G (2000) The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. *Mol. Cell*, 5:659–669.

Soares AR, Pereira PM, Santos B, Egas C, Gomes AC, Arrais J, Oliveira JL, Moura GR, and Santos MA (2009) Parallel DNA pyrosequencing unveils new zebrafish microRNAs. *BMC Genomics*, 10:195.

Sonnhammer EL and Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, 18:619–620.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, 12:1611–1618.

Stark A, Bushati N, Jan CH, Kheradpour P, Hodges E, Brennecke J, Bartel DP, Cohen SM, and Kellis M (2008) A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands. *Genes Dev.*, 22:8–13.

Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, and Kellis M (2007a) Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res.*, 17:1865–1879.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, Hodges E, Hinrichs AS, Caspi A, et al. (2007b) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450:219–232.

Steffen P, Voss B, Rehmsmeier M, Reeder J, and Giegerich R (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22:500–503.

Steil BP and Barton DJ (2009) Cis-active RNA elements (CREs) and picornavirus RNA replication. *Virus Res.*, 139:240–252.

Taft RJ, Glazov EA, Lassmann T, Hayashizaki Y, Carninci P, and Mattick JS (2009) Small RNAs derived from snoRNAs. *RNA*, 15:1233–1240.

Talavera G and Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, 56:564–577.

Tanzer A and Stadler PF (2004) Molecular evolution of a microRNA cluster. *J. Mol. Biol.*, 339:327–335.

Tapparel C, Junier T, Gerlach D, Cordey S, Van Belle S, Perrin L, Zdobnov EM, and Kaiser L (2007) New complete genome sequences of human rhinoviruses shed light on their phylogeny and genomic features. *BMC Genomics*, 8:224.

Tapparel C, Junier T, Gerlach D, Van-Belle S, Turin L, Cordey S, Mühlemann K, Regamey N, Aubert JD, Soccal PM, Eigenmann P, Zdobnov E, and Kaiser L (2009) New respiratory enterovirus and recombinant rhinoviruses among circulating picornaviruses. *Emerging Infect. Dis.*, 15:719–726.

Tarca AL, Carey VJ, Chen XW, Romero R, and Drăghici S (2007) Machine learning and its applications to biology. *PLoS Comput. Biol.*, 3:e116.

Tatusov RL, Koonin EV, and Lipman DJ (1997) A genomic perspective on protein families. *Science*, 278:631–637.

Terai G, Komori T, Asai K, and Kin T (2007) miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity. *RNA*, 13:2081–2090.

Thompson JD, Higgins DG, and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680.

Thomson T and Lin H (2009) The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annu. Rev. Cell Dev. Biol.*, 25:355–376.

Tilghman SM (1999) The sins of the fathers and mothers: genomic imprinting in mammalian development. *Cell*, 96:185–193.

Tucker BJ and Breaker RR (2005) Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, 15:342–348.

Tyagi S, Vaz C, Gupta V, Bhatia R, Maheshwari S, Srinivasan A, and Bhattacharya A (2008) CID-miRNA: a web server for prediction of novel miRNA precursors in human genome. *Biochem. Biophys. Res. Commun.*, 372:831–834.

van der Burgt A, Fiers MW, Nap JP, and van Ham RC (2009) In silico miRNA prediction in metazoan genomes: balancing between sensitivity and specificity. *BMC Genomics*, 10:204.

van Ooij MJ, Polacek C, Glaudemans DH, Kuijpers J, van Kuppeveld FJ, Andino R, Agol VI, and Melchers WJ (2006) Polyadenylation of genomic RNA and initiation of antigenomic RNA in a positive-strand RNA virus are controlled by the same cis-element. *Nucleic Acids Res.*, 34:2953–2965.

van Zon A, Mossink MH, Scheper RJ, Sonneveld P, and Wiemer EA (2003) The vault complex. *Cell. Mol. Life Sci.*, 60:1828–1837.

Vapnik VN (1998) *Statistical Learning Theory*. Wiley-Interscience, Malden, MA.

Vapnik VN and Chervonenkis AY (1974) *Teoriya raspoznavaniya obrazov: Statisticheskie problemy obucheniya. (Russian) [Theory of pattern recognition: Statistical problems of learning]*. Nauka, Moscow.

Walter P and Blobel G (1982) Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*, 299:691–698.

Wang P, Zou F, Zhang X, Li H, Dulak A, Tomko RJ, Lazo JS, Wang Z, Zhang L, and Yu J (2009) microRNA-21 negatively regulates Cdc25A and cell cycle progression in colon cancer cells. *Cancer Res.*, 69:8157–8165.

Wang X, Zhang J, Li F, Gu J, He T, Zhang X, and Li Y (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 21:3610–3614.

Washietl S, Hofacker IL, and Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U.S.A.*, 102:2454–2459.

Weber MJ (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, 272:59–73.

Weiss GM (2004) Mining with rarity: a unifying framework. *SIGKDD Explor. Newsl.*, 6.

Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Beukeboom LW, Desplan C, Elsik CG, Grimmelikhuijzen CJ, Kitts P, Lynch JA, Murphy T, Oliveira DC, Smith CD, et al. (2010) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science*, 327:343–348.

Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, and Peterson KJ (2009) The deep evolution of metazoan microRNAs. *Evol. Dev.*, 11:50–68.

Wienholds E, Koudijs MJ, van Eeden FJ, Cuppen E, and Plasterk RH (2003) The microRNA-producing enzyme Dicer1 is essential for zebrafish development. *Nat. Genet.*, 35:217–218.

Wightman B, Ha I, and Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75:855–862.

Wilm A, Higgins DG, and Notredame C (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, 36:e52.

Workman C and Krogh A (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.*, 27:4816–4822.

Wuchty S, Fontana W, Hofacker IL, and Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–165.

Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, and Turner DH (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–14735.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, and Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434:338–345.

Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, and Lander ES (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl. Acad. Sci. U.S.A.*, 104:7145–7150.

Xu JH, Li F, and Sun QF (2008a) Identification of microRNA precursors with support vector machine and string kernel. *Genomics Proteomics Bioinformatics*, 6:121–128.

Xu P, Vernooy SY, Guo M, and Hay BA (2003) The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Curr. Biol.*, 13:790–795.

Xu Y, Zhou X, and Zhang W (2008b) MicroRNA prediction with a novel ranking algorithm based on random walks. *Bioinformatics*, 24:i50–58.

Xue C, Li F, He T, Liu GP, Li Y, and Zhang X (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6:310.

Yang Y, Rijnbrand R, McKnight KL, Wimmer E, Paul A, Martin A, and Lemon SM (2002) Sequence requirements for viral RNA replication and VPg uridylylation directed by the internal cis-acting replication element (cre) of human rhinovirus type 14. *J. Virol.*, 76:7485–7494.

Yang Y, Yi M, Evans DJ, Simmonds P, and Lemon SM (2008) Identification of a conserved RNA replication element (cre) within the 3Dpol-coding sequence of hepatoviruses. *J. Virol.*, 82:10118–10128.

Yi R, Qin Y, Macara IG, and Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.*, 17:3011–3016.

Yin J, Paul AV, Wimmer E, and Rieder E (2003) Functional dissection of a poliovirus cis-acting replication element [PV-cre(2C)]: analysis of single- and dual-cre viral genomes and proteins that bind specifically to PV-cre RNA. *J. Virol.*, 77:5152–5166.

Yogo Y and Wimmer E (1972) Polyadenylic acid at the 3'-terminus of poliovirus RNA. *Proc. Natl. Acad. Sci. U.S.A.*, 69:1877–1882.

Yousef M, Jung S, Showe LC, and Showe MK (2008) Learning from positive examples when the negative class is undetermined–microRNA gene identification. *Algorithms Mol Biol*, 3:2.

Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, and Showe MK (2006) Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*, 22:1325–1334.

Zeng L, Carter AD, and Childs SJ (2009) miR-145 directs intestinal maturation in zebrafish. *Proc. Natl. Acad. Sci. U.S.A.*, 106:17793–17798.

Zeng Y and Cullen BR (2004) Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic Acids Res.*, 32:4776–4785.

Zhang BH, Pan XP, Cox SB, Cobb GP, and Anderson TA (2006) Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.*, 63:246–254.

Zhao Y, Ransom JF, Li A, Vedantham V, von Drehle M, Muth AN, Tsuchihashi T, McManus MT, Schwartz RJ, and Srivastava D (2007) Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. *Cell*, 129:303–317.

Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52.

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31:3406–3415.

Zuker M and Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148.

Part V

APPENDIX

A PUBLICATIONS

The following pages list my publications or publication in which I participated as a co-author. Pages coming from external sources are indicated by frames enclosing the respective pages.

A.1 COMPUTATIONAL MICRORNA GENE PREDICTION

A.1.1 *miROrtho: computational survey of microRNA genes*

Gerlach D, Kriventseva EV, Rahman N, Vejnar CE, and Zdobnov EM. miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.* (2009) 37:D111–117.

A.1.1.1 *Contributions*

Gerlach et al., 2009 presents a novel pipeline for predicting miRNA genes. The method was used to annotate miRNA genes in a set of over 40 animal genomes. The data is presented in web-accessible database called `miROrtho`: `http://cegg.unige.ch/mirortho`.

I designed the study, participated in all parts of the research, and drafted the manuscript. The web-interface was developed in collaboration with Nazim Rahman. The orthology assignment step of the pipeline was done by Evgenia Kriventseva with my feedback on the results.

A.1.1.2 *Main paper*

See pages 132–138 or at:
`http://nar.oxfordjournals.org/cgi/content/short/gkn707v1`

Published online 15 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D111–D117***
doi:10.1093/nar/gkn707

miROrtho: computational survey of microRNA genes

Daniel Gerlach[1,2], Evgenia V. Kriventseva[1], Nazim Rahman[1], Charles E. Vejnar[1,2] and Evgeny M. Zdobnov[1,2,3,*]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, [2]Swiss Institute of Bioinformatics, 1 Rue Michel-Servet, 1211 Geneva, Switzerland and [3]Imperial College London, South Kensington Campus, SW7 2AZ London, UK

Received August 19, 2008; Revised September 26, 2008; Accepted September 29, 2008

ABSTRACT

MicroRNAs (miRNAs) are short, non-protein coding RNAs that direct the widespread phenomenon of post-transcriptional regulation of metazoan genes. The mature ~22-nt long RNA molecules are processed from genome-encoded stem-loop structured precursor genes. Hundreds of such genes have been experimentally validated in vertebrate genomes, yet their discovery remains challenging, and substantially higher numbers have been estimated. The miROrtho database (http://cegg.unige.ch/mirortho) presents the results of a comprehensive computational survey of miRNA gene candidates across the majority of sequenced metazoan genomes. We designed and applied a three-tier analysis pipeline: (i) an SVM-based *ab initio* screen for potent hairpins, plus homologs of known miRNAs, (ii) an orthology delineation procedure and (iii) an SVM-based classifier of the ortholog multiple sequence alignments. The web interface provides direct access to putative miRNA annotations, ortholog multiple alignments, RNA secondary structure conservation, and sequence data. The miROrtho data are conceptually complementary to the miRBase catalog of experimentally verified miRNA sequences, providing a consistent comparative genomics perspective as well as identifying many novel miRNA genes with strong evolutionary support.

INTRODUCTION

MicroRNAs (miRNAs) represent an abundant class of short non-protein coding RNAs that direct post-transcriptional regulation of metazoan genes through repression of mRNA translation or transcript degradation. Since their initial discovery in *Caenorhabditis elegans*, the roles of miRNAs have been recognized as a widespread phenomenon, implicated in processes such as cell differentiation and cancer (1–6). Intensive studies have begun to unravel the mechanisms and characteristics of these single-stranded, ~22-nt long RNA molecules that are processed from genome-encoded precursor genes with a defining stem-loop RNA structure. Nevertheless, the discovery and characterization of novel miRNA genes have proved to be challenging both experimentally and computationally, and the miRNA gene repertoire therefore remains largely unexplored. The human genome tops the fast growing number of miRNA genes, with several hundreds now cataloged in the miRBase database of published miRNA sequences (7) and many more estimated (8,9).

The high-throughput experimental approaches usually identify only the short mature segments of the miRNA genes along with other types of endogenous small RNAs (10,11) and degradation products of mRNAs or structural RNAs. Robust computational post-processing of the experimentally derived sequences is therefore essential to identify the underlying miRNA genes. The widely applied discriminatory requirement of a characteristic stem-loop structure for the putative precursor is, however, insufficient as hairpin structures are common in eukaryotic genomes and are not a unique feature of miRNAs (12). Nonetheless, the rapid accumulation of genome-wide sequencing data provides another line of evolutionary evidence from comparative sequence analyses.

Computational screening methods that rely heavily on sequence conservation criteria, such as MirScan (13), were among the first to appear. These characteristically exhibit high specificity [e.g. predicting 35 new miRNA candidates in *C. elegans* (13) and 107 in human (14), many of which were experimentally confirmed], but their sensitivity, the ability to predict novel or divergent homologs in other organisms, is low. Methods that relax sequence conservation requirements in favor of conservation patterns specific to miRNAs (such as a more diverged loop sequence and a more conserved hairpin stem) gained

*To whom correspondence should be addressed. Tel: +41 22 379 59 73; Fax: +41 22 379 57 06; Email: evgeny.zdobnov@unige.ch

D112 *Nucleic Acids Research, 2009, Vol. 37, Database issue*

substantially higher sensitivity, e.g. Snarloop has been used to predict 214 candidate miRNAs in *C. elegans* (15) and miRSeeker (16) to predict 48 candidate miRNAs in *Drosophila melanogaster*. A similar approach was proposed that takes into account the shapes of conservation patterns of known miRNAs, e.g. phylogenetic shadowing (17,18). The first 7 nt from the second position of the 5′-end of the mature miRNA, termed the seed sequence, are presumed to be critical for the interaction between the miRNA and its targets (19–22). The intra-species abundance or inter-species conservation of such potential seeds have also been proposed as alternative starting points for miRNA gene hunting (23,24).

Secondary structure thermodynamic stability is another important characteristic that can be used to distinguish miRNAs from other hairpins (25). The recently developed software RNAz combines thermodynamic stability and conservation of secondary structure to predict non-coding RNAs (26) from multiple alignments of orthologous regions. Methods relying on phylogenetic conservation of miRNA structure and sequence are by definition restricted in terms of their predictive power. To overcome this limitation, several groups have developed *ab initio* approaches (12,27–32) to predict novel, non-conserved genes. However, these approaches often suffer from high rates of false positives.

Aiming to fuel further studies of microRNA'omes, we present here the database of computationally derived miRNA gene candidates using a novel comparative genomics approach coupled with machine-learning techniques that we consistently applied to a comprehensive set of available metazoan genomes. The three-tier pipeline consists of: (i) a custom designed SVM-based *ab initio* predictor, plus screening for known miRNA homologs, (ii) an orthology delineation procedure and (iii) an SVM-based classifier of the multiple sequence alignments of the putative orthologs. These data are conceptually complementary to the miRBase catalog of experimentally verified miRNA sequences (7). High-throughput experimental exploration of small RNAs requires rigorous follow-up bioinformatic analyses to claim evidence of microRNA genes. Decoupling experimental and bioinformatics approaches, the miROrtho data effectively provide independent supporting evidence for the numerous ongoing experimental interrogations of microRNA'omes.

MATERIAL AND METHODS

Ab initio predictors

The first tier of our analysis pipeline is a novel *ab initio* miRNA prediction procedure. We scanned the genomic sequences using RNALfold (33) for locally stable hairpins characteristic of miRNA precursors, requiring a length of 60–120 nt, a minimum free-folding energy less than −15 kcal/mol, a stem of 20–60 base pairs, a maximal interior loop size of 8 nt, and a maximum bulge loop size of 5 nt. The loop, however, was allowed to include short stem-loops e.g. hsa-let-7b. Those properties accommodate the vast majority of experimentally validated miRNAs (although there are exceptions, e.g. dme-mir-31b and dme-mir-1017). As stem-loop structures are abundant and not exclusive to miRNA genes, this step yields hundreds of millions of candidates: 1.3 million for the ∼170 Mb genome of fruitfly *Drosophila melanogaster*. The availability of many experimentally validated miRNAs revealed that although there are biases in biophysical properties of miRNA stem-loops in comparison to non-miRNA sequences, such as higher thermodynamic stability (25), no clear discriminatory features have yet been identified. We investigated a number of the most discriminating features, such as the minimum free-energy index (34) or the mean base pair distance in the ensemble of structures, and trained an SVM (support vector machine) classifier using LIBSVM (http://www.csie.ntu.edu.tw/∼cjlin/libsvm). The total number of features used for this first SVM was 253. The radial basis function kernel (RBF) was used on 1000 experimentally verified animal pre-miRNAs from miRBase (7) and a negative set of 3000 potent stem-loops from other confirmed ncRNAs [Rfam (35)]. Optimal parameters for the RBF kernel (C-SVC c = 2.0, gamma = 0.03 125) were estimated using a heuristic approach implemented in grid.py, which is a part of the LIBSVM package. A non-redundant training dataset was compiled using CD-Hit-EST (36) at a cutoff of 90% sequence identity. We tested the performance of the SVM on a test set of 237 miRNA sequences and 568 non-miRNA stem-loops which where not used for training the SVM model. Using the SVM posterior probability cutoff at 0.5, the accuracy was estimated to be 95.03%, the area under the ROC curve (receiver operating characteristic) was 0.984, corresponding to a sensitivity and specificity of 0.84 and 0.97, respectively. Using a 10-fold cross-validation procedure on the training data, we received an average AUC (area under the ROC curve) of 0.982. If the potent hairpins had >70% sequence overlap at the same locus, the one with the lower SVM score was discarded.

This single sequence SVM filter allows the space of likely candidates to be reduced by about 95%, yet still yields rather high numbers of gene candidates: 42 000 for *D. melanogaster*. The miRNA structure itself is likely to contribute to these elevated numbers: miRNAs have complementary arms in their stem-loop structure and the reverse complement of a precursor often also folds into a stable RNA hairpin. Nevertheless, we did not explicitly require a choice between the sense and the anti-sense candidates (if both of them passed the other filters) as there is evidence of miRNA loci with both strands yielding a functional miRNA, e.g. dme-mir-iab-4 and dme-mir-iab-4as.

Homology-based predictor

Screening for homologs of currently known miRNAs (miRBase 11.0) captures putative miRNAs that either did not pass the stem-loop screen, e.g. 13 (8%) of known *D. melanogaster* miRNAs, or failed the *ab initio* SVM classification, another 19 (13%). Our procedure initially performs a WU-BLAST (http://blast.wustl.edu) search using the default parameters, plus the DUST

filter and the hspsepSmax = 30 option, which defines the maximal separating distance between two high score pairs to allow for a varying loop while still matching the better conserved 5′ and 3′ arms. Next, blast hits longer than 20 nt are extended at both ends to match the length of the query sequence. These hits are further filtered using a minimum free energy filter (≤ -15 kcal/mol) and a RANDFOLD (25) filter ($P \leq 0.05$ on 100 sequence randomizations). We investigated the RNAshapes (37) filter, which predicts the probability of a sequence to fold into a simple stem-loop like structure, but it was not employed as several known miRNAs, e.g. hsa-let-7a-1, would not pass the filter. The candidate miRNAs were then aligned to the query sequence using MAFFT (38) and the conservation of the seed region was calculated by mapping the known mature miRNA region on the query miRNA to the alignment. The hits were then tested for the following criteria: a 100% conserved seed region, >90% conservation of the putative mature part, and a total hairpin identity >65%. As close paralogs (like hsa-let-7, mmu-let-7, etc) can map to the same locus when searched again one genome (e.g. the chimp), the matches were then clustered using GALAXY (http://main.g2.bx.psu.edu) and choosing one representative with the lowest *e*-value of all queries.

Orthology delineation

Groups of likely orthologous genes were automatically identified using a strategy employed previously for protein-coding genes (39) based on all-against-all sequence comparisons using the ParAlign algorithm (40) with NT2 substitution matrix; followed by clustering of best reciprocal hits (BRHs) from highest scoring ones to 10^{-6} *e*-value cutoff for triangulating BRHs or 10^{-10} cutoff for unsupported BRHs, and requiring a sequence alignment overlap of at least 20 nt across all members of a group. Furthermore, the orthologous groups were expanded by genes that are more similar to each other within a genome than to any gene in any of the other species, and by very similar copies that share over 97% sequence identity, which were identified initially using CD-Hit (36). The orthology filter allowed us to reduce the space of the miRNA candidates by a further 92%. Passing the orthology filter provides evolutionary support for the predicted miRNAs; however, detailed inspection highlighted the need for further rigorous sequence classification to remove questionable predictions.

Multi-species conservation classifier

We further analyzed the R-COFFEE (41) multiple sequence alignments of orthologous groups of putative miRNA sequences. From the alignments we gathered the 13 most descriptive features for conservation properties of sequence, energy and structures such as: GC content, number of taxa, mean pairwise sequence identity, number of consistent mutations, conservation of the mature part, etc. Those descriptors were chosen among a larger set of features, in order to optimally describe the typical conservation profile of a miRNA gene family and to reduce false positive predictions. Alignments that mapped to at least one known miRNA from miRBase 11.0 were used as the positive training and testing sets (344 and 100 alignments, respectively). Among those alignments which did not map to any known miRNA family, we randomly selected (with manual checking) the negative training and testing sets (344 and 100 alignments, respectively). The GIST SVM software package (http://www.cs.columbia.edu/compbio) was used for training, testing and classification using the default parameter. The final set of newly predicted miRNAs based on the alignment SVM was selected from all alignments which had SVM score $\geqslant 0.5$, a 100% conserved seed, a mature part >90% conserved and having representatives in at least four taxa. Performance estimation of the alignment SVM on the independent test set showed an accuracy of 91%, with the area under the ROC curve (AUC) of 0.97, and sensitivity and specificity of 0.9 and 0.92, respectively. The AUC for the 10-fold cross validation using the training data averaged to 0.998. The alignment SVM filter allowed us to reduce the space of the miRNA candidates by a further 98%, followed limited manual curation of novel miRNA candidates. We further analyzed the multiple alignments of novel miRNAs (without known homologs) to predict the mature part using a sliding 23-nt long sliding window and scanning for the region with the highest information content in the 5′ or the 3′ arms. The predictions, however, should be taken with caution without further experimental support.

DATABASE CONTENT

The miROrtho database (http://cegg.unige.ch/mirortho) presents computationally predicted putative miRNA genes for a comprehensive set of sequenced animal genomes (selection of genomes in Table 1), employing an in-house developed pipeline combining SVM-based classifiers and orthology delineation procedure adapted from OrthoDB (39). The alignments shown on the website were calculated using R-COFFEE (41), which combines MUSCLE (42), Probcons4RNA (43), MAFFT (38) and the secondary structures predicted by RNAplfold (33). Based on these alignments consensus secondary structures color-coded according to consistent/compensatory mutation were calculated using RNAalifold (44) which incorporates a ribosome scoring matrix suited for aligned RNA sequences. The database aims to provide a comprehensive comparative perspective on the animal repertoire of miRNA genes with direct reference to the putative ortholog multiple alignments, RNA secondary structure conservation, etc. As there seem to be numerous lineage specific miRNAs and miRNA-like sequences that are difficult to differentiate without experimental evidence, we see miROrtho as complementary to miRBase, the repository of experimentally verified miRNA sequences. Overall, miROrtho contains 7887 putative miRNA genes that are homologous to known miRNAs in miRBase 11.0, and 1437 confident predictions that are as yet without experimental support or homology to known miRNAs. Most experimental surveys provide support for mature miRNA sequences, while the identities of the underlying miRNA precursor genes remain somewhat uncertain.

D114 *Nucleic Acids Research, 2009, Vol. 37, Database issue*

Table 1. Analyzed genomes

Species name	Abbreviation	Size (Mb)	Number of miRNA genes			Source
			Homologs[a]	New[b]	miRBase 11.0	
Aedes aegypti	Aaeg	1384	58	1	0	AaegL1
Anopheles gambiae	Agam	273	55	1	45	AgamP3
Apis mellifera	Amel	235	60	1	54	Amel_4.0
Bombyx mori	Bmor	397	33	0	21	SW_scaffold_ge2k
Caenorhabditis elegans	Cele	100	149	0	154	WB170
Canis familiaris	Cfam	2532	383	138	203	CanFam 2.0
Ciona intestinalis	Cint	173	25	0	34	JGI2
Danio rerio	Drer	1626	324	22	337	ZFISH6
Drosophila ananassae	Dana	230	108	12	0	CAF1
Drosophila erecta	Dere	152	136	16	0	CAF1
Drosophila grimshawi	Dgri	200	110	13	0	CAF1
Drosophila melanogaster	Dmel	129	153	15	152	CAF1
Drosophila mojavensis	Dmoj	194	98	14	0	CAF1
Drosophila persimilis	Dper	188	108	16	0	CAF1
Drosophila pseudoobscura	Dpse	153	106	15	76	CAF1
Drosophila sechellia	Dsec	167	139	16	0	CAF1
Drosophila simulans	Dsim	142	131	15	0	CAF1
Drosophila virilis	Dvir	206	101	14	0	CAF1
Drosophila willistoni	Dwil	237	112	12	0	CAF1
Drosophila yakuba	Dyak	169	135	16	0	CAF1
Gallus gallus	Ggal	1100	168	49	149	WASHUC2
Gasterosteus aculeatus	Gacu	462	320	12	0	BROAD S1
Homo sapiens	Hsap	3665	626	151	678	NCBI36
Macaca mulatta	Mmul	3097	530	145	464	MMUL_1
Monodelphis domestica	Mdom	3606	205	82	119	monDom5
Mus musculus	Mmus	2661	505	117	472	NCBIM36
Ornithorhynchus anatinus	Oana	2073	207	57	0	Oana-5.0
Pan troglodytes	Ptro	3524	546	147	100	PanTro 2.1
Rattus norvegicus	Rnor	2719	440	110	287	RGSC 3.4
Strongylocentrotus purpuratus	Surc	907	13	0	0	Spur_v2.1
Takifugu rubripes	Trub	393	250	13	131	FUGU4
Tetraodon nigroviridis	Tnig	402	282	14	132	TETRAODON7
Tribolium castaneum	Tcas	200	37	1	0	Tcas_2.0
Xenopus tropicalis	Xtro	1511	351	24	184	JGI4.1

[a]Homologs to miRBase 11.0 miRNAs.
[b]New predictions that do not show any homology to any annotated miRNA.

In contrast, computational procedures rely on recognizing characteristic sequence and structural properties of the precursors, where even approximate prediction of mature miRNAs is rarely possible. This complementarity extends further, where computational predictions at different stringencies can either be used to prioritize experimental verification, or as direct independent support of miRNAs identified through high throughput experimental screens. Although miRBase accepts annotation of very close homologs of experimentally supported miRNAs, the comparative perspective is heavily biased towards favorite experimental model species. Such a bias is avoided in miROrtho through the consistent application of the same procedures across all the available genomes, delineating groups of orthologous miRNAs over distantly related organisms. The miROrtho methodology has also been applied to the task of miRNA gene annotation in a number of ongoing initial genome analyses, and this database will provide the supporting information for these predictions.

It should be noted that there is still no defining feature that clearly discriminates between *bona fide* miRNA precursors and other abundant genomic sequences capable of similar hairpin folding. Classification filters will therefore inevitably suffer from false negatives and false positives (see Materials and Methods section for estimates), leading to errors at each step along the pipeline. Even the most inclusive initial screen for locally stable stem-loop structures misses some miRNAs reported in miRBase as experimentally validated (e.g. dme-mir-1017). Despite the strict 97% specificity of our *ab initio* SVM, the abundance of false positives is clear and overloads the orthology filter. Computational methods developed for miRNA gene discovery are constantly improving, and will continue to do so as our knowledge of experimentally validated miRNAs grows.

WEB INTERFACE

The miROrtho database presents all predicted miRNA genes within the context of family groups of orthologous miRNAs. For each such family, we provide (Figure 1): (i) a table of annotated miRNA names and genomic coordinates, (ii) a multiple alignment of the miRNA sequences displaying RNA structure conservation, (iii) the minimum

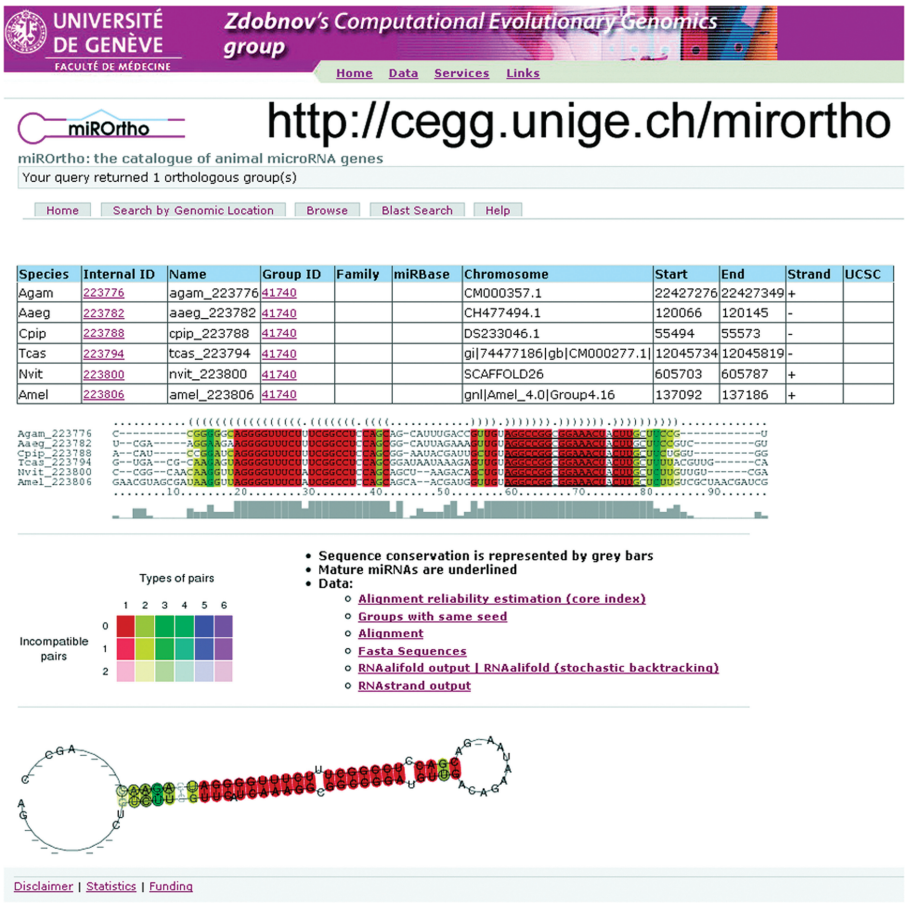


Figure 1. miROrtho screenshot showing a novel miRNA gene family. The results page consists of three parts: (i) a table with detailed information about the individual miRNAs; (ii) a multiple sequence alignment with the consensus secondary structure displayed above in dot-bracket format and conservation profile bars displayed below, with the sequence of the mature miRNAs underlined; (iii) the consensus secondary structure of the orthologous sequences. Both alignment and consensus secondary structure are color-coded according to consistent and compensatory base changes.

energy consensus miRNA hairpin fold, (iv) FASTA sequences and multiple alignment files. Color coding of the alignments and the depicted folds enables clear visualization of compensatory and consistent mutations within a given miRNA family. The mature miRNA sequences are underlined: as annotated in miRBase for known miRNAs or as predicted for novel families. Furthermore, we provide detailed folding information of individual pre-miRNAs including minimum free energy folding, the partition function folding and the centroid structure of the stem-loop. Three images show the secondary structure of a single pre-miRNA with the mature part annotated in red, color-coded according to base pairing probabilities and positional entropy per position. The data can be browsed by the species tree, or can be queried by annotation such as known families (e.g. let-7), identifiers or chromosomes. The predictions can be also searched by sequence homology using WU-BLAST (http://blast.wustl.edu).

ACKNOWLEDGEMENTS

We thank R.M. Waterhouse for help with the article, and the Vital-IT facility (http://www.vital-it.ch/vitalit-intro.htm). We would also like to acknowledge the sequencing centers that made the genome sequences that were used for this study, available before publication: The Baylor College of Medicine (www.hgsc.bcm.tmc.edu), the Washington University School of Medicine

(genome.wustl.edu), the Broad Institute (www.broad.mit.edu), the J. Craig Venter Institute (www.jcvi.org), the DOE Joint Genome Institute (www.jgi.doe.gov), the Sanger Center (www.sanger.ac.uk), the Institute for Genomic Research (www.tigr.org), Celera Genomics (www.celera.com), and Genoscope (www.genoscope.cns.fr).

FUNDING

Swiss National Science Foundation (SNF PDFMA3-118375 and 3100A0-112588). Funding for open access charges: Swiss National Science Foundation (SNF 3100A0-112588).

Conflict of interest statement. None declared.

REFERENCES

1. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
2. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
3. Du,T. and Zamore,P.D. (2005) microPrimer: the biogenesis and function of microRNA. *Development*, **132**, 4645–4652.
4. Calin,G.A. and Croce,C.M. (2006) MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, **6**, 857–866.
5. Zhang,B., Pan,X., Cobb,G.P. and Anderson,T.A. (2007) microRNAs as oncogenes and tumor suppressors. *Dev. Biol.*, **302**, 1–12.
6. Barbarotto,E., Schmittgen,T.D. and Calin,G.A. (2008) MicroRNAs and cancer: profile, profile, profile. *Int. J. Cancer*, **122**, 969–977.
7. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
8. Miranda,K.C., Huynh,T., Tay,Y., Ang,Y.S., Tam,W.L., Thomson,A.M., Lim,B. and Rigoutsos,I. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.
9. Berezikov,E., van Tetering,G., Verheul,M., van de Belt,J., van Laake,L., Vos,J., Verloop,R., van de Wetering,M., Guryev,V., Takada,S. *et al.* (2006) Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res.*, **16**, 1289–1298.
10. Kim,V.N. and Nam,J.W. (2006) Genomics of microRNA. *Trends Genet.*, **22**, 165–173.
11. Aravin,A. and Tuschl,T. (2005) Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett.*, **579**, 5830–5840.
12. Bentwich,I., Avniel,A., Karov,Y., Aharonov,R., Gilad,S., Barad,O., Barzilai,A., Einat,P., Einav,U., Meiri,E. *et al.* (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.*, **37**, 766–770.
13. Lim,L.P., Lau,N.C., Weinstein,E.G., Abdelhakim,A., Yekta,S., Rhoades,M.W., Burge,C.B. and Bartel,D.P. (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev.*, **17**, 991–1008.
14. Lim,L.P., Glasner,M.E., Yekta,S., Burge,C.B. and Bartel,D.P. (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.
15. Grad,Y., Aach,J., Hayes,G.D., Reinhart,B.J., Church,G.M., Ruvkun,G. and Kim,J. (2003) Computational and experimental identification of C. elegans microRNAs. *Mol. Cell*, **11**, 1253–1263.
16. Lai,E.C., Tomancak,P., Williams,R.W. and Rubin,G.M. (2003) Computational identification of Drosophila microRNA genes. *Genome Biol.*, **4**, R42.
17. Boffelli,D., McAuliffe,J., Ovcharenko,D., Lewis,K.D., Ovcharenko,I., Pachter,L. and Rubin,E.M. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.
18. Berezikov,E., Guryev,V., van de Belt,J., Wienholds,E., Plasterk,R.H. and Cuppen,E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.
19. Lewis,B.P., Shih,I.H., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
20. Doench,J.G. and Sharp,P.A. (2004) Specificity of microRNA target selection in translational repression. *Genes Dev.*, **18**, 504–511.
21. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.
22. Stark,A., Brennecke,J., Russell,R.B. and Cohen,S.M. (2003) Identification of Drosophila MicroRNA targets. *PLoS Biol.*, **1**, E60.
23. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
24. Weaver,D.B., Anzola,J.M., Evans,J.D., Reid,J.G., Reese,J.T., Childs,K.L., Zdobnov,E.M., Samanta,M.P., Miller,J. and Elsik,C.G. (2007) Computational and transcriptional evidence for microRNAs in the honey bee genome. *Genome Biol.*, **8**, R97.
25. Bonnet,E., Wuyts,J., Rouze,P. and Van de Peer,Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.
26. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
27. Sewer,A., Paul,N., Landgraf,P., Aravin,A., Pfeffer,S., Brownstein,M.J., Tuschl,T., van Nimwegen,E. and Zavolan,M. (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, **6**, 267.
28. Xue,C., Li,F., He,T., Liu,G.P., Li,Y. and Zhang,X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.
29. Nam,J.W., Kim,J., Kim,S.K. and Zhang,B.T. (2006) ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Res.*, **34**, W455–W458.
30. Helvik,S.A., Snove,O. Jr. and Saetrom,P. (2007) Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics*, **23**, 142–149.
31. Ng,K.L. and Mishra,S.K. (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, **23**, 1321–1330.
32. Jiang,P., Wu,H., Wang,W., Ma,W., Sun,X. and Lu,Z. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, **35**, W339–W344.
33. Hofacker,I.L., Priwitzer,B. and Stadler,P.F. (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.
34. Zhang,B.H., Pan,X.P., Cox,S.B., Cobb,G.P. and Anderson,T.A. (2006) Evidence that miRNAs are different from other RNAs. *Cell Mol. Life Sci.*, **63**, 246–254.
35. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
36. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
37. Steffen,P., Voss,B., Rehmsmeier,M., Reeder,J. and Giegerich,R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.
38. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.*, **9**, 286–298.
39. Kriventseva,E.V., Rahman,N., Espinosa,O. and Zdobnov,E.M. (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.*, **36**, D271–D275.
40. Saebo,P.E., Andersen,S.M., Myrseth,J., Laerdahl,J.K. and Rognes,T. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.*, **33**, W535–W539.

41. Wilm,A., Higgins,D.G. and Notredame,C. (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, **36**, e52.
42. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
43. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
44. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

A.1.2 *The genome sequence of taurine cattle: a window to ruminant biology and evolution*

Bovine Genome Sequencing Consortium (MicroRNA analysis: Anzola JM, Gerlach D, and Zdobnov EM). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* (2009) 324:522–528.

A.1.2.1 *Contributions*

The paper by the Bovine Genome Sequencing Consortium presents the complete genome of the species *Bos taurus* (cow).

I participated in the analysis of cow miRNA genes. The predictions were performed using the `miROrtho` pipeline. Another set of miRNAs produced by another group, was merged with my predictions summing up to a total of 496 bovine miRNAs grouped into 298 families.

A.1.2.2 *Main paper*

See pages 140–146 or at:
`http://www.sciencemag.org/cgi/content/full/324/5926/522`

second model, the two main conditions were parametrically modulated by the two categories, respectively (SOM, S5.1). The activation of the precuneus was higher for hard dominance-solvable games than for easy ones (Fig. 4A and table S10). The activation of the insula was higher for the highly focal coordination games than for less focal ones (Fig. 4B and table S11). Previous studies also found that precuneus activity increased when the number of planned moves increased (*40*, *41*). The higher demand for memory-related imagery and memory retrieval may explain the greater precuneus activation in hard dominance-solvable games. In highly focal coordination games, the participants may have felt quite strongly that the pool students must notice the same salient feature. This may explain why insula activation correlates with NCI.

Participants might have disagreed about which games were difficult. We built a third model to investigate whether the frontoparietal activation correlates with how hard a dominance-solvable game is and whether the activation in insula and ACC correlates with how easy a coordination game is. Here, the two main conditions were parametrically modulated by each participant's probability of obtaining a reward in each game (SOM, S2.2 and S5.2). We found a negative correlation between the activation of the precuneus and the participant's probability of obtaining a reward in dominance-solvable games (Fig. 4C and table S12), which suggests that dominance-solvable games that yielded lower payoffs presented harder mental challenges. In a previous study on working memory, precuneus activity positively correlated with response times, a measure of mental effort (*24*). Both findings are consistent with the interpretation that subjective measures reflecting harder tasks (higher efforts) correlate with activation in precuneus. A positive correlation between insula activation and the participant's probability of obtaining a reward again suggests that coordination games with a highly salient feature strongly activated the "gut feeling" reported by many participants (Fig. 4D and table S13). A previous study found that the subjective rating of "chills intensity" in music correlates with activation of insula (*42*). Both findings are consistent with the interpretation that the subjective intensity of how salient a stimulus is correlates with activation in insula.

As mentioned, choices were made significantly faster in coordination games than in dominance-solvable games. The results of the second and third models provide additional support for the idea that intuitive and deliberative mental processes have quite different properties. The "slow and effortful" process was more heavily taxed when the dominance-solvable games were harder. The "fast and effortless" process was more strongly activated when coordination was easy.

References and Notes

1. J. Schaeffer *et al.*, *Science* **317**, 1518 (2007).
2. Previous fMRI studies of game-playing include Gallagher et al. (*43*) and Bhatt and Camerer (*44*), but they address different issues. In particular, Bhatt and Camerer found higher insula and ACC activity when comparing choices to first-order beliefs in dominance-solvable games.
3. We are considering here coordination without visual or other contact. Nonhuman primates seem able to coordinate their actions (simultaneously pulling on bars to obtain food) when they are in visual contact (*45*).
4. J. Mehta, C. Starmer, R. Sugden, *Am. Econ. Rev.* **84**, 658 (1994).
5. T. Schelling, *J. Conflict Resolution* **2**, 203 (1958), p. 211.
6. D. Kahneman, *Am. Psychol.* **58**, 697 (2003).
7. K. Stanovich, R. West, *Behav. Brain Sci.* **23**, 645 (2000).
8. A. Rubinstein, *Econ. J.* **117**, 1243 (2007).
9. See (*46*). In our experiment, the average number of steps required to find out the game-theoretic solution for all 40 dominance-solvable games is 3.675.
10. R. Jung, R. Haier, *Behav. Brain Sci.* **30**, 135 (2007).
11. V. Goel, R. Dolan, *Neuropsychologia* **39**, 901 (2001).
12. I. Noveck, V. Goel, K. Smith, *Cortex* **40**, 613 (2004).
13. M. Atherton *et al.*, *Brain Res. Cogn. Brain Res.* **16**, 26 (2003).
14. P. Kyllonen, R. Christal, *Intelligence* **14**, 389 (1990).
15. M. D'Esposito, *Philos. Trans. R. Soc. London Ser. B* **362**, 761 (2007).
16. A. Baddeley, *Nat. Rev. Neurosci.* **4**, 829 (2003).
17. In coordination games, the participant has to encode and hold this information as well. However, because the targets of both players are the same, the demand on this capacity should be smaller.
18. E. Smith, J. Jonides, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12061 (1998).
19. N. Müller, R. Knight, *Neuroscience* **139**, 51 (2006).
20. E. Smith, J. Jonides, *Science* **283**, 1657 (1999).
21. T. Wager, E. Smith, *Cogn. Affect. Behav. Neurosci.* **3**, 255 (2003).
22. M. Berryhill, I. Olson, *Neuropsychologia* **46**, 1775 (2008).
23. A. Cavanna, M. Trimble, *Brain* **129**, 564 (2006).
24. M. Wallentin, A. Roepstorff, R. Glover, N. Burgess, *Neuroimage* **32**, 1850 (2006).
25. M. Wallentin, E. Weed, L. Østergaard, K. Mouridsen, A. Roepstorff, *Hum. Brain Mapp.* **29**, 524 (2008).
26. A. D. Craig, *Nat. Rev. Neurosci.* **3**, 655 (2002).
27. A. MacDonald III, J. Cohen, A. Stenger, C. Carter, *Science* **288**, 1835 (2000).
28. J. Decety *et al.*, *Neuroimage* **23**, 744 (2004).
29. J. S. Winston *et al.*, *Nat. Neurosci.* **5**, 277 (2002).
30. T. Singer *et al.*, *Science* **303**, 1157 (2004).
31. A. Bartels, S. Zeki, *Neuroreport* **11**, 3829 (2000).
32. J. Woodward, J. Allman, *J. Physiol. (Paris)* **101**, 179 (2007).
33. A. D. Craig, *Nat. Rev. Neurosci.* **10**, 59 (2009).
34. W. Seeley *et al.*, *J. Neurosci.* **27**, 2349 (2007).
35. J. Downar, A. Crawley, D. Mikulis, K. Davis, *Nat. Neurosci.* **3**, 277 (2000).
36. J. Downar, A. Crawley, D. Mikulis, K. Davis, *J. Neurophysiol.* **87**, 615 (2002).
37. K. Davis *et al.*, *J. Neurosci.* **25**, 8402 (2005).
38. K. Taylor, D. Seminowicz, K. Davis, *Hum. Brain Mapp.*, in press; published online 15 December 2008; 10.1002/hbm.20705.
39. See (*47*). The NCI can be interpreted as the probability that two randomly chosen individuals make the same choice relative to the probability of successful coordination if all choose randomly (SOM, S2.5).
40. S. Newman, P. Carpenter, S. Varma, M. Just, *Neuropsychologia* **41**, 1668 (2003).
41. J. Fincham *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3346 (2002).
42. A. Blood, R. Zatorre, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 11818 (2001).
43. H. Gallagher, A. Jack, A. Roepstorff, C. Frith, *Neuroimage* **16**, 814 (2002).
44. M. Bhatt, C. Camerer, *Games Econ. Behav.* **52**, 424 (2005).
45. K. Mendres, F. de Waal, *Anim. Behav.* **60**, 523 (2000).
46. C. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton Univ. Press, Princeton, NJ, 2003), Chap. 5.
47. N. Bardsley, J. Mehta, C. Starmer, R. Sugden, CeDEx Discussion Paper No. 2008-17 (Centre for Decision Research and Experimental Economics, Nottingham, UK, 2008); available at www.nottinghamnetlearning.com/economics/cedex/papers/2008-17.pdf.

48. We thank M. Hsu for helpful comments on the manuscript and J.-Y. Leu, J.T.-Y. Wang, D. Niddam, and participants at many seminars for discussions. Technical assistance from C.-R. Chou, C.-T. Chen, C.-H. Lan, S.-C. Lin, K.-L. Chen, Y.-Y. Chung, W.-Y. Lin, S. Hsu, R. Chen, and the National Taiwan University Hospital MRI Laboratory is greatly appreciated. This work was supported by the National Science Council of Taiwan (grant NSC 94-2415-H-002-004).

Supporting Online Material

www.sciencemag.org/cgi/content/full/324/5926/519/DC1
Materials and Methods
Figs. S1 to S9
Tables S1 to S18
References

8 September 2008; accepted 24 February 2009
10.1126/science.1165598

The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution

The Bovine Genome Sequencing and Analysis Consortium,* Christine G. Elsik,[1] Ross L. Tellam,[2] Kim C. Worley[3]

To understand the biology and evolution of ruminants, the cattle genome was sequenced to about sevenfold coverage. The cattle genome contains a minimum of 22,000 genes, with a core set of 14,345 orthologs shared among seven mammalian species of which 1217 are absent or undetected in noneutherian (marsupial or monotreme) genomes. Cattle-specific evolutionary breakpoint regions in chromosomes have a higher density of segmental duplications, enrichment of repetitive elements, and species-specific variations in genes associated with lactation and immune responsiveness. Genes involved in metabolism are generally highly conserved, although five metabolic genes are deleted or extensively diverged from their human orthologs. The cattle genome sequence thus provides a resource for understanding mammalian evolution and accelerating livestock genetic improvement for milk and meat production.

Domesticated cattle (*Bos taurus* and *Bos taurus indicus*) provide a significant source of nutrition and livelihood to nearly 6.6 billion humans. Cattle belong to a clade phylogenetically distant from humans and rodents, the Cetartiodactyl order of eutherian mammals, which

first appeared ~60 million years ago (*1*). Cattle represent the Ruminantia, which occupy diverse terrestrial environments with their ability to efficiently convert low-quality forage into energy-dense fat, muscle, and milk. These biological processes have been exploited by humans since domestication, which began in the Near East some 8000 to 10,000 years ago (*2*). Since then, over 800 cattle breeds have been established, representing an important world heritage and a scientific resource for understanding the genetics of complex traits.

The cattle genome was assembled with methods similar to those used for the rat and sea urchin genomes (*3*, *4*). The most recent assemblies, Btau3.1 and Btau4.0, combined bacterial artificial chromosome (BAC) and whole-genome shotgun (WGS) sequences. Btau3.1 was used for gene-specific analyses. Btau4.0, which includes finished sequence data and used different mapping methods to place the sequence on chromosomes, was used for all global analyses other than gene prediction. The contig N50 (50% of the genome is in contigs of this size or greater) is 48.7 kb for both assemblies; the scaffold N50 for Btau4.0 is 1.9 Mb. In the Btau4.0 assembly, 90% of the total genome sequence was placed on the 29 autosomes and X chromosome and validated (*3*). Of 1.04 million expressed sequence tag (EST) sequences, 95.0% were contained in the assembled contigs. With an equivalent gene distribution in the remaining 5% of the genome, the estimated genome size is 2.87 Gbp. Comparison with 73 finished BACs and single-nucleotide polymorphism (SNP) linkage data (*5*, *6*) confirmed this assembly quality with greater than 92% genomic coverage, and fewer than 0.8% of SNPs were incorrectly positioned at the resolution of these maps (*3*, *4*).

We used the cattle genome to catalog protein-coding genes, microRNA (miRNA) genes, and ruminant-specific interspersed repeats, and we manually annotated over 4000 genes. The consensus protein-coding gene set for Btau3.1 (OGSv1), from six predicted gene sets (*4*), consists of 26,835 genes with a validation rate of 82% (*4*). On this basis, we estimate that the cattle genome contains at least 22,000 protein-coding genes. We identified 496 miRNA genes of which 135 were unpublished miRNAs (*4*). About half of the cattle miRNA occur in 60 genomic miRNA clusters, containing two to seven miRNA genes separated by less than 10 kbp (fig. S2). The overall GC content of the cattle genome is 41.7%, with an observed-to-expected CpG ratio of 0.234, similar to that of other mammals.

The cattle genome has transposable element classes like those of other mammals, as well as large numbers of ruminant-specific repeats (table S4) that compose 27% of its genome. The

[1]Department of Biology, 406 Reiss, Georgetown University, 37th and O Streets, NW, Washington, DC 20057, USA. E-mail: ce75@georgetown.edu [2]CSIRO Livestock Industries, 306 Carmody Road, St. Lucia, QLD 4067, Australia. E-mail: ross.tellam@csiro.au [3]Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, MS BCM226, One Baylor Plaza, Houston, TX 77030, USA. E-mail: kworley@bcm.edu

*All authors with their affiliations and contributions are listed at the end of this paper.

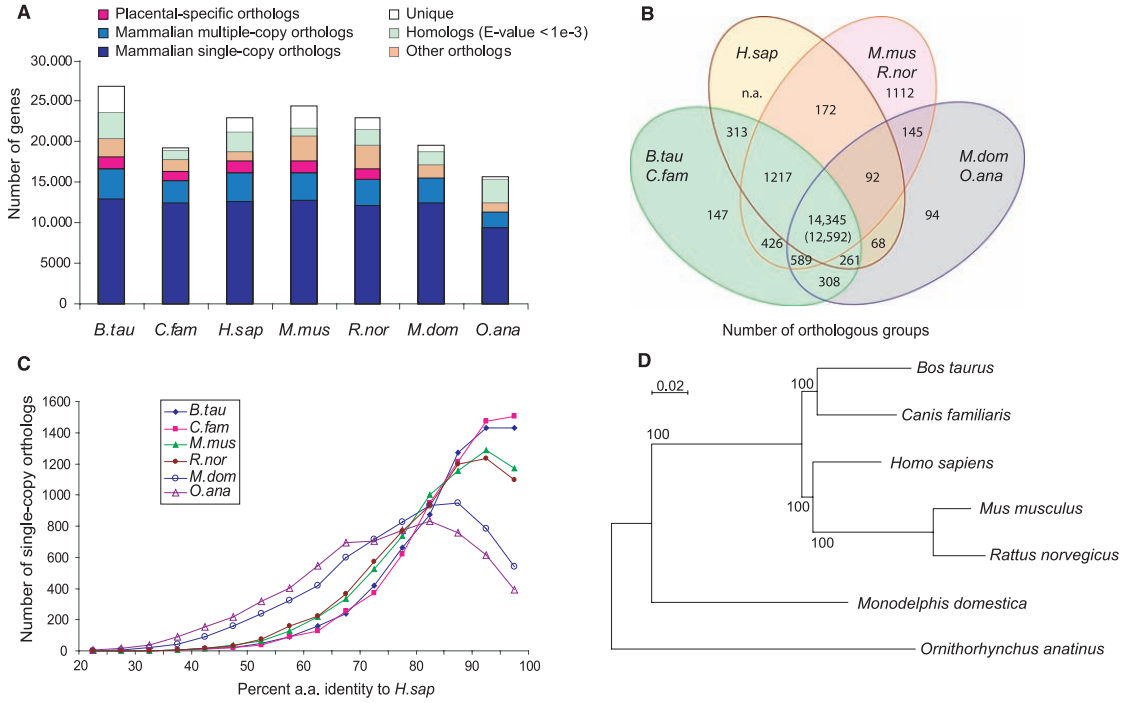


Fig. 1. Protein orthology comparison among genomes of cattle, dog, human, mouse, and rat (*Bos taurus*, *Canis familiaris*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, representing placental mammals), opossum (*Monodelphis domestica*, marsupial), and platypus (*Ornithorhynchus anatinus*, monotreme). (**A**) The majority of mammalian genes are orthologous, with more than half preserved as single copies (dark blue); a few thousand have species-specific duplications (blue); another few thousand have been lost in specific lineages (orange). We also show those lacking confident orthology assignment (green), and those that are apparently lineage specific [unique (white)]. Placental-specific orthologs are shown in pink. Single- or multiple-copy genes were defined on the basis of representatives in human, bovine, or dog; mouse or rat; and opossum or platypus. (**B**) Venn diagram showing shared orthologous groups (duplicated genes were counted as one) between laurasiatherians (cattle and dog), human, rodents (mouse and rat), and nonplacental mammals (opossum and platypus) on the basis of the presence of a representative gene in at least one of the grouped species [as in (A)]. (**C**) Distribution of ortholog protein identities between human and the other species for a subset of strictly conserved single-copy orthologs. (**D**) A maximum likelihood phylogenetic tree using all single-copy orthologs supports the accepted phylogeny and quantifies the relative rates of molecular evolution expressed as the branch lengths.

consensus sequence of Bov-B, a long interspersed nuclear element (LINE) lacked a functional open reading frame (ORF), which suggested that it was inactive (*7*). However, Bov-B repeats with intact ORF were identified in the genome, and their phylogeny (fig. S4) indicates that some are still actively expanding and evolving. Mapping chromosomal segments of high- and low-density ancient repeat content, L2/MIR [a LINE/SINE (short interspersed nuclear element) pair] and Bov-B, and more recent repeats, Bov-B/ART2A (Bov-B–derived SINE pair), revealed that the genome consists of ancient regions enriched for L2/MIR and recent regions enriched for Bov-B/ART2A (fig. S7). Exclusion of Bov-B/ART2A from contiguous blocks of ancient repeats suggests that evolution of the ruminant or cattle genome experienced invasions of new repeats into regions lacking ancient repeats. Alternatively, older repeats may have been destroyed by insertion of ruminant- or cattle-specific repeats. AGC trinucleotide repeats, the most common simple-sequence repeat (SSR) in artiodactyls (which include cattle, pigs, and sheep), are 90- and 142-fold overrepresented in cattle compared with human and dog, respectively (fig. S10). Of the AGC repeats in the cattle genome, 39% were associated with Bov-A2 SINE elements.

A comparative analysis examined the rate of protein evolution and the conservation of gene repertoires among orthologs in the genomes of dog, human, mouse, and rat (representing placental mammals); opossum (marsupial); and platypus (monotreme). Orthology was resolved for >75% of cattle and >80% of human genes (Fig. 1A). There were 14,345 orthologous groups with representatives in human, cattle, or dog; mouse or rat; and opossum or platypus, which represent 16,749 cattle and 16,177 human genes, respectively, of which 12,592 are single-copy orthologs. We also identified 1217 placental mammal–specific orthologous groups with genes present in human, cattle, or dog; mouse or rat; but not opossum or platypus. About 1000 orthologs shared between rodents and laurasiatherians (cattle and dog), many of which encode G protein–coupled receptors, appear to have been lost or may be misannotated in the human genome (Fig. 1B). Gene repertoire conservation among these mammals correlates with conservation at the amino acid–sequence level (Fig. 1C). The elevated rate of evolution in rodents relative to other mammals (*8*) was supported by the higher amino acid sequence identity between human and dog or cattle proteins relative to that between human and rodent proteins. However, maximum-likelihood analysis of amino acid substitutions in single-copy orthologs supports the accepted sister lineage relation of primates and rodents (*1*) (Fig. 1D).

Alternative splicing is a major mechanism for transcript diversification (*9*), yet the extent of its evolutionary conservation and functional impact remain unclear. We used the cattle genome to analyze the conservation of the most common form of alternative splicing, exon skipping, defined as a triplet of exons in which the middle exon is absent in some transcripts, in a set of 1930 exon-skipping events across human, mouse, dog, and cattle (*4*). We examined 277 cases, with different conservation patterns between human and mouse, in 16 different cattle tissues with reverse transcription polymerase chain reaction (*4*). These splicing events were divided into a shared set (163 in both human and mouse) and a nonshared set (114 in human but not in mouse). Of the 277, we detected exon-skipping for 188 cases in cattle (table S5), which suggested that the majority of genes with exon-skipping in human were present and regulated in cattle and that, if an event is shared between human and mouse, it was more likely to be found in cattle. It was estimated that at most 40% of exon-skipping is conserved among mammals; thus, our data agree with the upper bound from previous analyses with human and rodents [e.g., (*10*)].

We constructed a cattle-human Oxford grid (fig. S12) (*4*) to conduct synteny-based chromosomal comparisons, which reinforced that human genome organization is more similar to cattle's than rodents' because most cattle chromosomes primarily correspond to part of one human chromosome, albeit with multiple rearrangements [e.g., (*11*)]. In contrast, the cattle-mouse Oxford grid shows poorer chromosomal correspondence. Lineage-specific evolutionary breakpoints were identified for cattle, artiodactyls, and ferungulates (a group encompassing artiodactyls and carnivores, represented by cattle, pig, and dog) and are shown with cattle (fig. S11) and human sequence coordinates (Fig. 2) (*4*). Primate, dog, rodent, mouse, and rat lineage-specific breakpoint positions were similarly identified. A total of 124 evolutionary breakpoint regions (EBRs) were identified in the cattle lineage, of which 100 were cattle- or ruminant-specific and 24 were artiodactyl-specific (e.g., Fig. 2). Nine additional EBRs represent presumptive ferungulate-specific rearrangements. *Bos taurus* chromosome 16 (BTA16) is populated with four ferungulate-specific EBRs, which suggests that this region was rearranged before the Artiodactyla and Carnivora divergence (Fig. 2). Such conserved regions demonstrate that many inversions that occurred before the divergence of the carnivores and artiodactyls have probably been retained in the ancestral form within the human genome. In contrast to the cattle genome, a pig physical map identified only 77 lineage-specific EBRs. Interchromosomal rearrangements and inversions characterize most of the lineage-specific rearrangements observed in the cattle, dog, and pig genomes.

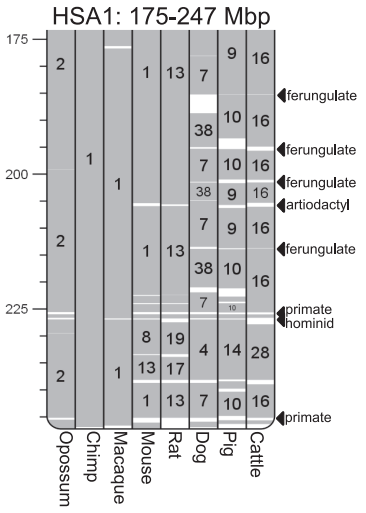


Fig. 2. Examples of EBRs. Ferungulate-, artiodactyl-, and primate-specific EBRs on HSA1 at 175 to 247 Mbp (other lineage-specific EBRs not shown). Homologous synteny blocks constructed for the macaque, chimp, cattle, dog, mouse, rat, and pig genomes were used for pairwise comparisons (*4*). White areas correspond to EBRs. Arrows to the right of the chromosome ideogram indicate positions of representative cattle-specific; artiodactyl-specific (specific to the chromosomes of pigs and cattle); ferungulate-specific (cattle, dog, and pig); primate-specific (human, macaque, and chimp); and hominoid-specific (human and chimp) rearrangements. Opossum is shown as an outgroup to the eutherian clade, which allows classification of ferungulate-specific EBRs.

Table 1. Changes in the number of genes in innate immune gene families. Many of the β-defensin genes are present in unassigned scaffolds, i.e., they are not yet part of the current assembly. The exact number of β-defensin genes is uncertain. Interferon subfamily pseudogenes predicted on the basis of frame-shift mutations or stop codons within the first 100 amino acids of the coding sequence have been excluded from the table. The IFNX genes represent a newly discovered subfamily of IFN and are so named for convenience. BPI, Bactericidal and/or permeability-increasing; RNase, ribonuclease; LBP, lipopolysaccharide-binding protein; ULBP, UL16-binding protein.

Gene family	Bovine	Human	Murine
Cathelicidin	10	1	1
RNase	21	13	25
BPI-like	13	9	11
BPI/*LBP*	3	2	2
β-Defensin	~106	39	52
Interferon subfamilies			
IFNK	1	1	1
IFNE	1	1	1
IFNB	6	1	1
IFNA	13	13	14
IFNW	24	1	0
IFNT	3	0	0
IFNX	3	0	0
IFNL	0	3	2
IFNZ	0	0	2
C-type lysozyme	10	1	3
ULBP[1]	30	3	1

[1](*31*).

An examination of repeat families and individual transposable elements within cattle-, artiodactyl- and ferungulate-specific EBRs showed a significantly higher density of LINE-L1 elements and the ruminant-specific LINE-RTE repeat family (*12*) in cattle-specific EBRs relative to the remainder of the cattle genome (table S6). In contrast, the SINE-BovA repeat family and the more ancient $tRNA^{Glu}$-derived SINE repeats (*13*) were present in lower density in cattle-specific EBRs, similar to other LINEs and SINEs (table S7). The differences in repeat densities were generally consistent in cattle-, artiodactyl- and ferungulate-specific EBRs, with the exception of the $tRNA^{Glu}$-derived and LTR-ERVL repeats, which are at higher densities in artiodactyl EBRs compared with the rest of the genome.

The $tRNA^{Glu}$-derived SINEs originated in the common ancestor of Suina (pigs and peccaries), Ruminantia, and Cetacea (whales) (*13*), which suggests that $tRNA^{Glu}$-derived SINEs were involved in ancestral artiodactyl chromosome rearrangements. Furthermore, the lower density of the more ancient repeat families in cattle-specific EBRs suggests that either more recently arising repeat elements were inserted into regions lacking ancient repeats or that older repeats were destroyed by this insertion (table S7). The repeat elements differing in density in EBRs were also found in regions of homologous synteny, which suggests that repeats may promote evolutionary rearrangements (see below). Differences in repeat density in cattle-specific EBRs are thus unlikely to be caused by the accumulation of repeats in EBRs after such rearrangements occur. We identified a cattle-specific EBR associated with a bidirectional promoter (figs. S14 and S15) that may affect control of the expression of the *CYB5R4* gene, which has been implicated in human diabetes and, therefore, may be important in the regulation of energy flow in cattle (*4*).

We identified 1020 segmental duplications (SDs) corresponding to 3.1% (94.4 Mbp) of the cattle genome (*4*). Duplications assigned to a chromosome showed a bipartite distribution with respect to length and percent identity (fig. S16), and interchromosomal duplications were shorter (median length 2.5 kbp) and more divergent (<94% identity) relative to intrachromosomal duplications (median length 20 kbp, ~97% identity) and tended to be locally clustered (fig. S17). Twenty-one of these duplications were >300 kbp and located in regions enriched for tandem duplications (e.g., BTA18) (fig. S18). This pattern is reminiscent of the duplication pattern of the dog, rat, and mouse but different from that of primate and great-ape genomes (*14*, *15*). On average, cattle SDs >10 kbp represent 11.7% of base pairs in 10-kbp intervals located within cattle-specific EBRs and 23.0% of base pairs located within the artiodactyl-specific EBRs. By contrast, in the remainder of the genome sequence assigned to chromosomes the fraction of SDs was 1.7% ($P < 1 \times 10^{-12}$). These data indicate that SDs play a role in promoting chromosome rearrangements by nonallelic homologous recombination [e.g., (*16*)] and suggest that either a significant fraction of the SDs observed in cattle occurred before the Ruminant-Suina split, and/or that the sites for accumulation of SDs are nonrandomly distributed in artiodactyl genomes.

SDs involving genic regions may give rise to new functional paralogs. Seventy-six percent (778 out of 1020) of the cattle SDs correspond to complete or partial gene duplications with high sequence identity (median 98.7%). This suggests that many of these gene duplications are specific to either the artiodactyla or the *Bos* lineage and tend to encode proteins that often interface with the external environment, particularly immune proteins and sensory and/or olfactory receptors. Several of these gene duplications are also duplicated in other mammalian lineages (e.g., cytochrome P-450, sulfotransferase, ribonuclease A, defensins, and pregnancy-associated glycoproteins). Paralogs located in segmental duplications that are present exclusively in cattle may have functional implications for the unique physiology, environment, and diet of cattle.

An overrepresentation of genes involved in reproduction in cattle SDs (tables S8 and S9) is associated with several gene families expressed in the ruminant placenta. These families encode the intercellular signaling proteins pregnancy-associated glycoproteins (on BTA29), trophoblast Kunitz domain proteins (on BTA13), and interferon tau (*IFNT*) (on BTA8). A gene family encoding prolactin-related proteins (on BTA23) was only identified in the assembly-dependent analysis of SDs. These genes regulate ruminant-specific aspects of fetal growth, maternal adaptations to pregnancy, and the coordination of parturition (*17*, *18*). Although type I interferon (IFN) genes are primarily involved in host defense (*19*), *IFNT* prevents regression of the corpus luteum during early pregnancy, which results in a uterine environment receptive to early conceptus development (*20*).

Signatures of positive selection (obtained by measurement of their rates of synonymous and nonsynonymous substitutions) identified 71 genes (*4*), including 10 immune-related genes (i.e., *IFNAR2*, *IFNG*, *CD34*, *TREM1*, *TREML1*, *FCER1A*, *IL23R*, *IL24*, *IL15*, and *LEAP2*). As previously mentioned, immune genes are overrepresented in SDs (see Table 1 and fig. S20). Examples of genes varying in cattle relative to mouse include a cluster of β-defensin genes, which encode antimicrobial peptides; the antimicrobial cathelicidin genes [which show increased sequence diversity of the mature cathelicidin peptides (*21*)]; and changes in the numbers of interferon genes (*22*) and the number and organization of genes involved in adaptive immune responses in cattle compared with human and mouse (*4*). This extensive duplication and divergence of genes involved in innate immunity may be because of the substantial load of microorganisms present in the rumen of cattle, which increases the risk of opportunistic infections at mucosal surfaces and positive selection for the traits that enabled stronger and more diversified innate immune responses at these locations. Another possibility is that immunity may have been under selection due to the herd structure, which can promote rapid disease transmission. Also, immune function–related duplicated genes have gained nonimmune functions, e.g., *IFNT* (see above), and the C-class lysozyme genes, which are involved in microbial degradation in the abomasum (see below).

There has been substantial reorganization of gene families encoding proteins present in milk. One such rearrangement affecting milk composition involves the histatherin (*HSTN*) gene within the casein gene cluster on BTA6 (fig. S21). In the cattle genome, *HSTN* is juxtaposed to a regulatory element (*BCE*) important (*23*) for β-casein (*CSN2*) expression, and as a probable consequence, *HSTN* is regulated like the casein genes during the lactation cycle. This rearrangement that led to the juxtaposition of *HSTN* next to the *BCE* is also the probable cause of deletion of one of the two copies of α-S2–like casein genes (*CSN1S2A*) present in other mammalian genomes (*24*). The biological implications of this change in casein gene copy number are not yet clear.

Additionally, the cattle serum amyloid A (*SAA*) gene cluster arose from both a laurasiatherian SD and a cattle-specific EBR, which resulted in two mammary gland–expressed *SAA3*-like genes, *SAA3.1* and *SAA3.2* on BTA29, and an *SAA3*-like gene on BTA15 (fig. S21). SAA3.2 has been shown to inhibit microbial growth (*25*). Two additional milk protein genes were associated with SDs: cathelicidin (*CATHL1*) and β_2-microglobulin (*B2M*)—part of the neonatal Fc receptor (FcRn) that transfers immunoglobulin IgG across epithelial cells of many tissues including the gut and mammary gland (*26*, *27*). IgG is the predominant immunoglobulin in cow's milk compared with IgA in human milk (*28*). Unlike humans, who acquire passive immunity from the mother via placental transfer of immunoglobulins during pregnancy, calves acquire passive immunity by ingestion of IgG in milk (*28*). *B2M* is also redistributed in epithelial cells upon calving, and it protects IgG from degradation (*26*). A genetic variant of *B2M* has negative effects on passive immune transfer (*29*). The additional copy of the gene encoding *B2M* might be associated with the abundance of IgG in cows' milk and an increased capacity for uptake in the neonatal gut. Considering that the passive transfer of immunity to the calf is one of the important functions of milk, it is striking that lactation-related genes affected by genomic rearrangements often encode immune-related proteins in milk.

Cattle metabolic pathways demonstrated a strong degree of conservation among the comprehensive set of genes involved in core mammalian metabolism (*4*) and permitted an examination of unique genetic events that may be related to ruminant-specific metabolic adaptations. However, among 1032 genes examined from the human metabolic pathways, five were deleted or extensively diverged in cattle: *PLA2G4C* (phospholipase A2, group IVC), *FAAH2* (fatty acid amide hydrolase 2), *IDI2* (isopentenyl-diphosphate delta isomerase 2), *GSTT2* (glutathione *S*-transferase

theta 2), and *TYMP* (thymidine phosphorylase), which may be adaptations that impact on fatty acid metabolism (*PLA2G4C* and *FAAH2)*; the mevalonate pathway (synthesis of dolichols, vitamins, steroid hormones, and cholesterol) (*IDI2*); detoxification (*GSTT2*); and pyrimidine metabolism (*TYMP*). Phylogenetic analysis shows that *PLA2G4C* was deleted ~87 to 97 million years ago in the laurasiatherian lineages (fig. S22). Strikingly, ~20% of the sequences from two abomasum (last chamber of the cattle stomach) EST libraries (a total of 2392 sequences) correspond to three C-type lysozyme genes. Lysozyme primarily functions in animals as an antibacterial protein, which suggests that they probably function in the abomasum (similar to the monogastric stomach) to degrade the cell walls of bacteria entering from the foregut (*30*). The cattle genome contains 10 C-type lysozyme genes (table S14 and fig. S23), and EST evidence (fig. S23) shows that six of the seven remaining C-type lysozyme genes are expressed primarily in the intestinal tract, which suggests additional roles for the encoded proteins in ruminant digestion.

In summary, the biological systems most affected by changes in the number and organization of genes in the cattle lineage include reproduction, immunity, lactation, and digestion. We highlighted the evolutionary activity associated with chromosomal breakpoint regions and their propensity for promoting gene birth and rearrangement. These changes in the cattle lineage probably reflect metabolic, physiologic, and immune adaptations due to microbial fermentation in the rumen, the herd environment and its influence on disease transmission, and the reproductive strategy of cattle. The cattle genome and associated resources will facilitate the identification of novel functions and regulatory systems of general importance in mammals and may provide an enabling tool for genetic improvement within the beef and dairy industries.

References and Notes

1. W. J. Murphy, P. A. Pevzner, S. J. O'Brien, *Trends Genet.* **20**, 631 (2004).
2. R. L. Willham, *J. Anim. Sci.* **62**, 1742 (1986).
3. Y. Liu *et al.*, *BMC Genomics* **10**, 180 (2009).
4. Materials, methods, and additional discussion are available on *Science* online.
5. H. Nilsen *et al.*, *Anim. Genet.* **39**, 97 (2008).
6. A. Prasad *et al.*, *BMC Genomics* **8**, 310 (2007).
7. H. S. Malik, T. H. Eickbush, *Mol. Biol. Evol.* **15**, 1123 (1998).
8. C. I. Wu, W. H. Li, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1741 (1985).
9. B. Modrek, C. J. Lee, *Nat. Genet.* **34**, 177 (2003).
10. R. Sorek, R. Shamir, G. Ast, *Trends Genet.* **20**, 68 (2004).
11. A. Everts-van der Wind *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 18526 (2005).
12. D. Kordis, F. Gubensek, *Gene* **238**, 171 (1999).
13. M. Shimamura, H. Abe, M. Nikaido, K. Ohshima, N. Okada, *Mol. Biol. Evol.* **16**, 1046 (1999).
14. J. A. Bailey, E. E. Eichler, *Nat. Rev. Genet.* **7**, 552 (2006).
15. J. A. Bailey *et al.*, *Science* **297**, 1003 (2002).
16. W. J. Murphy *et al.*, *Science* **309**, 613 (2005).
17. K. Hashizume *et al.*, *Reprod. Fertil. Dev.* **19**, 79 (2007).
18. J. H. Larson *et al.*, *Physiol. Genomics* **25**, 405 (2006).
19. S. Y. Zhang *et al.*, *Immunol. Rev.* **226**, 29 (2008).
20. R. M. Roberts, Y. Chen, T. Ezashi, A. M. Walker, *Semin. Cell Dev. Biol.* **19**, 170 (2008).
21. M. Scocchi, S. Wang, M. Zanetti, *FEBS Lett.* **417**, 311 (1997).
22. M. G. Katze, Y. He, M. Gale Jr., *Nat. Rev. Immunol.* **2**, 675 (2002).
23. C. Schmidhauser *et al.*, *Mol. Biol. Cell* **3**, 699 (1992).
24. M. Rijnkels, L. Elnitski, W. Miller, J. M. Rosen, *Genomics* **82**, 417 (2003).
25. A. J. Molenaar *et al.*, *Biomarkers* **14**, 26 (2009).
26. B. Mayer *et al.*, *J. Dairy Res.* **72** (suppl. S1), 107 (2005).
27. D. C. Roopenian, S. Akilesh, *Nat. Rev. Immunol.* **7**, 715 (2007).
28. T. J. Newby, C. R. Stokes, F. J. Bourne, *Vet. Immunol. Immunopathol.* **3**, 67 (1982).
29. M. L. Clawson *et al.*, *Mamm. Genome* **15**, 227 (2004).
30. D. M. Irwin, *J. Mol. Evol.* **41**, 299 (1995).
31. J. H. Larson *et al.*, *BMC Genomics* **7**, 227 (2006).
32. Funded by the National Human Genome Research Institute (NHGRI U54 HG003273); the U.S. Department of Agriculture's Agricultural Research Service (USDA-ARS agreement no. 59-0790-3-196) and Cooperative State Research, Education, and Extension Service National Research Initiative (grant no. 2004-35216-14163); the state of Texas; Genome Canada through Genome British Columbia; the Alberta Science and Research Authority; the Commonwealth Scientific and Industrial Research Organization of Australia (CSIRO); Agritech Investments Ltd., Dairy Insight, Inc., and AgResearch Ltd., all of New Zealand; the Research Council of Norway; the Kleberg Foundation; and the National, Texas, and South Dakota Beef Check-off Funds. The master accession for this WGS sequencing project is AAFC03000000. The individual WGS sequences are AAFC03000001 to AAFC03131728, and the scaffold records are CM000177 to CM000206 (chromosomes) and DS490632 to DS495890 (unplaced scaffolds).

The Bovine Genome Sequencing and Analysis Consortium

Principal Investigator: Richard A. Gibbs[1]
Analysis project leadership: Christine G. Elsik,[2,3] Ross L. Tellam[4]
Sequencing project leadership: Richard A. Gibbs,[1] Donna M. Muzny,[1] George M. Weinstock[5,1]
Analysis group organization: David L. Adelson,[6] Evan E. Eichler,[7,8] Laura Elnitski,[9] Christine G. Elsik,[2,3] Roderic Guigó,[10] Debora L. Hamernik,[11] Steve M. Kappes,[12] Harris A. Lewin,[13,14] David J. Lynn,[15] Frank W. Nicholas,[16] Alexandre Reymond,[17] Monique Rijnkels,[18] Loren C. Skow,[19] Ross L. Tellam,[4] Kim C. Worley,[1] Evgeny M. Zdobnov[20,21,22]
Sequencing project white paper: Richard A. Gibbs,[1] Steve M. Kappes,[12] Lawrence Schook,[13] Loren C. Skow,[19] George M. Weinstock,[5,1] James Womack[23]
Gene prediction and consensus gene set: Tyler Alioto,[10] Stylianos E. Antonarakis,[20] Alex Astashyn,[24] Charles E. Chapple,[10] Hsiu-Chuan Chen,[24] Jacqueline Chrast,[17] Francisco Câmara,[10] Christine G. Elsik[2,3] (leader), Olga Ermolaeva,[24] Roderic Guigó,[10] Charlotte N. Henrichsen,[17] Wratko Hlavina,[24] Yuri Kapustin,[24] Boris Kiryutin,[24] Paul Kitts,[24] Felix Kokocinski,[25] Melissa Landrum,[24] Donna Maglott,[24] Kim Pruitt,[24] Alexandre Reymond,[17] Victor Sapojnikov,[24] Stephen M. Searle,[25] Victor Solovyev,[26] Alexandre Souvorov,[24] Catherine Ucla,[20] George M. Weinstock,[5,1] Carine Wyss[20]
Experimental validation of gene set: Tyler Alioto,[10] Stylianos E. Antonarakis,[20] Charles E. Chapple,[10] Jacqueline Chrast,[17] Francisco Câmara,[10] Roderic Guigó[10] (leader), Charlotte N. Henrichsen,[17] Alexandre Reymond,[17] Catherine Ucla,[20] Carine Wyss[20]
MicroRNA analysis: Juan M. Anzola,[3] Daniel Gerlach,[20,21] Evgeny M. Zdobnov[20,21,22] (leader)
GC composition analysis: Eran Elhaik,[27,28] Christine G. Elsik[2,3] (leader), Dan Graur,[27] Justin T. Reese[2]
Repeat analysis: David L. Adelson[6] (leader), Robert C. Edgar,[29] John C. McEwan,[30] Gemma M. Payne,[30] Joy M. Raison[31]
Protein ortholog analysis: Thomas Junier,[19,20] Evgenia V. Kriventseva,[32] Evgeny M. Zdobnov[20,21,22] (leader)
Exon-skipping analysis: Jacqueline Chrast,[17] Eduardo Eyras,[33,34] Charlotte N. Henrichsen,[17] Mireya Plass,[34] Alexandre Reymond[17] (leader)
Evolutionary breakpoint analysis and Oxford grid: Ravikiran Donthu,[13] Denis M. Larkin,[13,14] Harris A. Lewin[13,14] (leader), Frank W. Nicholas[16]
Bidirectional promoter analysis: Laura Elnitski[9] (leader), Denis M. Larkin,[13,14] Harris A. Lewin,[13,14] James Reecy,[35] Mary Q. Yang[9]
Segmental duplication analysis: David L. Adelson,[6] Lin Chen,[7] Ze Cheng,[7] Carol G. Chitko-McKown,[36] Evan E. Eichler[7,8] (leader), Laura Elnitski,[9] Christine G. Elsik,[2,3] George E. Liu,[37] Lakshmi K. Matukumalli,[38,37] Jiuzhou Song,[39] Bin Zhu[39]
Analysis of gene ontology in segmental duplications: Christine G. Elsik,[2,3] David J. Lynn[15] (leader), Justin T. Reese[2]
Adaptive evolution: Daniel G. Bradley,[40] Fiona S.L. Brinkman,[15] Lilian P.L. Lau,[40] David J. Lynn[15] (leader), Matthew D. Whiteside[15]
Innate immunity: Ross L. Tellam[4] (leader), Angela Walker,[41] Thomas T. Wheeler[42]
Lactation: Theresa Casey,[43] J. Bruce German,[44,45] Danielle G. Lemay,[45] David J. Lynn,[15] Nauman J. Maqbool,[46] Adrian J. Molenaar,[42] Monique Rijnkels[18] (leader)
Metabolism: Harris A. Lewin[13,14] (leader), Seongwon Seo,[47] Paul Stothard[48]
Adaptive immunity: Cynthia L. Baldwin,[49] Rebecca Baxter,[50] Candice L. Brinkmeyer-Langford,[19] Wendy C. Brown,[51] Christopher P. Childers,[2] Timothy Connelley,[52] Shirley A. Ellis,[53] Krista Fritz,[19] Elizabeth J. Glass,[50] Carolyn T.A. Herzig,[49] Antti Iivanainen,[54] Kevin K. Lahmers,[51] Loren C. Skow[19] (leader)
Annotation data management: Anna K. Bennett,[2] Christopher P. Childers,[2] C. Michael Dickens,[3] Christine G. Elsik[2,3] (leader), James G.R. Gilbert,[25] Darren E. Hagen,[2] Justin T. Reese,[2] Hanni Salih[3]
Manual annotation organization: Jan Aerts,[55] Alexandre R. Caetano,[56] Brian Dalrymple,[4] Christine G. Elsik,[2,3] Jose Fernando Garcia,[57] Richard A. Gibbs,[1] Clare A. Gill,[3,58] Debora L. Hamernik,[11] Stefan G. Hiendleder,[59] Erdogan Memili,[60] Frank W. Nicholas,[16] James Reecy,[35] Monique Rijnkels,[18] Loren C. Skow,[19] Diane Spurlock,[35] Paul Stothard,[48] Ross L. Tellam,[4] George M. Weinstock,[5,1] John L. Williams,[61] Kim C. Worley[1]
cDNA tissues, libraries, and sequencing: Lee Alexander,[62] Michael J. Brownstein,[63] Leluo Guan,[48] Robert A. Holt[64] (leader), Steven J.M. Jones[64] (leader), Marco A. Marra[64] (leader), Richard Moore,[64] Stephen S. Moore[48] (leader), Andy Roberts,[62] Masaaki Taniguchi,[65,48] Richard C. Waterman[62]
Genome sequence production: Joseph Chacko,[1] Mimi M. Chandrabose,[1] Andy Cree[1] (leader), Marvin Diep Dao,[1] Huyen H. Dinh[1] (leader), Ramatu Ayiesha Gabisi,[1] Sandra Hines,[1] Jennifer Hume[1] (leader), Shalini N. Jhangiani,[1] Vandita Joshi,[1] Christie L. Kovar[1] (leader), Lora R. Lewis,[1] Yih-shin Liu,[1] John Lopez,[1] Margaret B. Morgan,[1] Donna M. Muzny[1] (leader), Ngoc Bich Nguyen,[1] Geoffrey O. Okwuonu,[1] San Juana Ruiz,[1] Jireh Santibanez,[1] Rita A. Wright[1]
Sequence finishing: Christian Buhay[1] (leader), Yan Ding,[1] Shannon Dugan-Rocha[1] (leader), Judith Herdandez,[1] Michael Holder,[1] Aniko Sabo[1]
Automated BAC assembly: Amy Egan,[1] Jason Goodell,[1] Katarzyna Wilczek-Boney[1]
Sequence production informatics: Gerald R. Fowler[1] (leader), Matthew Edward Hitchens,[1] Ryan J. Lozado,[1] Charles Moen,[1] David Steffen,[66,1] James T. Warren,[1] Jingkun Zhang[1]
BAC mapping: Readman Chiu,[64] Steven J.M. Jones,[64] Marco A. Marra[64] (leader), Jacqueline E. Schein[64]
Genome assembly: K. James Durbin,[67,1] Paul Havlak,[68,1] Huaiyang Jiang,[1] Yue Liu,[1] Xiang Qin,[1] Yanru Ren,[1] Yufeng Shen,[1,69] Henry Song,[1] George M. Weinstock,[5,1] Kim C. Worley[1] (leader)
Sequence library production: Stephanie Nicole Bell,[1] Clay Davis,[1] Angela Jolivet Johnson,[1] Sandra Lee,[1] Lynne V. Nazareth[1] (leader), Bella Mayurkumar Patel,[1] Ling-Ling Pu,[1] Selina Vattathil,[1] Rex Lee Williams Jr.[1]
BAC production: Stacey Curry,[1] Cerissa Hamilton,[1] Erica Sodergren[5,1] (leader)
Sequence variation detection: Lynne V. Nazareth,[1] David A. Wheeler[1]
Markers and mapping: David L. Adelson,[6] Jan Aerts,[55] Wes Barris,[4] Gary L. Bennett,[36] Brian Dalrymple,[4] André Eggen,[70] Clare A. Gill,[3,58] Ronnie D. Green,[71] Gregory P. Harhay,[36] Matthew Hobbs,[72] Oliver Jann,[50] Steve M. Kappes[12] (leader), John W. Keele,[36] Matthew P. Kent,[73] Denis M. Larkin,[13,14] Harris A. Lewin,[13,14] Sigbjørn Lien,[73] John C. McEwan,[30] Stephanie D. McKay,[74] Sean McWilliam,[4] Stephen S. Moore,[48] Frank W. Nicholas,[16] Gemma M. Payne,[30] Abhirami Ratnakumar,[75,4] Hanni Salih,[3] Robert D. Schnabel,[74] Timothy Smith,[36] Warren M. Snelling,[36] Tad S. Sonstegard,[37] Roger T. Stone,[36] Yoshikazu Sugimoto,[76] Akiko Takasuga,[76] Jeremy F. Taylor,[74] Ross L. Tellam,[4] Curtis P. Van Tassell,[37] John L. Williams[61]
Genomic DNA: Michael D. MacNeil[62]
Manual annotation: Antonio R.R. Abatepaulo,[77] Colette A. Abbey,[3] Jan Aerts,[55] Virpi Ahola,[78] Iassudara G. Almeida,[57] Ariel F. Amadio,[79] Elen Anatriello,[77] Suria M. Bahadue,[2] Cynthia L. Baldwin,[49] Rebecca Baxter,[50] Anna K. Bennett,[2] Fernando H. Biase,[13] Clayton R. Boldt,[3] Candice L. Brinkmeyer-Langford,[19] Wendy C. Brown,[51] Alexandre R. Caetano,[56] Jeffery A. Carroll,[80] Wanessa A. Carvalho,[77] Theresa Casey,[43] Eliane P. Cervelatti,[57] Elsa Chacko,[81] Jennifer E. Chapin,[3] Ye Cheng,[35] Christopher P. Childers,[2] Jungwoo Choi,[3] Adam J. Colley,[82] Timothy Connelley,[52] Tatiana A. de Campos,[56] Marcos De Donato,[83]

Isabel K.F. de Miranda Santos,[56,77] Carlo J.F. de Oliveira,[77] Heather Deobald,[84] Eve Devinoy,[85] C. Michael Dickens,[3] Kaitlin E. Donohue,[2] Peter Dovc,[86] Annett Eberlein,[87] Shirley A. Ellis,[53] Carolyn J. Fitzsimmons,[59] Alessandra M. Franzin,[77] Krista Fritz,[19] Gustavo R. Garcia,[77] Jose Fernando Garcia,[57] Sem Genini,[61] J. Bruce German,[44,45] James G.R. Gilbert,[25] Clare A. Gill,[3,58] Cody J. Gladney,[3] Elizabeth J. Glass,[50] Jason R. Grant,[48] Marion L. Greaser,[88] Jonathan A. Green,[74] Darryl L. Hadsell,[18] Darren E. Hagen,[2] Hatam A. Hakimov,[89] Rob Halgren,[43] Jennifer L. Harrow,[25] Elizabeth A. Hart,[25] Nicola Hastings,[90,50] Marta Hernandez,[91] Carolyn T.A. Herzig,[49] Stefan G. Hiendleder,[59] Matthew Hobbs,[72] Zhi-Liang Hu,[35] Antti Iivanainen,[54] Aaron Ingham,[4] Terhi Iso-Touru,[78] Catherine Jamis,[2] Oliver Jann,[50] Kirsty Jensen,[50] Dimos Kapetis,[61] Tovah Kerr,[51] Sari S. Khalil,[2] Hasan Khatib,[92] Davood Kolbehdari,[48,93] Charu G. Kumar,[13] Dinesh Kumar,[94,35] Richard Leach,[50] Justin C-M Lee,[2] Danielle G. Lemay,[45] Changxi Li,[95,48] George E. Liu,[37] Krystin M. Logan,[96] Roberto Malinverni,[61] Nauman J. Maqbool,[46] Elisa Marques,[48] William F. Martin,[45] Natalia F. Martins,[56] Sandra R. Maruyama,[77] Raffaele Mazza,[97] Kim L. McLean,[84] Juan F. Medrano,[98] Erdogan Memili,[60] Adrian J. Molenaar,[42] Barbara T. Moreno,[57] Daniela D. Moré,[77] Carl T. Muntean,[3] Hari P. Nandakumar,[19] Marcelo F.G. Nogueira,[99] Ingrid Olsaker,[100] Sameer D. Pant,[82] Francesca Panzitta,[61] Rosemeire C.P. Pastor,[57] Mario A. Poli,[101] Nathan Poslusny,[2] Satyanarayana Rachagani,[35] Shoba Ranganathan,[81,102] Andrej Razpet,[86] James Reecy,[35] Penny K. Riggs,[3,58] Monique Rijnkels,[18] Gonzalo Rincon,[98] Nelida Rodriguez-Osorio,[60,103] Sandra L. Rodriguez-Zas,[13] Natasha E. Romero,[3] Anne Rosenwald,[2] Lillian Sando,[4] Sheila M. Schmutz,[84] Seongwon Seo,[47] Libing Shen,[2] Laura Sherman,[48] Loren C. Skow,[19] Bruce R. Southey,[104] Diane Spurlock,[35] Ylva Strandberg Lutzow,[4] Jonathan V. Sweedler,[104] Imke Tammen,[72] Masaaki Taniguchi,[65,48] Ross L. Tellam,[4] Bhanu Prakash V.L. Telugu,[74] Jennifer M. Urbanski,[2] Yuri T. Utsunomiya,[57] Chris P. Verschoor,[82] Ashley J. Waardenberg,[4,105] Angela Walker,[41] Zhiquan Wang,[48] Robert Ward,[106] Rosemarie Weikard,[87] Thomas H. Welsh Jr.,[3,58] Thomas T. Wheeler,[42] Stephen N. White,[51,107] John L. Williams,[61] Laurens G. Wilming,[25] Kris R. Wunderlich,[3] Jianqi Yang,[108] Feng-Qi Zhao[109]

[1]Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [2]Department of Biology, 406 Reiss, Georgetown University, 37th & O Streets NW, Washington, DC 20057, USA. [3]Department of Animal Science, Texas A&M University, 2471 TAMU, College Station, TX 77843–2471, USA. [4]Livestock Industries, Commonwealth Scientific and Industrial Research Organization (CSIRO), 306 Carmody Road, St. Lucia, Queensland, 4067, Australia. [5]The Genome Center at Washington University, Washington University School of Medicine, 4444 Forest Park Avenue, St. Louis, MO 63108, USA. [6]School of Molecular and Biomedical Science, School of Agriculture, Food and Wine, The University of Adelaide, Adelaide, SA, 5005, Australia. [7]Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195–5065, USA. [8]Howard Hughes Medical Institute, Seattle, WA 98195, USA. [9]National Human Genome Research Institute, National Institutes of Health, 5625 Fishers Lane, Rockville, MD 20878, USA. [10]Center for Genomic Regulation and Grup de Recerca en Informática Biomédica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain. [11]U.S. Department of Agriculture (USDA), Cooperative State Research, Education, & Extension Service, 1400 Independence Avenue SW, Stop 2220, Washington, DC 20250–2220, USA. [12]National Program Staff, USDA–Agricultural Research Service, 5601 Sunnyside Avenue, Beltsville, MD 20705, USA. [13]Department of Animal Sciences, University of Illinois at Urbana–Champaign, 1201 West Gregory Drive, Urbana, IL 61801, USA. [14]Institute for Genomic Biology, University of Illinois at Urbana–Champaign, 1201 West Gregory Drive, Urbana, IL 61801, USA. [15]Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada. [16]Faculty of Veterinary Science, University of Sydney, Sydney, NSW, 2006, Australia. [17]Center for Integrative Genomics, University of Lausanne, Lausanne, 1015, Switzerland. [18]Children's Nutrition Research Center, USDA–Agricultural Research Service, Department of Pediatrics–Nutrition, Baylor College of Medicine, 1100 Bates Street, Houston, TX 77030–2600, USA. [19]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [20]Department of Genetic Medicine and Development, University of Geneva Medical School, 1 rue Michel-Servet, Geneva, 1211, Switzerland. [21]Swiss Institute of Bioinformatics, 1 rue Michel-Servet, Geneva, 1211, Switzerland. [22]Division of Molecular Biosciences, Imperial College London, South Kensington Campus, London, SW7 2AZ, UK. [23]Department of Veterinary Pathobiology, Texas A&M University, College Station, TX 77843, USA. [24]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892, USA. [25]Informatics Department, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK. [26]Department of Computer Science, University of London, Royal Holloway, Egham, Surrey, TW20 0EX, UK. [27]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA. [28]McKusick—Nathans Institute of Genetic Medicine, BRB 579, Johns Hopkins University School of Medicine, 733 North Broadway, Baltimore, MD 21205, USA. [29]45 Monterey Drive, Tiburon, CA 94920, USA. [30]Animal Genomics, AgResearch, Invermay, PB 50034, Mosgiel, 9053, New Zealand. [31]eResearch SA, University of Adelaide, North Terrace, Adelaide, SA, 5005, Australia. [32]Department of Structural Biology and Bioinformatics, University of Geneva Medical School, 1 rue Michel-Servet, Geneva, 1211, Switzerland. [33]Catalan Institution for Research and Advanced Studies, 08010 Barcelona, Catalonia, Spain. [34]Computational Genomics, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain. [35]Department of Animal Science, Iowa State University, 2255 Kildee Hall, Ames, IA 50011–3150, USA. [36]Meat Animal Research Center, USDA–Agricultural Research Service, Clay Center, NE 68933, USA. [37]Bovine Functional Genomics Laboratory, USDA–Agricultural Research Service, Beltsville Agricultural Research Center (BARC)–East, Beltsville, MD 20705, USA. [38]Department of Bioinformatics and Computational Biology, George Mason University, 10900 University Blvd, Manassas, VA 20110, USA. [39]Department of Bioengineering, University of Maryland, College Park, MD 20742, USA. [40]Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland. [41]Department of Veterinary Pathobiology, 245 Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA. [42]Dairy Science and Technology Section, AgResearch, Ruakura Research Centre, East Street, Private Bag 3123, Hamilton, 3240, New Zealand. [43]Department of Animal Science, Michigan State University, East Lansing, MI 48824–1225, USA. [44]Nestlé Research Centre, Vers chez les Blanc CH, Lausanne 26, 1000, Switzerland. [45]Department of Food Science and Technology, University of California–Davis, Davis, CA 95616, USA. [46]Bioinformatics, Mathematics and Statistics, AgResearch, Ruakura Research Centre, East Street, Private Bag 3123, Hamilton, 3240, New Zealand. [47]Division of Animal Science and Resource, Chungnam National University, Daejeon, 305-764, Korea. [48]Department of Agricultural, Food and Nutritional Science, University of Alberta, 410 AgFor Centre, Edmonton, AL, T6G 2P5, Canada. [49]Department of Veterinary and Animal Sciences, University of Massachusetts, Amherst, MA 01003, USA. [50]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, Midlothian, EH25 9PS, UK. [51]Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, WA 99164, USA. [52]Division of Infection and Immunity, The Roslin Institute, Royal (Dick) School of Veterinary Science, University of Edinburgh, Roslin, Midlothian, EH25 9RG, UK. [53]Immunology Division, Institute for Animal Health, Compton, RG20 7NN, UK. [54]Department of Basic Veterinary Sciences, University of Helsinki, Post Office Box 66, Helsinki, FIN-00014, Finland. [55]Genome Dynamics and Evolution, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK. [56]Embrapa Recursos Genéticos e Biotecnologia, Final Avenida W/5 Norte, Brasilia, DF, 70770-900, Brazil. [57]Animal Production and Health Department, UNESP—Sao Paulo State University, Aracatuba, SP, 16050-680, Brazil. [58]Texas AgriLife Research, College Station, TX 77843, USA. [59]JS Davies Epigenetics and Genetics Group, School of Agriculture, Food & Wine and Research Centre for Reproductive Health, The University of Adelaide, Roseworthy Campus, Roseworthy, SA, 5371, Australia. [60]Department of Animal and Dairy Sciences, Mississippi Agricultural and Forestry Experiment Station, Mississippi State University, Mississippi State, MS 39762, USA. [61]Parco Tecnologico Padano, Via Einstein, Polo Universitario, Lodi, 26900, Italy. [62]Fort Keogh Livestock and Range Research Laboratory, USDA-Agricultural Research Service, Miles City, MT 59301, USA. [63]Laboratory of Genetics, National Institute of Mental Health, NIH, Building 49, B1EE16, 49 Convent Drive, Bethesda, MD 20892, USA. [64]Genome Sciences Centre, British Columbia Cancer Agency, 675 West 10th Avenue, Vancouver, BC, V5Z 1L3, Canada. [65]Division of Animal Sciences, National Institute of Agrobiological Sciences, Tsukuba, Ibaraki, 305-8602, Japan. [66]Bioinformatics Research Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [67]Department of Biomolecular Engineering, University of California at Santa Cruz, Santa Cruz, CA 95064, USA. [68]Department of Computer Science, University of Houston, Houston, TX 77204–3010, USA. [69]Department of Computer Science and Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA. [70]INRA, Animal Genetics and Integrative Biology, Bovine Genetics and Genomics, 78350 Jouy-en-Josas, France. [71]Pfizer Animal Genetics, Pfizer Animal Health, New York, NY 10017, USA. [72]Faculty of Veterinary Science, University of Sydney, Camden, NSW, 2570, Australia. [73]Centre for Integrative Genetics and Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Arboretveien 6, Ås, 1432, Norway. [74]Division of Animal Sciences, University of Missouri, 920 East Campus Drive, Columbia, MO 65211, USA. [75]Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala Biomedical Centre Husargatan 3, Uppsala, 75 123, Sweden. [76]Shirakawa Institute of Animal Genetics, Nishigo, Fukushima 961-8061, Japan. [77]Department of Biochemistry and Immunology, Ribeirão Preto Medical School, University of São Paulo, Av Bandeirantes 3900, Ribeirão Preto, SP, 14049-900, Brazil. [78]Biotechnology and Food Research, MTT Agrifood Research Finland, Jokioinen, FI-31600, Finland. [79]EEA Rafaela, Instituto Nacional de Tecnología Agropecuaria (INTA), Ruta 34 Km 227, Rafaela, Santa Fe, 2300, Argentina. [80]Livestock Issues Research Unit, USDA–Agricultural Research Service, Lubbock, TX 79403, USA. [81]Department of Chemistry and Biomolecular Sciences & ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, 2109, NSW, Australia. [82]Department of Animal and Poultry Science, University of Guelph, Guelph, ON, N1G2W1, Canada. [83]Instituto de Investigaciones en Biomedicina y Ciencias Aplicadas, Universidad de Oriente, Avenida Universidad, Cumana, Sucre, 6101, Venezuela. [84]Department of Animal and Poultry Science, University of Saskatchewan, Saskatoon, SK, S7N 5A8, Canada. [85]INRA–UR1196, Génomique et Physiologie de la Lactation, F78352 Jouy-en-Josas, France. [86]Department of Animal Science, University of Ljubljana, Groblje 3, Domzale, SI-1230, Slovenia. [87]Research Unit Molecular Biology, Research Institute for the Biology of Farm Animals (FBN), Dummerstorf, 18196, Germany. [88]Department of Animal Sciences, University of Wisconsin–Madison, 1805 Linden Drive, Madison, WI 53706, USA. [89]Department of Molecular and Cellular Biology, University of Guelph, Guelph, ON, N1G 2W1, Canada. [90]Cell Biology and Biophysics, European Molecular Biology Laboratory (EMBL)–Heidelberg, Meyerhofstrasse 1, Heidelberg, Germany. [91]Laboratory of Molecular Biology, Instituto Tecnologico Agrario de Castilla y Leon (ITACyL), Carretera de Burgos km 119, Valladolid, 47071, Spain. [92]Department of Dairy Science, University of Wisconsin, Madison, WI 53706, USA. [93]Monsanto Company, 3302 SE Convenience Blvd, Ankeny, IA 50021, USA. [94]Genes & Genetic Resources Molecular Analysis Lab, National Bureau of Animal Genetic Resources, Baldi Bye Pass, Karnal, Haryana, 132001, India. [95]Lacombe Research Centre, Agriculture and Agri-Food Canada, Lacombe, AL, T4L 1W1, Canada. [96]Biomedical Sciences, University of Guelph, Guelph, ON, N1G 2W6, Canada. [97]Zootechnics Institute, Università Cattolica del Sacro Cuore, via Emilia Parmense 84, Piacenza, 29100, Italy. [98]Department of Animal Science, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA. [99]Departamento de Ciências Biológicas, Faculdade de Ciências e Letras, UNESP–São Paulo State University, Av Dom Antônio 2100, Vila Tênis Clube, Assis, SP, 19806-900, Brazil. [100]Department of Basic Sciences and Aquatic Medicine, Norwegian School of Veterinary Science, Post Office Box 8146 Dep, Oslo, NO-0033, Norway. [101]Instituto de Genética Ewald Favret, Instituto Nacional de Tecnología Agropecuaria (INTA), Las Cabañas y de Los Reseros s/n CC25, Castelar, Buenos Aires, B1712WAA, Argentina. [102]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore, 117597, Singapore. [103]Grupo CENTAURO, Universidad de Antioquia, Medellín, Colombia. [104]Department of Chemistry, University of Illinois, Urbana, IL 61801, USA. [105]Eskitis Institute for Cell and Molecular Therapies, Griffith University, Nathan, QLD, 4111, Australia.

[106]Nutrition and Food Sciences, Utah State University, Logan, UT 84322, USA. [107]Animal Disease Research Unit, USDA–Agricultural Research Service, Pullman, WA 99164, USA. [108]Department of Pharmacology, 2-344 BSB, University of Iowa, 51 Newton Road, Iowa City, IA 52242, USA. [109]Department of Animal Science, 211 Terrill, University of Vermont, 570 Main Street, Burlington, VT 05405, USA.

Supporting Online Material
www.sciencemag.org/cgi/content/full/324/5926/522/DC1
Materials and Methods
Figs. S1 to S23
Tables S1 to S14
References

10 December 2008; accepted 16 March 2009
10.1126/science.1169588

Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds

The Bovine HapMap Consortium*

The imprints of domestication and breed development on the genomes of livestock likely differ from those of companion animals. A deep draft sequence assembly of shotgun reads from a single Hereford female and comparative sequences sampled from six additional breeds were used to develop probes to interrogate 37,470 single-nucleotide polymorphisms (SNPs) in 497 cattle from 19 geographically and biologically diverse breeds. These data show that cattle have undergone a rapid recent decrease in effective population size from a very large ancestral population, possibly due to bottlenecks associated with domestication, selection, and breed formation. Domestication and artificial selection appear to have left detectable signatures of selection within the cattle genome, yet the current levels of diversity within breeds are at least as great as exists within humans.

The emergence of modern civilization was accompanied by adaptation, assimilation, and interbreeding of captive animals. In cattle (*Bos taurus*), this resulted in the development of individual breeds differing in, for example, milk yield, meat quality, draft ability, and tolerance or resistance to disease and pests. However, despite mapping and diversity studies (*1*–*5*) and the identification of mutations affecting some quantitative phenotypes (*6*–*8*), the detailed genetic structure and history of cattle are not known.

*The full list of authors with their contributions and affiliations is included at the end of the manuscript.

Cattle occur as two major geographic types, the taurine (humpless—European, African, and Asian) and indicine (humped—South Asian, and East African), which diverged >250 thousand years ago (Kya) (*3*). We sampled individuals representing 14 taurine (n = 376), three indicine (n = 73) (table S1), and two hybrid breeds (n = 48), as well as two individuals each of *Bubalus quarlesi* and *Bubalus bubalis*, which diverged from *Bos taurus* ~1.25 to 2.0 Mya (*9*, *10*). All breeds except Red Angus (n = 12) were represented by at least 24 individuals. We preferred individuals that were unrelated for ≥4 generations; however, each breed had one or two sire, dam, and progeny trios to allow assessment of genotype quality.

Single-nucleotide polymorphisms (SNPs) that were polymorphic in many populations were primarily derived by comparing whole-genome sequence reads representing five taurine and one indicine breed to the reference genome assembly obtained from a Hereford cow (*10*) (table S2). This led to the ascertainment of SNPs with high minor allele frequencies (MAFs) within the discovery breeds (table S5). Thus, as expected, with trio progeny removed, SNPs discovered within the taurine breeds had higher average MAFs

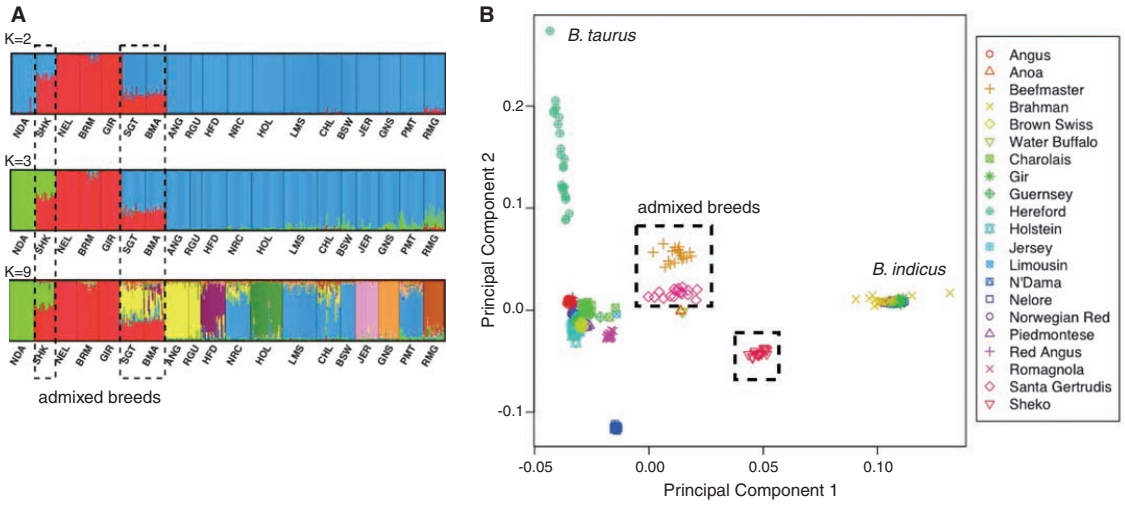


Fig. 1. (**A**) Population structure assessed by InStruct. Bar plot, generated by DISTRUCT, depicts classifications with the highest probability under the model that assumes independent allele frequencies and inbreeding coefficients among assumed clusters. Each individual is represented by a vertical bar, often partitioned into colored segments with the length of each segment representing the proportion of the individual's genome from K = 2, 3, or 9 ancestral populations. Breeds are separated by black lines. NDA, N'Dama; SHK, Sheko; NEL, Nelore; BRM, Brahman; GIR, Gir; SGT, Santa Gertrudis; BMA, Beefmaster; ANG, Angus; RGU, Red Angus; HFD, Hereford; NRC, Norwegian Red; HOL, Holstein; LMS, Limousin; CHL, Charolais; BSW, Brown Swiss; JER, Jersey; GNS, Guernsey; PMT, Piedmontese; RMG, Romagnola. (**B**) Principal components PC1 and PC2 from all SNPs. Taurine breeds remain separated from indicine breeds, and admixed breeds are intermediate.

A.1.2.3 *Supplementary information*

An extract of the supplementary material for the paper "The genome sequence of taurine cattle: a window to ruminant biology and evolution" covering my work is presented on pages 148–157. The original source and can be found here:
`http://www.sciencemag.org/cgi/data/324/5926/522/DC1/2`

bovine selenoproteins, we used TBLASTN to search for human-cow homologous proteins where a cysteine residue in human aligns to a TGA in the bovine genome. Of the seven sequences which passed our thresholds (E-value ≤0.1, no more than one in-frame STOP, and a potential SECIS element less than 10,000 bp downstream of the end of the BLAST high scoring pair), three were human non-selenoprotein members of the GPX family, which aligned to bovine GPX selenoproteins. The remaining four were compared with the NCBI's NR database but no alignments from any other species were found to support the alignment of a cysteine residue with a TGA codon found in human versus cow.

U12 Introns

U12 introns in the bovine genome were identified with the union of several computational approaches and this set was manually refined to eliminate obvious errors: (i) human introns from U12DB (*S37*) were mapped to the bovine genome with GMAP (*S38*) with 100 bp of exonic flanking sequence; (ii) human RefSeq, cow mRNA and cow ESTs were mapped to Btau3.1 scaffolds with GMAP (*S38*) and the resulting introns scored and classified as U12 or U2, and; (iii) Geneid *ab initio* predictions allowing for U12 introns were made with permissive parameters and the predicted U12 introns (~2,900 total) were filtered according to alignment of flanking sequence to cattle mRNAs, cattle ESTs and the NCBI NR protein database (*S18*). Alignment support was considered positive if >67% of the 10 translated amino acids on either side of the predicted splice junction (14/20 in total) were aligned to the same or similar residues. The combined set of U12 introns was subjected to manual inspection to eliminate obvious mapping errors.

MicroRNAs

Bovine microRNAs (miRNAs) were independently predicted by two approaches (prediction Sets 1 and 2 below) and then combined to create a non-redundant set.

microRNA Prediction Set 1

First, a set of bovine predicted miRNAs was generated by comparison with known miRNAs as follows. Metazoan miRNA were downloaded from miRBase version 11.0 (*S39*) in FASTA format containing the respective start and end positions of their mature parts. WU-BLAST (*S40*) was used to search each of the known miRNAs with the default parameters plus a DUST filter and the *hspsepSmax 30* option for defining the maximum separating distance between two high score pairs (this allows for a varying pre-miRNA loop while still matching the better conserved 5' and 3' arms). BLASTN matches longer than 20 bp were extended at both ends to match the length of the query sequence. To remove unstable or spurious hits three filter parameters where calculated for each putative pre-miRNA. These included a minimum free energy filter (≤-15 kcal/mol), a RANDFOLD (*S41*) filter estimating the stability of the folding compared to dinucleotide shuffled folded sequences (100 randomizations, p-value ≤0.05) and finally a RNAshapes filter (*S42*) was used to predict the probability of the sequence folding into a simple stem-loop like shaped structure. Nevertheless, the RNAshapes filter was not applied on the final predictions as some known miRNAs, like hsa-let-7a, are known to not meet criteria for stable stem-loop structures when subjected to a minimum free energy folding algorithm, such as RNAfold (*S43*). The putative miRNAs that passed these filters were aligned to their query miRNAs with MAFFT (*S44, S45*) and the conservation of the seed region was calculated by mapping the known mature miRNA region on the query miRNA to the alignment. Criteria for

bovine miRNA predictions were 100% conservation of the seed (nucleotides 2-7 of the mature miRNA) and more than 90% sequence identity over the full mature miRNA. As several miRNA (like hsa-let-7, mmu-let-7, etc) can map to the same locus, all predictions were clustered with GALAXY (*S45*). From a single locus the match with the highest conservation of the mature miRNA and the highest overall percent alignment identity over the entire putative pre-miRNA was used as a single representative sequence for that locus. Including 10 additional pre-miRNAs found after an update to miRBase 12.0, this approach yielded a total of 361 pre-miRNAs in the bovine genome with homology to known miRNAs in other animal genomes.

In addition to predicting bovine miRNAs on the basis of homology to known miRNAs, we used a comparative approach on the basis of a Support Vector Machine (SVM) model of hairpin-like structures followed by an orthology assignment step. This method allows prediction of novel miRNAs that do not show sequence homology to known miRNAs. The complete method is described in (*S46*); what follows is a brief outline of the basic principles. First, an *ab-intio* SVM model was created to score stem-loop like sequences extracted from the genomic sequence with RNAfold (*S43*). Second, an orthology assignment pipeline grouped putative precursors from over 40 animal species, then precursors within groups were aligned. In a third step the orthologous groups were again subjected to an SVM model designed to distinguish alignments of orthologous miRNA sequences from other ncRNA alignments or false positive predictions, taking into account typical conservation patterns in pre-miRNA sequence alignments. This approach yielded 135 putative novel bovine miRNAs that are not yet found in miRBase 12.0 with the direct homology approach. The ortholog-based and novel bovine predicted miRNAs were combined to form Set 1.

microRNA Prediction Set 2

Mature miRNAs and stem-loop precursors were downloaded from miRbase v. 10 (*S39*). Mature miRNAs and their respective precursors were combined into a single sequence with the mature region in lower case format. Each precursor miRNA sequence was aligned with Btau3.1 with WU-BLAST (*S40*). BLAST was performed by seeding only the mature region of the precursor miRNA to minimize false positives, and then allowing seed extensions outside the mature region. BLAST output was parsed and a sequence corresponding to each hit was extracted from the assembly, extending the extracted sequence to the length of the original query. A global alignment between query (precursor miRNA) and subject sequence (extracted region) was constructed with T-COFFEE (*S47*), and the number of substitutions was determined. The free energy of folding of the subject sequence were computed with RNAfold (*S43*). A PRSS analysis between the two sequences was performed with 1,000 iterations in order to assess the statistical significance of the alignment and confirm that the two sequences were homologous. PRSS is part of the Fasta sequence comparison package (*S48*), and works by constructing local alignments between a query and a database of shuffled subject sequences to generate a distribution of alignment scores, which is used to compute an E-value for the alignment of the query to the actual subject. In our case, the query was the precursor miRNA and the subject was the extracted region of the assembly. A RANDfold analysis of the subject sequence was performed in order to determine how likely the sequence resembled a miRNA. Most of the known miRNAs are in a structural conformation corresponding to a free energy of folding that is considerably lower than that for shuffled sequences with the same nucleotide composition, indicating a tendency in the sequence towards a stable secondary structure (*S41*). RANDfold was run with 1,000 iterations per sequence, and the results were tabulated. Putative miRNA homologs were

kept if they were at least 95% identical to the known miRNA or if the following conditions were met: (i) similarity score of at least 65% throughout the entire global alignment; (ii) free energy of folding ≤-20 kcal/mol or lower; (iii) PRSS score ≤1e-05, and; (iv) RANDfold score ≤0.015. In many cases there were more than one miRNA per genomic locus. This was particularly true for miRNAs that are known to have several paralogs, or due to orthologous genes that are redundant in the database used (miRBase v.10). All overlapping miRNAs were clustered on the basis of genomic location and sequences within the cluster were scored based on similarity to known query miRNA. Only the sequence with the greatest similarity was used in further analysis. Putative microRNAs were analyzed with RepeatMasker to remove repetitive and transposable elements.

Merging microRNA Set 1 and microRNA Set 2
The precursor miRNA sequences from the two miRNA prediction sets were compared with WU-BLAST with default settings, except the hspsepSmax parameter was set to 30. The results were parsed for hits on the same strand with 100% sequence identity and more than 50 bp alignment length. Sequences that met these alignment criteria were considered identical miRNA loci if their start and end coordinates in the genome did not differ by more than 25 bp. This step prevents merging of paralogous loci, but allows for variation in length of precursor miRNAs predicted with different methods. Most of the predictions missed in Set 2 were located in the unassigned scaffolds and/or were new miRNAs present in miRBase version 12.0.

Bovine Official Gene Set
The OGSv1 used in global analyses and annotation was the GLEAN5 consensus set. OGSv2 was generated by: (i) rerunning GLEAN with new cDNA evidence and; (ii) incorporating manual annotations, selenoproteins and U12 intron data into the consensus gene set. Specifically, manual annotations, selenoprotein and U12 intron data were used to replace GLEAN5 gene models or add additional gene models.

Manual Annotation
Manual annotation was performed by a group of approximately 150 scientists who typically had experience with specific genes. The aims of the manual annotation effort were to confirm or correct OGSv1 automated gene models, identify genes missing from OGSv1, and identify changes in genes or gene families that comment on ruminant biology and evolution. A total of approximately 4,000 gene models were manually inspected. The initial step was to obtain the sequence of a bovine EST/ cDNA or a human or mouse protein ortholog from RefSeq (*S49*), Ensembl (*S50*), UCSC Genome Browser (*S51*) or Uniprot (*S52*). This sequence was used to search the OGSv1 translated protein database with BLASTP or BLASTX (*S22*). The most significant Expect values and bit scores for the bovine ortholog were generally well separated from secondary hits. Reciprocal BLAST analysis was performed to validate the ortholog. For gene families syntenic position was also used to define orthology. Genes missing from OGSv1 were identified in the assembly by comparing bovine EST/cDNA or protein homologs to the assembly with BLASTN or TBLASTN. Gene models were then annotated with one of three methods. Some participants used manual methods and web tools such as BLAST at NCBI, Ensembl and Bovine Genome Database to annotate a gene, and submitted annotations to a manual submission website at the Bovine Genome Database (*S53*). Other people participated in an annotation jamboree held at the Sanger Institute, and used the Otterlace annotation software

MicroRNAs
We identified 361 bovine miRNA genes with homology to experimentally verified microRNAs in the miRBase 12.0 and 135 novel microRNAs with a comparative genomic approach (*S46*). The 496 bovine miRNAs were grouped into 298 homolog families. About half of the bovine miRNA occur in 60 genomic miRNA clusters, in which 2 to 7 miRNA genes are separated by less than 10 kbp (Fig. S2). A notable exception was a 43 kbp cluster on BTA21 harboring approximately 40 contiguous microRNA genes that is orthologous to a large cluster on human 14q32.31 (*S98*). This region is imprinted in the mouse (*S99*).

GC Content
Animal genomes are not uniform in their long-range sequence composition, but are composed of a mosaic of sequence stretches of variable lengths that differ widely in their GC compositions. Whether these stretches meet the criteria of isochores [*sensu* (*S100*)], or should better be referred to as GC-content domains (*S101*) is a matter of debate (*S57, S102-S104*). In animal genome sequences studied to date, the distribution of GC-content domain lengths (plotted on a log-log scale) was found to follow a heavy-tail distribution with power-law decay exponents ranging from –1.12 to –1.15. The genome of the *B. taurus* genome is no exception and the compositionally homogeneous segments in its genome, as in all other genomes studied so far, do not have a characteristic length; rather, there is an abundance of short segments and only a small number of longer segments.

A comparison of the distributions of GC-content lengths among *B. taurus* (Btau4.0), *H. sapiens* (NCBI Build 36.3), and *M. musculus* (NCBI Build M37.1) is shown in Fig. S3. Interestingly, the bovine has the lowest abundance of small size GC-content domains (<2 kbp) relative to the other three genomes. The GC contents of their small domains span from 7% to 82%. In contrast, the mid- and long-size GC-content domains (3 kbp - 1 Mbp) in *B. taurus* are more frequent than in human but the long size domains are less frequent than in mouse. Only a small fraction (3%) of the homogeneous domains are longer than 300 kbp, however their mean GC content (39.6%) is significantly lower than the mean GC content for the entire genome (41.7%).

Phylogenetic Analysis of LINE Elements
The maximum likelihood tree of BovB elements, with 11 terminal clusters with branch lengths less than 0.02, indicates that a number of recent retrotransposition events of BovB have occurred, which is evidence for continued activity of BovB retrotransposons (Fig. S4). A similar analysis for L1_BT repeats is shown in Fig. S5.

Correlation of Repeat Elements
Figure S6 shows correlations among the repeat groups, gene density, GC content, and SD. A chromosome map of high and low density ancient repeats, L2/MIR (a LINE/SINE pair) and BovB, and more recent repeats, BovB/Art2A (BovB derived SINE pair) is shown in Fig. S7.

Simple Sequence Repeats
Figures S8 and S9 show the relative frequencies of different dinucleotide and trinucleotide repeats, respectively. Comparative frequencies of trinucleotide SSRs in the human, canine, bovine and ovine sequences are shown in figure S10. The latter information was obtained from

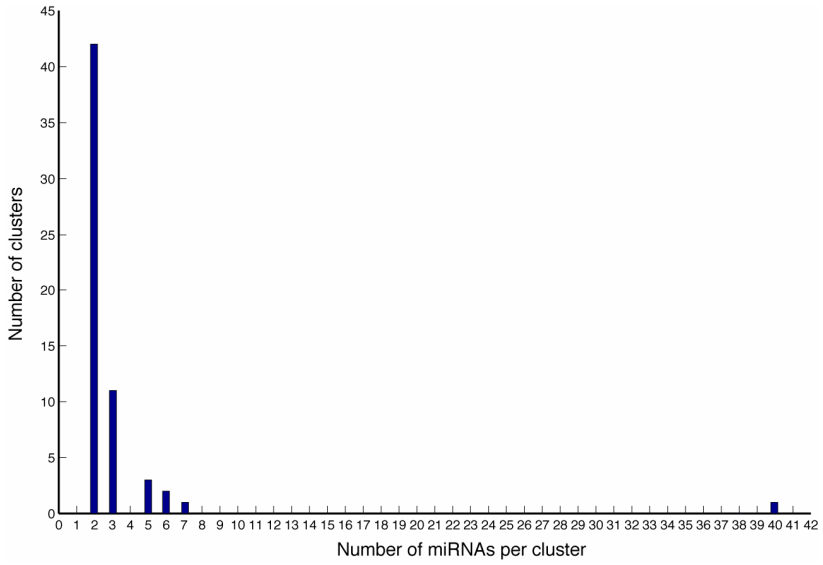


Fig. S2. Number of clusters of miRNA containing two or more miRNA. A total of 496 bovine miRNAs were grouped into 298 homolog families. miRNAs were considered to form a cluster if they were separated by less than 10 kbp. About half of the bovine miRNA occur in genomic clusters containing 2 to 7 miRNA genes with the exception of one large 43 kbp long cluster on BTA21:59,594,412-59,637,311 (Btau3.1) containing 40 miRNAs.

Supplementary References and Notes

S1. Rat Genome Sequencing Consortium, *Nature* **428**, 493-521 (2004).
S2. Sea Urchin Genome Sequencing Consortium, *Science* **314**, 941-952 (2006).
S3. Y. Liu *et al.*, *BMC Genomics* **10**, 180 (2009)
S4. P. Havlak *et al.*, *Genome Res* **14**, 721-732 (2004).
S5. A. Everts-van der Wind *et al.*, *Proc Natl Acad Sci U S A* **102**, 18526-18531 (2005).
S6. W. M. Snelling *et al.*, *Genome Biol* **8**, R165 (2007).
S7. H. Nilsen *et al.*, *Anim Genet* **39**, 97-104 (2008).
S8. A. Prasad *et al.*, *BMC Genomics* **8**, 310 (2007).
S9. Y. Kapustin, A. Souvorov, T. Tatusova, D. Lipman, *Biol Direct* **3**, 20 (2008).
S10. B. Kiryutin, A. Souvorov, *ISMB*, (2005).
S11. A. Souvorov, T. Tatusova, D. Lipman, *ISMB*, (2004).
S12. B. J. Haas *et al.*, *Nucleic Acids Res* **31**, 5654-5666 (2003).
S13. V. Curwen *et al.*, *Genome Res* **14**, 942-950 (2004).
S14. E. Birney, M. Clamp, R. Durbin, *Genome Res* **14**, 988-995 (2004).
S15. G. S. Slater, E. Birney, *BMC Bioinformatics* **6**, 31 (2005).
S16. A. A. Salamov, V. V. Solovyev, *Genome Res* **10**, 516-522 (2000).
S17. V. Solovyev, in *Handbook of Statistical Genetics,* D. J. Balding, M. Bishop, C. Cannings, Eds. (John Wiley & Sons, Chichester, England ; Hoboken, NJ, 2007), pp. 97-159.
S18. E. W. Sayers *et al.*, *Nucleic Acids Res*, (2008).
S19. http://www.repeatmasker.org.
S20. E. Blanco, G. Parra, R. Guigo, *Curr Protoc Bioinformatics* **chap. 4**, Unit 4 3 (2007).
S21. G. Parra *et al.*, *Genome Res* **13**, 108-117 (Jan, 2003).
S22. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389-3402 (1997).
S23. C. G. Elsik *et al.*, *Genome Biol* **8**, R13 (2007).
S24. A. Bairoch, B. Boeckmann, S. Ferro, E. Gasteiger, *Brief Bioinform* **5**, 39-55 (2004).
S25. M. S. Boguski, T. M. Lowe, C. M. Tolstoshev, *Nat Genet* **4**, 332-333 (1993).
S26. G. Pertea *et al.*, *Bioinformatics* **19**, 651-652 (Mar 22, 2003).
S27. W. R. Pearson, D. J. Lipman, *Proc Natl Acad Sci U S A* **85**, 2444-2448 (1988).
S28. http://bovinegenome.org/bovine_genome_consortium/datasets.html.
S29. Chicken Genome Sequencing Consortium, *Nature* **432**, 695-716 (2004).
S30. R. Guigo *et al.*, *Proc Natl Acad Sci U S A* **100**, 1140-1145 (2003).
S31. A. Reymond *et al.*, *Genomics* **79**, 824-832 (2002).
S32. S. Rozen, H. J. Skaletsky, in *Bioinformatics Methods and Protocols: Methods in Molecular Biology,* S. Krawetz, S. Misener, Eds. (Humana Press, Totowa, NJ, 2000), pp. 365-386.
S33. D. S. Gerhard *et al.*, *Genome Res* **14**, 2121-2127 (2004).
S34. R. L. Strausberg *et al.*, *Proc Natl Acad Sci U S A* **99**, 16899-16903 (2002).
S35. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, D. L. Wheeler, *Nucleic Acids Res* **36**, D25-30 (2008).
S36. P. Carninci *et al.*, *Genome Res* **10**, 1617-1630 (2000).
S37. T. S. Alioto, *Nucleic Acids Res* **35**, D110-115 (2007).
S38. T. D. Wu, C. K. Watanabe, *Bioinformatics* **21**, 1859-1875 (2005).
S39. S. Griffiths-Jones, H. K. Saini, S. van Dongen, A. J. Enright, *Nucleic Acids Res* **36**, D154-158 (2008).

40. http://blast.wustl.edu.
S41. E. Bonnet, J. Wuyts, P. Rouze, Y. Van de Peer, *Bioinformatics* **20**, 2911-2917 (2004).
S42. P. Steffen, B. Voss, M. Rehmsmeier, J. Reeder, R. Giegerich, *Bioinformatics* **22**, 500-503 (2006).
S43. I. L. Hofacker, *Curr Protoc Bioinformatics* **chap. 12**, Unit 12.2 (2004).
S44. K. Katoh, H. Toh, *BMC Bioinformatics* **9**, 212 (2008).
S45. B. Giardine *et al.*, *Genome Res* **15**, 1451-1455 (2005).
S46. D. Gerlach, E. V. Kriventseva, N. Rahman, C. E. Vejnar, E. M. Zdobnov, *Nucleic Acids Res*, (Oct 15, 2008).
S47. C. Notredame, D. G. Higgins, J. Heringa, *J Mol Biol* **302**, 205-217 (2000).
S48. W. R. Pearson, *Methods Enzymol* **183**, 63-98 (1990).
S49. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res* **35**, D61-65 (2007).
S50. P. Flicek *et al.*, *Nucleic Acids Res* **36**, D707-714 (2008).
S51. D. Karolchik *et al.*, *Nucleic Acids Res* **36**, D773-779 (2008).
S52. UniProt Consortium, *Nucleic Acids Res* **37**, D169 (2008).
S53. http://BovineGenome.org.
S54. S. M. Searle, J. Gilbert, V. Iyer, M. Clamp, *Genome Res* **14**, 963-970 (2004).
S55. S. E. Lewis *et al.*, *Genome Biol* **3**, RESEARCH0082 (2002).
S56. P. Bernaola-Galvan, R. Roman-Roldan, J. L. Oliver, *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **53**, 5181-5189 (1996).
S57. N. Cohen, T. Dagan, L. Stone, D. Graur, *Mol Biol Evol* **22**, 1260-1272 (2005).
S58. R. C. Edgar, E. W. Myers, *Bioinformatics* **21 Suppl 1**, i152-158 (2005).
S59. A. L. Price, N. C. Jones, P. A. Pevzner, *Bioinformatics* **21 Suppl 1**, i351-358 (2005).
S60. R. C. Edgar, *Nucleic Acids Res* **32**, 1792-1797 (2004).
S61. A. Stamatakis, T. Ludwig, H. Meier, *Bioinformatics* **21**, 456-463 (2005).
S62. T. H. Jukes, C. R. Cantor, in *Mammalian Protein Evolution,* H. N. Munro, Ed. (Academic Press, New York, 1969), pp. 21-123.
S63. K. Tamura, J. Dudley, M. Nei, S. Kumar, *Mol Biol Evol* **24**, 1596-1599 (2007).
S64. http://www.clcbio.com.
S65. L. Kraemer *et al.*, *BMC Bioinformatics* **10**, 41 (2009).
S66. http://espressosoftware.com/pages/sputnik.jsp.
S67. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**, 403-410 (1990).
S68. J. Jurka *et al.*, *Cytogenet Genome Res* **110**, 462-467 (2005).
S69. G. Talavera, J. Castresana, *Syst Biol* **56**, 564-577 (2007).
S70. S. Guindon, O. Gascuel, *Syst Biol* **52**, 696-704 (2003).
S71. Mouse Genome Sequencing Consortium, *Nature* **420**, 520-562 (2002).
S72. A. Reymond *et al.*, *Genomics* **78**, 46-54 (2001).
S73. M. Ashburner *et al.*, *Nat Genet* **25**, 25-29 (2000).
S74. A. Kasprzyk *et al.*, *Genome Res* **14**, 160-169 (2004).
S75. W. J. Kent, *Genome Res* **12**, 656-664 (2002).
S76. S. J. Humphray *et al.*, *Genome Biol* **8**, R139 (2007).
S77. W. J. Murphy *et al.*, *Science* **309**, 613-617 (2005).
S78. D. M. Larkin *et al.*, *Genome Research* **19**, 770-777 (2009).
S79. D. Thissen, L. Steinberg, D. Kuang, *Journal of Educational and Behavioral Statistics* **27**, 77-83 (2002).

S80. J. H. Edwards, *Ann Hum Genet* **55**, 17-31 (1991).
S81. N. D. Trinklein *et al.*, *Genome Res* **14**, 62-66 (2004).
S82. M. Q. Yang, L. M. Koehly, L. L. Elnitski, *PLoS Comput Biol* **3**, e72 (2007).
S83. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res* **11**, 1005-1017 (2001).
S84. X. She, Z. Cheng, S. Zollner, D. M. Church, E. E. Eichler, *Nat Genet* **40**, 909-914 (2008).
S85. J. A. Bailey *et al.*, *Science* **297**, 1003-1007 (2002).
S86. D. J. Lynn *et al.*, *Mol Syst Biol* **4**, 218 (2008).
S87. J. Felsenstein. (Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005).
S88. http://evolution.genetics.washington.edu/phylip.html.
S89. Z. Yang, *Mol Biol Evol* **24**, 1586-1591 (2007).
S90. L. D. Hurst, *Trends Genet* **18**, 486 (2002).
S91. H. Mi, N. Guo, A. Kejariwal, P. D. Thomas, *Nucleic Acids Res* **35**, D247-252 (2007).
S92. Rhesus Macaque Genome Sequencing and Analysis Consortium, *Science* **316**, 222-234 (2007).
S93. S. Seo, H. A. Lewin, *BMC Systems Biology* **3**, 33 (2009).
S94. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res* **22**, 4673-4680 (1994).
S95. E. Eyras *et al.*, *BMC Bioinformatics* **6**, 131 (2005).
S96. A. Levine, R. Durbin, *Nucleic Acids Res* **29**, 4006-4013 (2001).
S97. N. Sheth *et al.*, *Nucleic Acids Res* **34**, 3955-3967 (2006).
S98. E. A. Glazov, S. McWilliam, W. C. Barris, B. P. Dalrymple, *Mol Biol Evol* **25**, 939-948 (May, 2008).
S99. H. Seitz *et al.*, *Genome Res* **14**, 1741-1748 (2004).
S100. G. Bernardi, *Gene* **241**, 3-17 (2000).
S101. E. S. Lander *et al.*, *Nature* **409**, 860-921 (2001).
S102. O. Clay, G. Bernardi, *Mol Biol Evol* **22**, 2315-2317 (2005).
S103. W. Li, *Gene* **300**, 129-139 (Oct 30, 2002).
S104. W. Li, P. Bernaola-Galvan, P. Carpena, J. L. Oliver, *Comput Biol Chem* **27**, 5-10 (2003).
S105. http://cegg.unige.ch/cow_genome.
S106. Z. Kan, P. W. Garrett-Engele, J. M. Johnson, J. C. Castle, *Nucleic Acids Res* **33**, 5659-5666 (2005).
S107. B. Modrek, C. J. Lee, *Nat Genet* **34**, 177-180 (2003).
S108. Q. Pan *et al.*, *Trends Genet* **21**, 73-77 (2005).
S109. R. Sorek, R. Shamir, G. Ast, *Trends Genet* **20**, 68-71 (2004).
S110. T. A. Thanaraj, F. Clark, J. Muilu, *Nucleic Acids Res* **31**, 2544-2552 (2003).
S111. http://evolutionhighway.ncsa.uiuc.edu.
S112. http://oxgrid.angis.org.au/cattle/.
S113. J. Xie *et al.*, *Proc Natl Acad Sci U S A* **101**, 10750-10755 (2004).
S114. http://imgt.cines.fr.
S115. S. A. Ellis, K. T. Ballingall, *Immunol Rev* **167**, 159-168 (1999).
S116. S. A. Ellis *et al.*, *Immunogenetics* **50**, 319-328 (1999).
S117. J. Birch, G. Codner, E. Guzman, S. A. Ellis, *Immunogenetics* **60**, 267-273 (2008).
S118. http://www.ebi.ac.uk/ipd/mhc/bola/).
S119. J. Birch, C. De Juan Sanjuan, E. Guzman, S. A. Ellis, *Immunogenetics* **60**, 477-483 (2008).

S120. C. P. Childers *et al.*, *Anim Genet* **37**, 121-129 (2006).
S121. A. P. Ambagala, Z. Feng, R. G. Barletta, S. Srikumaran, *Immunogenetics* **54**, 30-38 (2002).
S122. P. D. Karp *et al.*, *Nucleic Acids Res* **33**, 6083-6089 (2005).
S123. D. M. Irwin, *J Mol Evol* **41**, 299-312 (1995).
S124. We thank funding officers for their ongoing efforts: J. Peterson (Project Officer), C. Bennet, A. Felsenfeld, M. Guyer, J. Malone, L. Wang, and K. Wetterstrand of NHGRI; R. D. Green (Project Officer) and S. M. Kappes of USDA ARS; D. Hamernik (Project Officer) of USDA CSREES; C. Bell (Genome Canada); R. Baker and A. Crawford (AgResearch Ltd.); B. Church (ASRA); E. Dressler of the National Beef Council; K. A. Eversole of Eversole & Associates; S. Moore; W. Roberts (State of Texas); R. Tellam (Project Officer) of CSIRO; and R. Wortham (Texas Beef Council).
S125. Analysis of the bovine genome was supported by the following sources of funding for individual consortium members. C. Elsik: USDA National Research Initiative (NRI) grant 2007-35616-17882, Kleberg Foundation, Georgetown University, the Texas Agricultural Experiment Station; H. Lewin: USDA Cooperative State Research Education and Extension Service, Livestock Genome Sequencing Initiative (AG 2005-34480-15939); E. Eichler: NIH grant HG002385; M. Rijnkels: USDA Agricultural Research Service Cooperative Agreement 6250-51000-048; E. Zdobnov and S. Antonarakis: The Swiss National Science Foundation; D. Lynn, M. Whiteside, and F. Brinkman: Genome Canada, Genome BC and the Michael Smith Foundation for Health Research; D. Bradley and L. Lau: Science Foundation Ireland; F. Brinkman: Canadian Institutes of Health Research; E. Memili: Mississippi Agricultural and Forestry Experiment Station; A. Iivanainen: The Academy of Finland (122540/2007) and Research Funds of The University of Helsinki (914/51/2006); C. Gill: USDA NRI grant 2007-35604-17870; S. Hiendleder and C. Fitzsimmons: JS Davies Bequest; J. Taylor and R. Schnabel: USDA NRI grants 2005-35205-15448, 2005-35604-15615, 2006-35205-16701 and 2006-35616-16697. This research was also supported in part by the Intramural Research Program of the NIH National Library of Medicine, EADGENE (EU Contract No. FOOD-CT-2004-506416), and Institute Strategic Grant funding from the Biotechnology and Biological Sciences Research Council.
S126. We thank the following individuals for contributions to this research: Jonathan Usmar for preparing the Oxford Grid; Shelia Alexander, Cal Davison, R.J. Hubbard, Whisper Kelly, Vicki Leesburg, Kathy Meidinger, Ryan Rienstra, Brooke Shipp, and Heidi Stroh for collection of tissue samples for cDNA sequencing. We thank the following individuals for contributions to sequence production: Carlana Clashette Allen, Ugonna Sharon Anosike, Ashton Vashawn Bell, Carla Bickman, Veronica Cardena,Kelvin Carter, Alejandra Chavez, Dean Chavez, Hau-Seng Chu, Raynard Cockrell, Mary Louise Davila, Latarsha Davy-Carroll, Shawn Denson, Victor E. Ebong, Sonia Fernandez, Pushpa Ranjani Fernando, Courtney Sherell Francis, Jason Garner, Ricardo M. Garcia III, Tiffany Evette Garrett, Brandy A. Harbes, Ebere Sylvia Onyirioha Hawkins, Marilyn Hogues, Barbara Hollins, LaToya Howell, Bennie Johnson, Laquisha Monique King, Haika Kisamo, Liza Alvarez Lago, Chuan-Yar Lai, Fremiet Lara, Fitzherbert Henderson Legall III, Thanh-Kim Thi Le, Dhammika Liyanage, Pamela London, Lorna M. Lorensuhewa, Renita C. Madu, Evangelina Martinez, Tittu Mathew, Christian Mercado, Iracema Cavazos Mercado, Mala Munidasa, Dinh Ngoc Ngo Phong Quoc Nguyen,

75

Ogechi O. Nwaokelemeh, Melissa Obregon, Evelyn Odeh,Chibueze G. Onwere, Andrea Alexandra Parra, Heidie A. Paul, Agapito Perez, Yolanda Yaneth Perez, Eltrick L. Primus, Maria Puazo, Juana B. Quiroz, Dina Rabata, Moazzam Sana,Brian W. Schneider, Ida Sisson, Richard P. Sorelle, Rosenie Thelus, Nicole Thomas, Rachel Diane Thorn, Reshaunda Devon Thornton, Zulma Y. Trejos, Kamran Usmani, Courtney Sherell White, Aneisa C. Williams. We thank the following individuals for contributions to sequence finishing: Guan Chen, Marcus D. Coyle, Alicia C. Hawes, Laronda R. Jackson, Ziad Mohid Khan, Zhangwan Li, Wen Liu, Lesette M. Perez, Hua Shen, Suzhen Wang, Qiaoyan Wang, Jennifer Eunyoung Watt, Jianling Zhou.

S127. Mention of trade names or commercial products is solely for the purpose of providing information and does not imply recommendation, endorsement or exclusion of other suitable products by any of the institutional participants.

A.1.3 *Integration of microRNA miR-122 in hepatic circadian gene expression*

Gatfield D, Le Martelot G, Vejnar CE, Gerlach D, Schaad O, Fleury-Olela F, Ruskeepää AL, Oresic M, Esau CC, Zdobnov EM, and Schibler U. Integration of microRNA miR-122 in hepatic circadian gene expression. *Genes Dev.* (2009) 23:1313–1326.

A.1.3.1 *Contributions*

Gatfield et al., 2009 presents the integration of the liver specific miRNA miR-122 in hepatic circadian gene expression.

I contributed to the study in analyzing secondary structures flanking the mir-122 locus in mouse (see on page 174) and investigating the pri-miRNA locus for further functional regions. Additionally, I prepared figures showing the conservation of the mir-122 promotor region in 32 mammal species (see on page 173).

A.1.3.2 *Main paper*

See pages 159–171 or at:
`http://genesdev.cshlp.org/content/23/11/1313.long`

Integration of microRNA miR-122 in hepatic circadian gene expression

David Gatfield,[1,10] Gwendal Le Martelot,[1,8] Charles E. Vejnar,[2,3,8] Daniel Gerlach,[2,3] Olivier Schaad,[4] Fabienne Fleury-Olela,[1] Anna-Liisa Ruskeepää,[5] Matej Oresic,[5] Christine C. Esau,[6] Evgeny M. Zdobnov,[2,3,7] and Ueli Schibler[1,9]

[1]Department of Molecular Biology, Sciences III, University of Geneva, 30, CH-1211 Geneva, Switzerland; [2]Department of Genetic Medicine and Development, University of Geneva Medical School, CH-1211 Geneva, Switzerland; [3]Swiss Institute of Bioinformatics, CH-1211 Geneva, Switzerland; [4]Genomics Platform, University of Geneva Medical School, CH-1211 Geneva, Switzerland; [5]VTT Technical Research Centre of Finland, FI-02044 VTT, Finland; [6]Regulus Therapeutics, Carlsbad, California 92008, USA; [7]Imperial College London, SW7 2AZ London, United Kingdom

In liver, most metabolic pathways are under circadian control, and hundreds of protein-encoding genes are thus transcribed in a cyclic fashion. Here we show that rhythmic transcription extends to the locus specifying miR-122, a highly abundant, hepatocyte-specific microRNA. Genetic loss-of-function and gain-of-function experiments have identified the orphan nuclear receptor REV-ERBα as the major circadian regulator of mir-122 transcription. Although due to its long half-life mature miR-122 accumulates at nearly constant rates throughout the day, this miRNA is tightly associated with control mechanisms governing circadian gene expression. Thus, the knockdown of miR-122 expression via an antisense oligonucleotide (ASO) strategy resulted in the up- and down-regulation of hundreds of mRNAs, of which a disproportionately high fraction accumulates in a circadian fashion. miR-122 has previously been linked to the regulation of cholesterol and lipid metabolism. The transcripts associated with these pathways indeed show the strongest time point-specific changes upon miR-122 depletion. The identification of *Pparβ/δ* and the peroxisome proliferator-activated receptor α (PPARα) coactivator *Smarcd1/Baf60a* as novel miR-122 targets suggests an involvement of the circadian metabolic regulators of the PPAR family in miR-122-mediated metabolic control.

[*Keywords*: Circadian; miRNA; miR-122; metabolism; clock; PPAR]

Supplemental material is available at http://www.genesdev.org.

Received January 12, 2009; revised version accepted April 20, 2009.

Light-sensitive organisms possess a circadian timekeeping system that serves to synchronize gene expression and physiology with geophysical time (Reppert and Weaver 2002; Gachon et al. 2004). Current models of the mammalian molecular clocks are based on two interlocked transcriptional feedback loops (Sato et al. 2006): a positive limb, in which the heterodimeric BMAL1:CLOCK transcription factor mediates the transcriptional activation of cryptochrome (*Cry1* and *Cry2*) and period genes (*Per1* and *Per2*), and a negative limb, in which PER:CRY complexes repress the BMAL1:CLOCK-mediated transcription of their own genes. Coordination between the two limbs is accomplished by nuclear receptors of the REV-ERB and ROR families (Preitner et al. 2002; Reppert and Weaver 2002; Sato et al. 2004). Cyclic *Rev-erbα* transcription is regulated by the mechanisms described above for *Cry* and *Per* genes, and the circadian accumulation of the repressor REV-ERBα results in the rhythmic repression of target genes, such as *Bmal1*, carrying retinoid-related orphan receptor elements (ROREs) (Ueda et al. 2002). In addition to these transcriptional feedback loops, numerous post-translational modifications of core clock proteins are known to contribute to the rhythm-generating clockwork circuitry (Gallego and Virshup 2007).

The cyclic expression of clock output genes can be governed directly by core clock components via E-box or RORE sequences (Ueda et al. 2002), or transcription factors such as PAR bZip proteins whose genes are regulated by these mechanisms (Gachon et al. 2004). However, despite the similar molecular makeup of the core oscillator in different organs, its outputs vary substantially between tissues (e.g., Storch et al. 2002). Gene expression profiling in liver has suggested that, depending on the algorithms used for the identification of cyclically expressed genes, 2%–10% of the transcriptome may be under circadian control (Panda et al. 2002; Storch et al.

[8]These authors contributed equally to this work.
Corresponding authors.
[9]E-MAIL ueli.schibler@unige.ch; FAX 41-22-3796868.
[10]E-MAIL david.gatfield@unige.ch; FAX 41-22-3796868.
Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.1781009.

2002; Kornmann et al. 2007a; Miller et al. 2007). Many of these genes are involved in hepatocyte-specific metabolic pathways.

In part, the synergistic activation of genes by circadian and tissue-specific transcription factors may account for the rhythmic expression of cell type-specific transcripts. However, tissue-specific post-transcriptional regulation of gene expression may also participate in this endeavor. It is estimated that in mammals ~30% of all mRNAs are subject to regulation by microRNAs (miRNAs) (Lewis et al. 2005), and miRNAs have been implicated in the post-transcriptional control of cellular proliferation, development, and differentiation (Bushati and Cohen 2007). miRNAs are short (~22 nucleotides [nt]), endogenous RNAs that promote translational repression and/or destabilization of target mRNAs (Bushati and Cohen 2007; Liu 2008). Target recognition occurs via base-pairing interactions with the 3′ untranslated region (UTR). Usually the 5′ portion of the miRNA forms a perfect hybrid with a 6- to 8-nt seed site, whereas the remainder of the miRNA undergoes interactions of only partial complementarity with the 3′UTR of its target mRNA (Lewis et al. 2005). The mismatches and gaps between miRNA and mRNA duplexes render the de novo prediction of miRNA targets challenging. Generally, a given miRNA can be expected to fine-tune the production of large sets of proteins within the cell (Baek et al. 2008; Liu 2008; Selbach et al. 2008).

Given the large fraction of mRNAs targeted by miRNAs, it is likely that miRNAs also modulate clock and clock output functions (Cheng et al. 2007; Xu et al. 2007; Yang et al. 2008). We wished to examine this conjecture and initiated our studies with miR-122, a miRNA that has been proposed to constitute up to 70% of all miRNA molecules in hepatocytes (Lagos-Quintana et al. 2002). The knockdown of miR-122 expression in mice and monkeys has previously been recognized to result in a down-regulation of cholesterol and lipid metabolizing enzymes and a reduction in plasma cholesterol levels (Krutzfeldt et al. 2005; Esau et al. 2006; Elmen et al. 2008a,b). Both lipid and cholesterol metabolism are well known for their daytime-dependent regulation, similar to many other hepatic functions that require coordination of food intake with nutrient-processing and energy homeostasis (Panda et al. 2002).

Here, we show that transcription of the miR-122 locus is under circadian control, involving the transcriptional repressor REV-ERBα. Thus, pri- and pre-miRNA precursors oscillated about fourfold to 10-fold in abundance during the day but accumulated at nearly constant levels in the livers of *Rev-erbα* knockout mice. However, due to its high stability mature miR-122 levels were virtually constant throughout the day. Despite the apparent invariable temporal accumulation of miR-122, the identification of its target mRNAs suggested that miR-122 nevertheless participates in the circadian control of many transcripts involved in hepatic metabolism. Among the miR-122 targets we found the mRNAs encoding peroxisome proliferator-activated receptor β/δ (PPARβ/δ) and SMARCD1/BAF60a, which are themselves circadian regulators of metabolism (Yang et al. 2006; Seedorf and Aberle 2007; Li et al. 2008).

Results

The miR-122 locus is transcribed in a circadian fashion

In a search for miRNAs that could shape the circadian expression of target mRNAs, we analyzed the expression of various miRNAs in mouse liver at different time points (Zeitgeber time, ZT) around the day. Several miRNAs (miR-19, miR-20, miR-22, miR-24, miR-30, miR-92, miR-126-3p), some of which had been predicted to target clock components (Lewis et al. 2005), only showed modest, if any, circadian changes in expression, as judged by Northern blot analysis (Supplemental Fig. 1). However, analysis of miR-122, the most abundant miRNA in liver, revealed that pre-mir-122 oscillated with an approximately fivefold daily amplitude in abundance, whereas mature miR-122 levels remained nearly constant over the day (Fig. 1A,B). Pre-mir-122 is a 66-nt hairpin-shaped precursor molecule from which the endonuclease Dicer cleaves the mature 22-nt miR-122. The mature miRNA is then incorporated into the RNA-induced silencing complex (RISC). The same expression pattern for pre-mir-122 was detected with a probe recognizing the strand complementary to the miRNA (known as the miRNA* sequence) (Fig. 1). The observed circadian changes in pre-mir-122 levels could be the result of either circadian synthesis or circadian processing into mature miRNA. To distinguish between these possibilities we analyzed the circadian levels of the miR-122 primary transcript, pri-mir-122, a ~5-kb precursor (Chang et al. 2004), from which the pre-miRNA is cleaved by the Drosha-containing microprocessor complex.

As shown in Figure 1A (bottom panels), pri-mir-122 accumulation was highly circadian (~10-fold amplitude), showing a similar phase as pre-mir-122 (i.e., minimal levels at ZT8-12 and maximal levels at ZT24). We wanted to test if high-amplitude circadian precursors were specific for miR-122 or were a common feature of miRNAs. Two other loci tested, pri-mir-17-92 and pri-mir-22, did not show the circadian pattern observed for pri-mir-122 (Supplemental Fig. 1C,D). This suggested that specifically the miR-122 locus was transcribed in a circadian fashion. The two intermediates in miR-122 biogenesis can be expected to be short-lived and reflect the rate at which the gene is transcribed. In contrast, the absence of cyclic expression at the level of mature miR-122 was probably due to its high metabolic stability. Indeed, based on Northern blot experiments, we estimated that the ratio of miR-122/pre-mir-122 steady-state levels (which is largely determined by the ratio of the half-lives of the two species) is in the range of 400:1. If one assumes that the pre-mir-122 half-life is a few minutes, this means that the miR-122 half-life is probably well beyond 24 h.

The orphan nuclear receptor REV-ERBα drives circadian mir-122 transcription

We wished to study the molecular mechanism accounting for circadian mir-122 transcription. The phase of

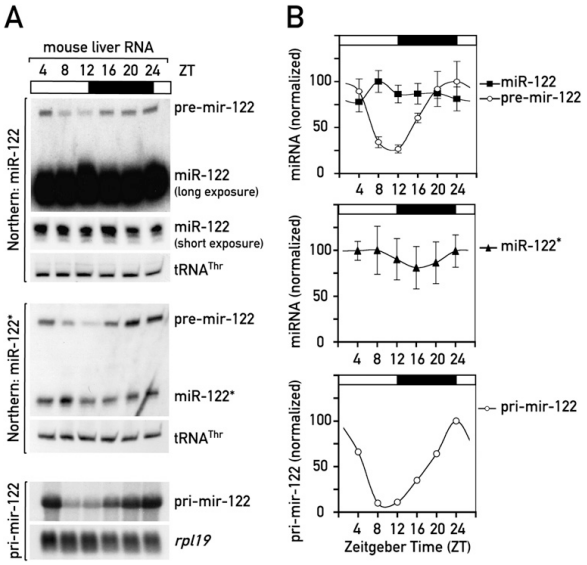


Figure 1. miR-122 precursors are circadian in mouse liver. (*A*) Northern blot analysis of miR-122 and its precursor RNAs using whole-cell RNA from male C57BL/6 mice sacrificed at the indicated ZT values around the clock. An RNA pool from three mice was used per time point, $tRNA^{Thr}$ and *rpl19* mRNA served as loading controls in denaturing polyacrylamide (*top* and *middle* panels) and agarose gel electrophoresis (*bottom* panels), respectively. (*Top* panels) miR-122 and pre-mir-122. (*Middle* panels) pre-mir-122 and miR-122*. miR-122* is the antisense "passenger strand" that is incorporated into RISC at low levels. (*Bottom* panels) pri-mir-122. (*B*, *top* and *middle* panels) miR-122, miR-122* and pre-mir-122 levels, normalized to $tRNA^{Thr}$, from Northern blots in which single animals were analyzed (data not shown). Mean values ± SEM. (*Bottom* panel) Quantification of pri-mir-122 levels, normalized to the circadianly invariant *rpl19*, from the Northern blot shown in *A*.

pri-/pre-mir-122 expression suggested that the circadian transcriptional repressor REV-ERBα might be involved: REV-ERBα protein expression peaks at around ZT8, leading to minimal transcript levels for REV-ERBα target genes at around ZT12 (Preitner et al. 2002; Ueda et al. 2002). Consistent with the hypothesis of miR-122 being a REV-ERBα target gene, the mir-122 promoter contains two conserved ROREs ~120–160 base pairs (bp) upstream of the transcriptional start site (Fig. 2A; see also Supplemental Fig. 2 for an alignment of the promoter region in 32 mammalian species). More importantly, the amplitude of cyclic pri-mir-122 accumulation was severely blunted in the livers of *Rev-erbα* knockout animals (Fig. 2B,C), and mature miR-122 accumulated to 1.6-fold higher levels (Fig. 2D). The residual amplitude in mir-122 transcription was possibly caused by REV-ERBβ, a highly related paralog of REV-ERBα (Preitner et al. 2002). A second mouse model, in which REV-ERBα was overexpressed specifically in hepatocytes (Kornmann et al. 2007a), showed the converse effect (i.e., 1.7-fold reduced miR-122 levels). In summary, these findings supported a model according to which the miR-122 locus is regulated by the circadian clock component REV-ERBα.

Does the miR-122 locus specify multiple functional RNAs?

Since the accumulation of pri-mir-122, but not that of mature miR-122, was rhythmic, we considered that this locus produced additional biologically active RNAs with shorter half-lives than miR-122. In fact, several pri-miRNAs are polycistronic and produce multiple miRNAs (Sewer et al. 2005). Although mature miR-122 shows a sequence conservation of 100% from fish to humans (Gerlach et al. 2009), its pri-miRNA gene structure is conserved only in mammals. In these organisms, the transcription start site is flanked by elements of a classical RNA polymerase II (PolII)-dependent promoter, which drives transcription of the ~5-kb capped and polyadenylated pri-mir-122 containing the pre-mir-122 hairpin at its 3′-end (Fig. 2A; Supplemental Figs. 2, 3A; Chang et al. 2004). Overall, the pri-mir-122 sequence is poorly conserved, and we did not detect additional potential miRNAs (or conserved open reading frames) within the primary transcript. A thorough bioinformatics search for conserved RNA secondary structures within the pri-mir-122 genomic locus in the genomes of six mammalian species also failed to identify additional RNA structures that could carry a function (Supplemental Fig. 3). Thus, it appeared likely that a potential biological function associated with the circadian control of pri-mir-122 transcription was mediated by miR-122 itself.

Genome-wide identification of miR-122 targets

As miR-122 was produced in a circadian fashion, we wondered whether it might assume rhythmic functions despite its long half-life. We decided to approach this question in an unbiased way by identifying putative miR-122 targets. In particular, we wished to determine whether there are targets whose daily rhythms are influenced by miR-122. To deplete miR-122, we injected antisense oligonucleotides (ASOs) intraperitoneally into mice (termed 122ASO in the following sections) and used genome-wide Affymetrix oligonucleotide arrays to determine the impact this had on hepatic mRNA levels. As

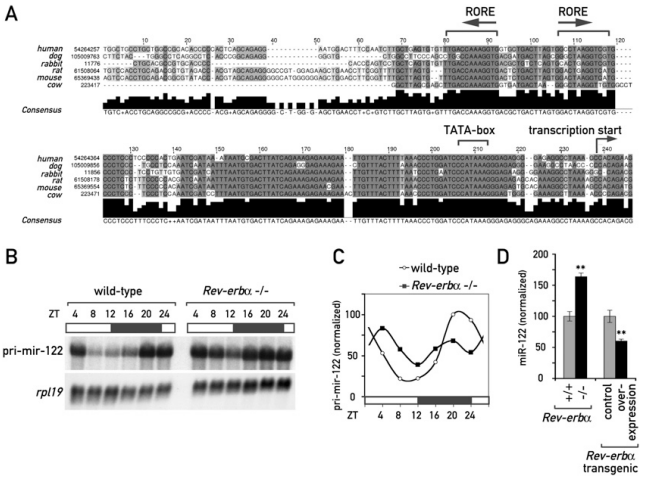


Figure 2. REV-ERBα is involved in circadian control of the miR-122 locus. (*A*) Alignment of the genomic sequence upstream of the predicted transcriptional start site of pri-mir-122 in six mammalian species (extracted from the University of California at Santa Cruz alignment; see Supplemental Fig. 3). The predicted ROREs, TATA-box, and transcriptional start site are indicated. (*B*) Northern blot analysis of pri-mir-122 in total RNA samples from *Rev-erbα* knockout and littermate control mice sacrificed at the indicated ZT values around the clock. For each time point, an RNA pool of three female mice was used. (*C*) Quantification of the Northern blot shown in *B*; values are pri-mir-122 normalized to *rpl19*. (*D*) miR-122 levels in total liver RNA from individual animals (mixed ZTs) of the indicated genotypes were quantified by Northern blot (data not shown). Control animals were set to 100%. Data are mean ± SEM ($n = 36$ for *Rev-erbα*$^{-/-}$ vs. *Rev-erbα*$^{+/+}$ and $n = 18$ for REV-ERBα overexpression vs. control); (**) $P < 5*10^{-5}$ (two-tailed Student's *t*-test).

controls, we used animals treated with ASOs targeting a miRNA not expressed in liver (miR-124; samples 124ASO in the following sections) or with PBS alone. Mice were sacrificed at time-points ZT0 and ZT12, when pri-mir-122 transcription was highest and lowest, respectively.

The efficiency of miR-122 depletion was between 89% and 99% as judged by Northern blot hybridization (Fig. 3A,B). Residual miR-122 levels were consistently lower for mice sacrificed at ZT12, when miR-122 production was low, suggesting that miRNA stability was decreased by 122ASOs. Importantly, the abundance of the unrelated miRNA let-7a remained unchanged (Fig. 3A), demonstrating the specificity of the ASO. To functionally assess if miR-122 was sufficiently depleted to derepress its targets, we determined the mRNA levels of the formerly suggested targets *glycogen synthase 1* (*Gys1*) and *aldolase A* (*AldoA*) by quantitative RT–PCR (qPCR). Similar to what had been observed previously (Krutzfeldt et al. 2005; Esau et al. 2006), these mRNAs were up-regulated twofold to sevenfold (Fig. 3C).

miRNAs initially have been proposed to mediate translational repression of their target mRNAs. This is often accompanied by a decrease in mRNA abundance (Baek et al. 2008; Selbach et al. 2008). Transcriptomal profiling using microarrays is therefore a convenient means to

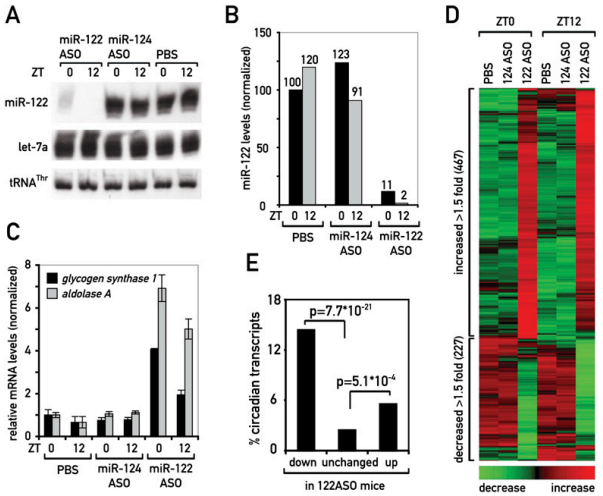


Figure 3. Analysis of miR-122 targets at two time points ZT0 and ZT12. (*A*) Northern blot analysis of miR-122, let-7a, and tRNA^{Thr} of mice treated with miR-122 ASO, miR-124 ASO, or PBS. Pools of RNA of three mice were loaded per lane. (*B*) Quantification of Northern blot shown in *A*. (*C*) qPCR analysis of RNAs from individual mice treated with the ASOs or PBS, as indicated. Probes used were for the known miR-122 targets *glycogen synthase 1* and *aldolase A*, normalized to 45S pre-rRNA. Values are mean ± SEM ($n = 3$). (*D*) Heat map of the probe sets up- and down-regulated in 122ASO-treated animals relative to both control groups, 124ASO- and PBS-treated animals (cutoff 1.5). The heat scale at the *bottom* of the panel represents changes on a linear scale, where green and red represent minimal and maximal expression, respectively. (*E*) Enrichment for circadian transcripts in the up- and down-regulated fractions in 122ASO mice. *P*-values were determined by a χ^2 test.

identify potential miRNA targets. Obviously, this technology is unable to detect miRNA targets whose translational attenuation is not accompanied by increased degradation.

Using Affymetrix microarray hybridization, we detected signals for a total of 22,384 probe sets, representing 11,638 transcripts. Among these, we found 343 transcripts (represented by 467 probe sets) that were up-regulated, and 188 transcripts (227 probe sets) that were down-regulated at at least one of the two time points in 122ASO-treated animals, when we applied a 1.5-fold expression change cutoff (Fig. 3D). We next analyzed whether transcripts up-regulated in 122ASO livers were enriched for potential miR-122 targets. For the prediction of potential miR-122-binding sites we applied a model that takes into account both the presence of miRNA seed sites and the energy of miRNA:mRNA duplexes, ensuring that energetically stable miRNA–target interactions are considered. Using this thermodynamic model (with an energy cutoff of −15 kcal/mol), we observed that 52% of transcripts in the up-regulated fraction contained a predicted miR-122-binding site (Supplemental Fig. 4). With only 22% of transcripts in the unchanged and 14% in the down-regulated fraction, this enrichment in the up-regulated fraction was statistically highly significant (up vs. unchanged: P-value $\sim 10^{-39}$; up vs. down: P-value $\sim 10^{-17}$). The differences between the unchanged and down-regulated fractions, however, were barely significant (P-value ~0.02). With a less elaborate model that only considers seed site presence, the enrichment for potential miR-122 targets in the up-regulated fraction was significant as well (Supplemental Fig. 4).

We next wished to determine, whether transcripts showing a time point-specific regulation by miR-122 could be clustered into particular metabolic pathways. To this end, we selected the transcripts that showed regulation upon 122ASO treatment exclusively at one of the two time points. Genome ontology (GO) analyses in the down-regulated fraction revealed that the genes involved in lipid and cholesterol metabolism (which had been reported previously to be most responsive to miR-122 depletion) also showed the strongest temporal regulation ($P \sim 10^{-10}$). Thus, the down-regulation of these mRNAs was significantly stronger at ZT12 than at ZT0 (Supplemental Fig. 5A). For up-regulated genes, transcripts belonging to GO:9607 "response to biotic stimuli" were most overrepresented ($P \sim 10^{-7}$). Their up-regulation occurred mainly at ZT0 and less so at ZT12 (Supplemental Fig. 5B). These observations suggested a considerable amount of cross-talk between circadian gene expression and miR-122, and encouraged us to analyze the effect of miR-122 depletion on circadian gene expression in greater detail.

Circadian transcripts are highly enriched among miR-122 targets

We wished to focus on transcripts that were direct potential targets of miR-122 and that showed circadian expression. For the genome-wide analysis of cyclic transcripts, we used previously reported transcriptome profiling experiments (Kornmann et al. 2007a). This work analyzes the hepatic transcriptome in a transgenic mouse model in which REV-ERBα can be conditionally overexpressed in liver in a doxycycline-dependent manner (tet-off system). In the presence of doxycycline, the hepatic circadian clock is functional in these animals, as the *Rev-erbα* transgene is constitutively repressed. The gene expression profiles from these animals, sampled over a 48-h period (with a resolution of 4 h), have been used to identify the circadian hepatic transcriptome using stringent algorithms (Kornmann et al. 2007a,b). In the absence of doxycycline, REV-ERBα overexpression arrests the endogenous liver clock. Thus, most circadian genes lose rhythmicity, with the notable exception of a small fraction of transcripts whose rhythms are driven by systemic cues rather than local oscillators (Kornmann et al. 2007a,b). In these mice, REV-ERBα overexpression also led to reduced miR-122 levels (Fig. 2D). It may thus be assumed that the derepression of miR-122 targets contributed to the gene expression changes observed upon REV-ERBα overexpression. We therefore compared the gene expression changes common to REV-ERBα overexpression and 122ASO administration. Of the transcripts whose abundance changed under both conditions, the majority (79.2%) indeed showed regulation in the same direction and only few (20.8%) showed reverse regulation (Supplemental Fig. 6). These observations lend further support to a role of REV-ERBα in miR-122 regulation.

We next analyzed the probe sets representing transcripts with circadian accumulation. Using stringent algorithms, these corresponded to ~2.8% of the liver transcriptome (Kornmann et al. 2007a,b). We found that the up- and down-regulated fractions in the 122ASO mice were significantly enriched for circadian transcripts: 14.4% of the down-regulated, 5.5% of the up-regulated, but only 2.4% of the unchanged mRNAs were among those classified as circadian (Fig. 3E). We thus concluded that the effects of depleting miR-122 were biased toward a misregulation of circadian transcripts. Since the enrichment was particularly high in the down-regulated fraction that contained indirect miR-122 targets, there were possibly common circadian regulatory mechanisms in control of this group. Indeed, almost a quarter of the transcripts in this fraction belonged to lipid/cholesterol metabolizing enzymes.

Identification of circadian mRNAs with functional miR-122-binding sites

We next investigated in more detail the group of transcripts with circadian accumulation that were up-regulated upon miR-122 depletion, as this subset was likely to contain the direct miR-122 targets (Fig. 4A). Within this group, 16 transcripts (specified by 19 probe sets) contained potential miR-122-binding sites in their 3′UTRs and were therefore candidates for circadianly expressed miR-122 targets (Fig. 4A, bold type). Many of them were also up-regulated in REV-ERBα-overexpressing

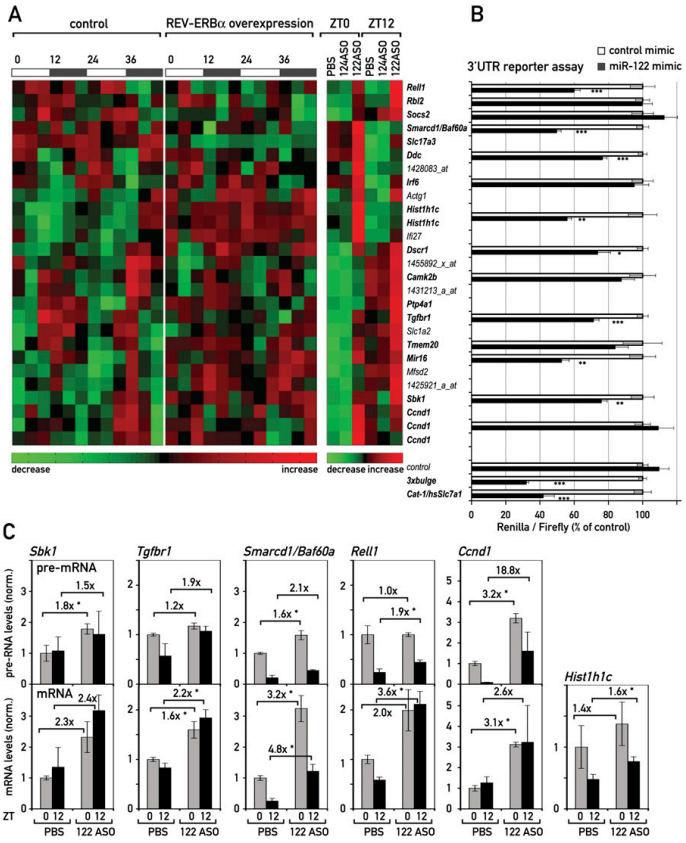


Figure 4. Circadian genes are miR-122 targets. (*A*) Heat map of the circadian probe sets (*left* and *middle* panel; taken from Kornmann et al. 2007b) that are up-regulated in 122ASO mice (*right* panel). *Smarcd1/Baf60a* was just below the stringent criteria used for circadian expression in the microarray data of Kornmann et al. (2007b), but was also included in the figure as it was confirmed as robustly circadian by qPCR (see Fig. 5). Heat scales at the *bottom* of the panels represent changes on a linear scale with green and red representing minimal and maximal expression, respectively. Transcripts in bold type contain potential miR-122 seed sites in their 3′UTRs. (*B*) The effect of miR-122 mimics in a 3′UTR luciferase assay. Control has only the vector 3′UTR, containing no seed sites. 3xbulge and *Cat-1/hsSlc7a1* are positive controls for 3′UTRs known to be regulated by miR-122. Values are mean ± SEM ($n \geq 6$ per transfection). (*) $P < 10^{-2}$; (**) $P < 10^{-3}$; (***) $P < 10^{-4}$ (two-tailed Student's *t*-test). (*C*) qPCR analysis in 122ASO mice and PBS controls of pre-mRNA (*top* panels) and mRNA (*bottom* panels) levels of selected transcripts from *A*. *Hist1h1c* is an intron-less gene; hence, pre-mRNA levels were not measured. Note that *Ccnd1* is also changed on the pre-mRNA level and is hence probably up-regulated by an indirect, transcriptional effect. Data are mean values of three mice per condition ±SEM. (*) $P < 10^{-2}$ (two-tailed Student's *t*-test).

animals (Fig. 4A, middle panel). We wished to verify that the changes in mRNA abundance detected by microarray analysis were potentially the direct result of miR-122 derepression, as opposed to more complicated indirect effects. Therefore, we tested the impact of miR-122 on the 3′UTRs of several candidate transcripts in cotransfection experiments. To this end, we cloned the candidate 3′UTRs into a vector carrying a renilla luciferase reporter gene, and transfected these constructs together with synthetic miRNA mimics into Hela cells, which do not express endogenous miR-122. We then measured the ability of a miR-122 mimic to inhibit the expression of luciferase when its open reading frame was followed by a particular 3′UTR. Two 3′UTRs known to be regulated by miR-122 served as positive controls: an artificial 3′UTR containing three optimized miR-122-binding sites (3xbulge) (Pillai et al. 2005), and the 3′UTR of *Cat-1/human Slc7a1*, a well-known miR-122 target (Chang et al. 2004; Bhattacharyya et al. 2006). These two 3′UTRs mediated a miR-122-dependent repression by about 68% and 58%, respectively. In contrast, luciferase expression from reporters harboring the vector-based 3′UTR devoid of miR-122 seed sites was not affected (Fig. 4B; Supplemental Fig. 7). Of the circadian transcripts up-regulated in 122ASO mice, we found that the 3′UTRs of *Rell1* (receptor expressed in lymphoid tissues-like 1), *Smarcd1/Baf60a* (SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily d, member 1/BRG1-associated factor 60a), *Ddc* (dopa decarboxylase), *Hist1h1c* (histone cluster 1, H1c), *Dscr1* (down syndrome critical region protein 1), *Tgfbr1* (TGF-β receptor type 1), *Mir16* (membrane-interacting protein of RGS16), and *Sbk1* (SH3-binding kinase 1) conferred sensitivity toward miR-122 (Fig. 4B). A complete compilation of the >30 3′UTRs we tested, including those of several newly identified miR-122 targets, is given in Supplemental Figure 7.

miR-122 contributes to circadian mRNA expression

For some selected targets, we wanted to verify that their up-regulation in the 122ASO mice was indeed caused by

post-transcriptional, rather than indirect transcriptional mechanisms. Since miRNAs are thought to act on processed mRNAs, a derepression mediated by the 122ASO should manifest itself on the level of the mature mRNA, but not on that of its pre-mRNA. Indirect effects, however, can be expected to occur through changes in transcription rates, caused by the up-regulation of activators or repressors whose production depends on miR-122. These changes should also be visible on the pre-mRNA level. Hence, we designed qPCR probes enabling us to measure mRNA and intron-containing pre-mRNA levels of several of the identified targets. Our analyses showed that the up-regulation of mature mRNA levels for the transcripts *Sbk1, Tgfbr1, Smarcd1/Baf60a, Rell1*, and *Hist1h1c* was similar, or even greater, than assessed by the microarray analysis. The effects of the 122ASO on pre-mRNA levels, however, were less pronounced (Fig. 4C). In contrast, a transcript such as *Ccnd1* fulfills the criteria for being indirectly affected. Thus, while *Ccnd1* was also circadian and up-regulated in 122ASO mice (Fig. 4A), it did not confer sensitivity to miR-122 in the 3′UTR assay (Fig. 4B). In keeping with this observation, the changes in *Ccnd1* expression were already observed on the level of *pre-Ccnd1* mRNA accumulation (Fig. 4C).

To evaluate more precisely which influence miR-122 had on shaping the rhythmic accumulation of these transcripts, we extended our analyses to 122ASO mice that had been sacrificed at six time points around the clock. Using RNA pools from three to four animals per time point and for both control and 122ASO mice (see Supplemental Fig. 8), we observed similar increases in target mRNA accumulation as in the previous two time point experiments (Fig. 5A; Supplemental Fig. 8D). In addition, it was apparent that miR-122 depletion had striking effects on the circadian amplitude (*Smarcd1/Baf60a, Ddc, Hist1h1c*), magnitude (*Rell1*) and phase (*Smarcd1/Baf60a, Hist1h1c*, and *Ddc*) of accumulation (Fig. 5A, bottom panels). For several transcripts (*Smarcd1/Baf60a, Ddc*, and *Hist1h1c*) we also observed that derepression caused an especially strong up-regulation at around ZT4 (Fig. 5A, bottom panels). This time point corresponds to a few hours after maximal mir-122 transcription (see Fig. 1B). Moreover, despite a particularly efficient miR-122 depletion at ZT12 (Fig. 3A,B; Supplemental Fig. 8), derepression clearly had a milder effect at this time point (Fig. 5A, bottom panels). For some of the miR-122 targets, these time-dependent effects were already observed in the microarray data (Fig. 4A). Due to their low abundance, the detection of the corresponding pre-mRNAs was less robust than that of the mature transcripts (Fig. 5A, top panels). Nevertheless, it was evident that (with the exception of *Ccnd1*) differences between 122ASO and control mice could not be accounted for by different transcription rates. These findings indicated that miR-122 probably assumes rhythmic functions despite its constant levels (see the Discussion). Importantly, the circadian clock per se did not appear to be affected by 122ASO treatment, as the mRNA levels of core clock and clock output genes were

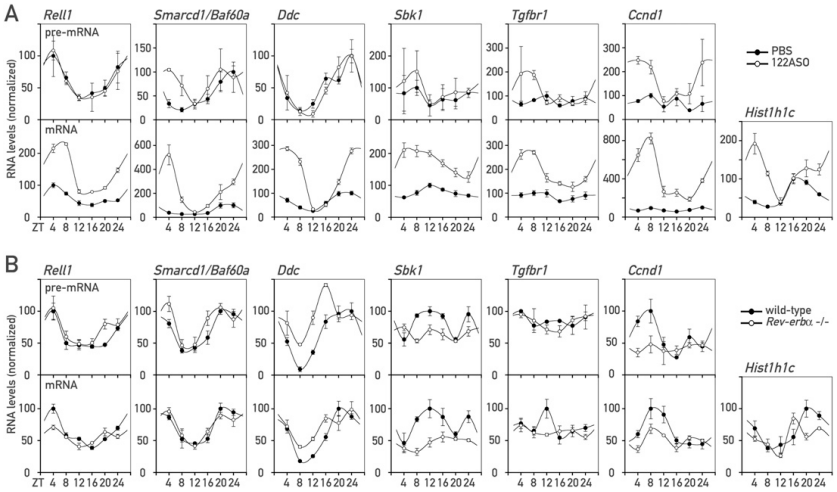


Figure 5. miR-122 targets in 122ASO and *Rev-erbα* knockout mice around the clock. (*A*) pre-mRNA (*top* panels) and mRNA (*bottom* panels) levels for the indicated transcripts in 122ASO and PBS-injected control mice around the clock. For each data point, transcript levels were measured in triplicate by qPCR using a pool of total liver RNA isolated from three to four mice. Due to low abundance, the detection of pre-mRNA levels was less robust, as indicated by generally larger error bars (standard deviations) in the qPCR analysis. (*B*) As in *A*, pre-mRNA (*top* panels) and mRNA (*bottom* panels) levels measured around the clock in *Rev-erbα* knockout and wild-type littermate animals, using a pool of whole-cell liver RNA isolated from five female mice.

essentially unchanged in 122ASO mice (Supplemental Fig. 9).

As shown in Figure 2, the circadian amplitude of mir-122 transcription was blunted in *Rev-erbα* knockout mice (Fig. 2B,C), leading to ~1.6-fold higher miR-122 levels (Fig. 2D). We wanted to examine whether these alterations in miR-122 production were sufficient to perturb the rhythmic expression of any of the targets analyzed above. When measuring mRNA and pre-mRNA abundances around the clock in *Rev-erbα* knockout and control animals, we observed the strongest post-transcriptional perturbations of rhythms for *Rell1*, a ubiquitously expressed member of the tumor necrosis factor (TNF) receptor family (Cusick et al. 2006). The amplitude of *Rell1* mRNA, but not that of its pre-mRNA, was blunted in the *Rev-erbα* knockout (1.7-fold amplitude) as compared with the control mice (2.6-fold amplitude) (Fig. 5B). This further supported our conclusion that *Rell1* was an example for an mRNA whose circadian rhythm was partially shaped by post-transcriptional mechanisms. In view of its up-regulation in 122ASO mice on the mRNA but not the pre-mRNA level (Figs. 4C, 5A) and the fact that the *Rell1* 3′UTR conferred sensitivity to miR-122 in the reporter assay (Fig. 4B), it is likely that miR-122 was directly implicated in this process. The cyclic accumulation of other miR-122 targets, such as *Smarcd1/Baf60a*, was unchanged in *Rev-erbα* knockout mice, whereas some changes already occurred on the pre-mRNA level (Fig. 5B).

Cross-talk between miR-122 and PPARs

The down-regulation of enzymes associated with lipid and cholesterol metabolism (see Krutzfeldt et al. 2005; Esau et al. 2006; Elmen et al. 2008a; this study) in miR-122-depleted mice implies that the corresponding mRNAs are regulated by indirect mechanisms. However, the direct miR-122 targets responsible for these control mechanisms remained to be identified. We suspected that these direct targets were also expressed in a circadian manner, since the down-regulation of mRNAs encoding lipid and cholesterol enzymes was daytime-dependent (Supplemental Fig. 5A). Interestingly, recent work has suggested that SMARCD1/BAF60a, a component of the SWI/SNF chromatin-remodeling complex, specifically regulates hepatic lipid metabolizing genes (Li et al. 2008). In our experiments, *Smarcd1/Baf60a* mRNA appeared as a circadian and direct miR-122 target, as it was robustly up-regulated in 122ASO mice (Figs. 4A,C, 5A) and as its 3′UTR was responsive to miR-122 in our cotransfection experiments (Figs. 4B, 6A). Li et al. (2008) further demonstrated that SMARCD1/BAF60a interacts and cooperates with the metabolic regulator PPARα, and that SMARCD1/BAF60a and PPARα share a large number of target genes.

PPARs belong to the nuclear hormone receptor superfamily and are well-known metabolic regulators. They are activated upon binding to their mainly amphipathic ligands, which are mostly derived from dietary fat or endogenous fatty acid metabolism. Of the three PPAR

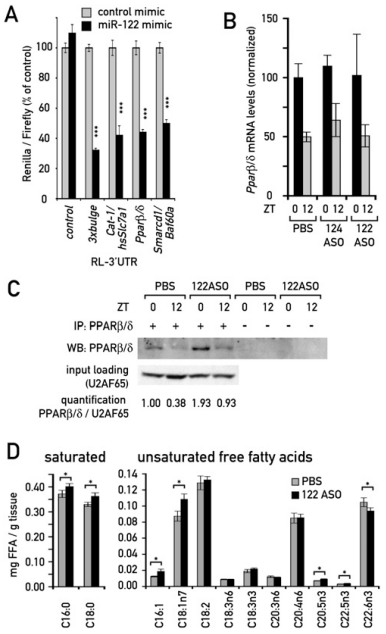


Figure 6. Cross-talk between miR-122 and PPAR receptors. (*A*) The effect of the miR-122 mimic in a 3′UTR luciferase assay as in Fig. 4B, using the *Pparβ/δ* and *Smarcd1/Baf60a* 3′UTRs. Values are mean ± SEM ($n \geq 9$ per transfection). (***) $P < 10^{-5}$ (two-tailed Student's *t*-test). (*B*) Expression levels of *Pparβ/δ* mRNA quantified from Northern blots. Data are mean ± SEM ($n = 3$ animals per condition). (*C*) Immunoprecipitation-Western blot of PPARβ/δ protein from 122ASO and PBS-treated mice, as described in the Materials and Methods. Each immunoprecipitation was performed from a pool of extracts from three mice. U2AF65 protein levels in the input of the same pool served as a loading control. (*D*) FFA levels in liver pieces from 122ASO- and PBS-injected animals, as determined by GC/MS. Values are mean ± SEM ($n = 6$). (*) $P < 0.05$ (two-tailed Student's *t*-test).

isotypes, PPARα, and the less-studied PPARβ/δ, serve predominantly catabolic functions, whereas PPARγ mainly promotes lipid storage in adipose tissue. In liver, all PPARs show circadian expression (Yang et al. 2006). Although we did not find PPAR transcripts misregulated using microarrays with RNA from 122ASO mice, we noticed that the *Pparβ/δ* 3′UTR contained four miR-122 seed sites that could be predicted to confer strong targeting by miR-122. We therefore tested if the *Pparβ/δ* 3′UTR showed sensitivity to miR-122 in our cotransfection experiments. Indeed, this 3′UTR caused a miR-122-dependent reduction of luciferase activity by 56%, which was among the highest down-regulation effects we observed in these assays. Only the two positive controls, the artificial 3xbulge and the *Cat-1/human Slc7a1* 3′UTR showed a slightly stronger repression (Fig. 6A; Supplemental Fig. 7). Consistent with these findings, we observed that whereas *Pparβ/δ* mRNA levels remained

unchanged upon miR-122 depletion (Fig. 6B), the protein was up-regulated around twofold to threefold, as judged by Western blot experiments with 122ASO liver extracts (Fig. 6C). These findings strongly suggested that *Pparβ/δ* was a bona fide miR-122 target that thus far had been overlooked, supposedly because it is not regulated on the level of mRNA stability.

Unsaturated fatty acids are probably the most important endogenous PPAR ligands, and their levels are known to be tightly regulated in vivo. The perturbation of lipid metabolism associated with miR-122 depletion may thus also lead to changes in PPAR ligand availability. We therefore determined the concentrations of free fatty acids (FFAs) in livers from 122ASO and control mice by GC/MS. Several unsaturated FFA species were indeed significantly changed, including palmitoleic acid (C16:1; up by 51% in 122ASO mice) and vaccenic acid (C18:1n7; up by 24%) (Fig. 6D). The latter constitutes a significant proportion of the total unsaturated FFA pool and has previously been proposed as a PPARβ/δ ligand (Fyffe et al. 2006). We therefore deemed it likely that PPAR activity in 122ASO mice was additionally modulated by changes in ligand concentration.

We conclude that miR-122 has several ties to the PPAR family of nuclear receptors, via *Pparβ/δ*, *Smarcd1/Baf60a*, and possibly ligand availability. Given the important functions PPARs possess in regulating metabolism in liver, these connections are very likely to contribute to the overall metabolic phenotype observed in 122ASO mice.

Discussion

Circadian mir-122 transcription and function

In the present study, we show that miRNA miR-122 expression and function are embedded in the output system of the circadian clock. Thus, we found that the miR-122 locus was transcribed in a circadian manner, manifesting itself in rhythmic pri-mir-122 and pre-mir-122 expression. Based on genetic loss-of-function and gain-of-function experiments we concluded that the orphan receptor REV-ERBα is the dominant regulator of circadian mir-122 transcription. On a genome-wide scale, we observed that the portion of the transcriptome sensitive to miR-122 depletion was highly enriched for circadian mRNAs, and it appeared that these were biased toward specific circadian phases (Supplemental Fig. 10). This temporal gating was particularly evident for mRNAs encoding cholesterol and lipid metabolizing enzymes, which were identified previously as indirectly regulated miR-122 targets (Krutzfeldt et al. 2005; Esau et al. 2006; Elmen et al. 2008a,b). Further analyses of individual up-regulated transcripts around the clock enabled us to identify several circadian transcripts that were likely candidates for direct miR-122 targets. The rhythmic accumulation of these mRNAs showed changes in amplitude (*Smarcd1/Baf60a*, *Ddc*, and *Hist1h1c*), magnitude (*Rell1*), and phase (*Smarcd1/Baf60a* and *Hist1h1c*) upon miR-122 depletion. In *Rev-erbα* knockout animals, miR-122 synthesis was nearly constant over the day and steady-state miR-122 levels were 1.7-fold elevated. REV-ERBα regulates many clock-controlled genes directly by repressing their transcription in a cyclic manner (see also Supplemental Fig. 6; G Le Martelot, T Claudel, O Schaad, B Kornmann, G Lo Sasso, A Moschetta, and U Schibler, in prep.). Irrefutable evidence that changes in miR-122 levels and/or production account for the circadian misregulation of target transcripts in *Rev-erbα* knockout mice is therefore difficult to obtain. However, as indicated by our analysis of pre-mRNA and mRNA expression, miR-122 misregulation is likely to be responsible for the altered circadian amplitude of *Rell1* mRNA accumulation in *Rev-erbα* knockout mice. Interestingly, *Rev-erbα* knockout mice show a cholesterol- and lipid-related phenotype opposite to 122ASO mice (G Le Martelot, T Claudel, O Schaad, B Kornmann, G Lo Sasso, A Moschetta, and U Schibler, in prep.). Again, REV-ERBα probably regulates these pathways mainly by more direct, transcriptional mechanisms, but miR-122 up-regulation is likely to contribute to these phenotypes as well.

The regulation of lipid metabolism by miR-122 may involve PPAR receptors

The direct miR-122 targets involved in hepatic lipid metabolism have not yet been identified. The decrease in hepatic fatty acid and cholesterol synthesis and the increase in hepatic fatty acid oxidation are paralleled by an increased activation of AMP-activated protein kinase (AMPK) in 122ASO mice (Esau et al. 2006). Thus, miR-122 may act through the modulation of this central sensor of metabolism. Our experiments also uncovered several connections of miR-122 to the nuclear receptors of the PPAR family, which are well-known regulators of metabolism. Specifically, we found that upon miR-122 inactivation, PPARβ/δ protein was up-regulated by around twofold to threefold. The *Pparβ/δ* 3′UTR contains seed sites for miR-122, and among the >30 3′UTRs we tested, it conferred one of the strongest levels of miR-122-mediated repression. In liver, PPARα and PPARβ/δ, the two PPARs executing catabolic functions, are both expressed in a circadian manner with a phase difference of ~8 h (Yang et al. 2006). Since PPARα is the predominant isoform in this organ, hepatic functions of PPARβ/δ have not yet been studied in detail. Although PPAR functions can vary in different tissues, it is interesting to note that recently an interaction between the PPARβ/δ and AMPK pathways was shown in muscle. Thus, a constantly active VP16-PPARβ/δ transgene led to constitutive AMPK stimulation (Narkar et al. 2008). Therefore, it is tempting to speculate that at least in part the AMPK activation (Esau et al. 2006) could be the result of higher PPARβ/δ protein levels in the livers of miR-122-depleted mice.

The newly identified miR-122 target *Smarcd1/Baf60a* provides a second link to PPARs. *Smarcd1/Baf60a* is a core subunit of the SWI/SNF chromatin remodeling complexes and was very recently identified in a screen for transcription factors whose activity is augmented by

PPARγ coactivator-1α (PGC-1α) (Li et al. 2008). In this study, SMARCD1/BAF60a overexpression in hepatocytes was shown to have surprisingly specific effects on the transcriptional activation of genes involved in fatty acid oxidation, and many of these were also activated by a synthetic PPARα agonist. In addition, SMARCD1/BAF60a was found to physically interact with PPARα and to be required for its function. Both proteins are corecruited with PGC-1α to PPAR response element (PPRE)-containing promoters.

A third connection of miR-122 to PPARs was provided by the observation that the livers of miR-122-depleted animals contained higher levels of FFAs, known to serve as PPAR ligands. All in all, our results suggest that PPARs might act as mediators to link miR-122 function to the control of circadian gene expression and hepatic lipid metabolism, although the detailed genetic and biochemical dissection of this network will require many additional experiments.

Speculations about the circadian action of miR-122

Generally, RISC-bound miRNAs are thought to be long-lived (Lee et al. 2003; Lund et al. 2004), and we estimated the miR-122 half-life to exceed 24 h. In the light of the large, stable steady-state pool of miR-122 the question thus arises of how circadian miR-122 production could nevertheless have an impact on its targets. Below we present three possible mechanisms through which miR-122 could modulate circadian gene expression on the post-transcriptional level. The first involves the attenuation of basal mRNA accumulation and/or translation by invariant miR-122 activity, the second implies chemically different subpopulations of miR-122 with distinct purposes, and the third a different availability of newly assembled and old RISC complexes for being loaded on target mRNAs.

Strictly speaking, there is no requirement for a miRNA to be circadian itself in order to contribute to the circadian accumulation of a target transcript. For example, the constant repression of basal levels of translation from such mRNA could strongly increase the circadian amplitude of the produced protein, as depicted in Figure 7A. Such a mechanism could have the additional benefit of conferring robustness to low protein expression in the trough: Rather than relying on very low transcription rates, which inevitably contain a stochastic component, low expression levels might thus be achieved more

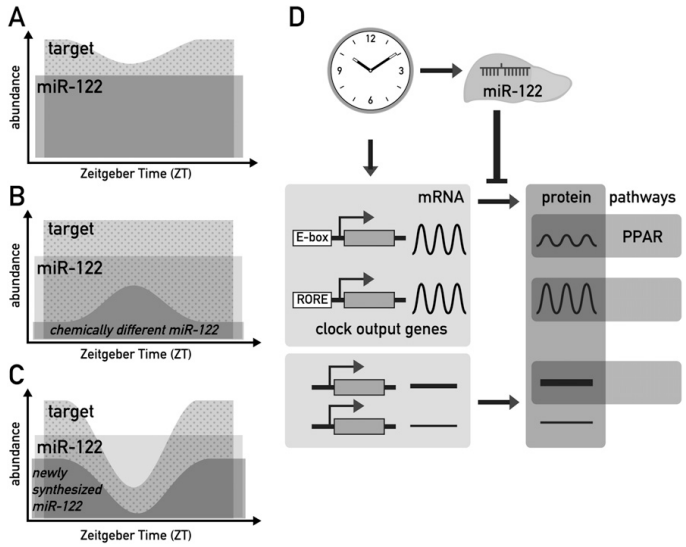


Figure 7. Models for how miR-122 could impart on circadian gene expression of its targets. (*A*) Even constant miR-122 levels (dark gray) could shape the circadian rhythm of a target (light gray) by constantly repressing basal levels of translation (represented by the overlap of the two areas). Only the amount of target mRNA represented by the dotted area would be available for translation into protein. As shown in the cartoon, this mechanism could increase the amplitude of cycling and convert a low-amplitude mRNA rhythm into a higher amplitude protein rhythm. In addition, this regulation could confer robustness to low protein expression levels in the trough, as described in the Discussion. The mechanism depicted in this cartoon would require a high affinity of the miRNA–target interaction and an excess of targets over the miRNA. Considering that miR-122 probably has hundreds of targets, of which many contain several seed sites, this assumption is quite plausible even for this highly abundant miRNA (*B*) Conceivably, chemically distinct, short-lived miR-122 subpopulations (dark gray) could exist within the pool of bulk miR-122. If these distinct miR-122 species also had specific functional properties, this speculative model would imply that target mRNAs would be subject to circadian repression. Consequently, the transcript available to produce protein (dotted area) would show circadian oscillations. (*C*) Conceivably, newly assembled RISC complexes could immediately get committed to their target mRNAs and remain stably associated with them, as described in the Discussion. The availability of such newly assembled miR-122 RISC would be expected to closely follow circadian miR-122 production (dark gray). If targets are transcribed circadianly as well, the phase relationship of the two rhythmic processes will determine to what extent a target will encounter miR-122 RISCs in the cell, and what influence this has on the circadian amplitude, magnitude, and phase of the produced protein (dotted area). As in *A*, this model would demand that the miRISC–mRNA affinity be high and that the targets are in excess. (*D*) Model for the integration of miR-122 in circadian hepatic gene expression. MiR-122 is depicted as a tissue-specific modulator of circadian output genes, with PPAR-dependent regulation of gene expression as one of the regulated output pathways.

precisely by simultaneously producing a mRNA and its inhibiting miRNA, which will partially annul each other. A related role for miRNAs in denoising and conferring robustness to gene expression has previously been suggested in developmental timing (Stark et al. 2005; Cohen et al. 2006; Li et al. 2006, 2009). With regard to liver-specific miR-122, this mechanism could also represent a way of modulating the circadian rhythm of outputs in a tissue-specific manner.

As mentioned above the high metabolic stability of miR-122 prevents its cyclic accumulation. However, our experiments do not exclude that functionally distinct, less stable subpopulations exist within the large pool of miR-122 molecules. This speculative scenario is schematically depicted in Figure 7B (see the figure legend for explanation). Although we currently have no direct evidence for such distinct miR-122-containing RISC subpopulations, they could be produced by miRNA editing, RISC protein composition or subcellular localization. It is interesting to note in this context that miR-122 was recently shown to undergo cytoplasmic 3′ adenylation, affecting miR-122 stability (Katoh et al. 2009). Hence, different miR-122 subpopulations with varying metabolic stabilities may indeed coexist. *Nocturnin*, a rhythmically expressed deadenylase, is also involved in the regulation of lipid metabolism (Green et al. 2007). Although bona fide target mRNAs have not yet been identified for this enzyme, the regulation of poly(A) length is known to contribute to translational repression also in the case of miRNA-mediated mechanisms (Liu 2008; Eulalio et al. 2009). Circadian deadenylation may thus also contribute to the post-transcriptional control of protein synthesis. It should be emphasized in this context that almost half of the cycling liver proteins identified by mass spectrometry are translated from stably expressed mRNAs (Reddy et al. 2006).

Recent work has suggested that the ternary RISC–miRNA–target complex is remarkably stable, allowing for the immunopurification of RISC-bound targets (e.g., Beitzinger et al. 2007; Karginov et al. 2007). One might therefore speculate that mainly uncommitted, "fresh" miRNA-loaded RISCs are available for the silencing of newly synthesized targets, whereas "old" RISCs, which are already engaged in silencing, are less so (Fig. 7C). Since "fresh" miR-122 RISC is produced in a circadian fashion, the extent of target capture and silencing may well be daytime-specific. If targets are transcribed circadianly as well, it becomes evident that the phase relationship of the two rhythmic processes will determine to what extent a target will encounter miR-122 RISCs in the cell. For several of the circadian miR-122 target profiles we determined in 122ASO livers (e.g., *Smarcd1/Baf60a* and *Ddc*; see Fig. 5A), the factor of up-regulation upon miR-122 depletion was indeed especially high around ZT4, just after the peak of miR-122 production. These transcripts were transcribed in phase with miR-122, and miR-122 could function to buffer against and counteract too extreme target oscillations. Future experiments will need to address whether and to what extent the three miR-122-related mechanisms contribute to the post-transcriptional regulation of circadian output rhythms.

Materials and methods

Animal care and treatment

Animal studies were conducted in accordance with the regulations of the veterinary office of the State of Geneva. Mice were maintained under standard animal housing conditions (12-h light/12-h dark cycles; free access to food/water). *Rev-erbα* knockout/transgenic mice have been described (Preitner et al. 2002; Kornmann et al. 2007a). ASO treatment was performed in 11-wk male C57BL/6 mice (Elevage Janvier) by intraperitoneal injection. ASOs were chimeric 2′-fluoro/2′-O-methoxyethyl-modified oligonucleotides with a completely modified phosphorothioate backbone. The exact chemistry is available on request. Mice received four doses of 20 mg of ASO per kilogram of body weight in 150 μL, or 150 μL of saline alone (PBS control), over the course of 2 wk. Two days to 3 d after the last injection, animals were sacrificed at the respective ZTs, and livers were snap-frozen in liquid nitrogen.

RNA analysis

RNA was prepared as in Kornmann et al. (2007a), except that the LiCl wash was omitted to prevent loss of small RNAs. mRNA Northern blots were performed as in Kornmann et al. (2007a). Single-stranded ^{32}P-labeled DNA probes were generated by linear PCR using standard methods. Templates were obtained by PCR amplification from liver cDNA or genomic DNA using gene-specific oligonucleotides (Supplemental Table 1). For miRNA Northern blots 10–30 μg of total RNA per sample were separated by 15% denaturing PAGE/1× TBE, electroblotted (36 min; 3.3 mA/cm^2; 0.5× TBE; 4°C) to Genescreen Plus (NEN) membrane, and immobilized by UV and baking. Hybridizations with radioactively labeled oligonucleotide probes were performed overnight in 5× SSC, 20 mM Na phosphate at pH 7.2, 7% SDS, 2× Denhardt's solution at 50°C, followed by four 15-min washes (3× SSC, 25 mM Na phosphate at pH 7.5, 5% SDS, 10× Denhardt's) and a 5-min wash with 1× SSC and 1% SDS. The sequences of oligonucleotide probes are listed in Supplemental Table 1. Quantification of Northern blots was performed by phosphorimaging using Quantity One Software (Bio-Rad).

Global transcriptome profiling using Affymetrix oligonucleotide microarrays

Whole-cell liver RNAs from ASO-injected mice (ZT0 and ZT12) were analyzed individually on a total of 18 microarrays. Five micrograms of RNA were employed for the synthesis of biotinylated cRNA, of which 8.75 μg were hybridized to Affymetrix Mouse Genome 430 2.0 arrays according to the supplier's instructions. To identify differentially expressed transcripts, pairwise comparisons were carried out using Affymetrix GCOS 1.2 software. Transcripts were considered as expressed if they were detectable in at least two of three replicates in at least one of the experimental conditions. To compare two experimental conditions, each of the triplicates of one condition was compared with the triplicates of the other condition, resulting in nine pairwise comparisons. This approach is based on the Mann-Whitney pairwise comparison test, and allows the ranking of results by concordance and the calculation of significance (*P*-value) for each identified change in gene expression (Hubbell et al. 2002; Liu et al. 2002). Genes whose concordance in the pairwise comparisons exceeded the imposed threshold of 77% (seven of nine comparisons) were considered to be statistically significant. Transcripts were considered as up- or down-regulated in 122ASO samples when their accumulation had an average

change of at least 1.5-fold with regard to both control samples, 124ASO, and PBS. The extraction of circadian genes from Affymetrix data sets (Fig. 4A) has been described previously (Kornmann et al. 2007b). The ArrayExpress repository (http://www.ebi.ac.uk/arrayexpress) accession number for the microarray data is E-TABM-692.

qPCR analysis

cDNA was synthesized from 2 μg of DNase-digested whole-cell RNA using random hexamers and SuperScript II reverse transcriptase (Invitrogen) following the supplier's instructions. cDNAs were PCR-amplified (7900HT Sequence Detection Systems, Applied Biosystems) using TaqMan Universal Master Mix, No AmpErase UNG (Applied Biosystems), and raw threshold cycle (Ct) values were calculated with SDS 2.0 software (Applied Biosystems). Mean levels were calculated from triplicate PCR assays for each sample and normalized to those obtained for the control transcripts *Eef1a1*, *Gapdh*, *GusB*, and 45S pre-rRNA. RT^- samples were included to exclude contaminations with genomic DNA. For primers and probes, see Supplemental Table 1.

miR-122 target predictions and enrichment statistics

We relied on the Ensembl version 50 mapping of the Affymetrix probes to transcripts. Up-regulated, down-regulated, and unchanged transcripts were selected as described above. The seed sequence of the miRNA was defined as 6–8 bases from the second position of the miRNA 5′-end, not allowing mismatches except a single G:U in 7-mers and 2 G:U in 8-mers. Duplex energies were computed with the cofolding function from the RNA Vienna Package (Hofacker 2003). The statistical significance of the putative miR-122 target site enrichment in the up, equal, and down fractions was evaluated using a χ^2 test.

Plasmids, clonings, and analysis of 3′UTRs

3′UTR sequences were amplified by PCR from mouse liver cDNA or genomic DNA with specific oligonucleotides (Supplemental Table 1) and cloned 3′ to the renilla luciferase (RL) sequence in vector pRL-control. The identity of the UTRs was verified by sequencing. Plasmids pRL-control and pRL-Cat-1 are as in Bhattacharyya et al. (2006) and pRL-3xbulge is similar to the homonymous plasmid in Pillai et al. (2005), except that bulges match miR-122 instead of let-7. For normalization, a CMV-driven firefly luciferase-expressing plasmid on the basis of pEGFP-C1 was used. Details on all plasmids are available on request. For 3′UTR assays, 2 ng of pRL, 40 ng of FL plasmid, and 10 pmol of miRNA mimic (miR-122 and control mimic cel-miR-67 from Dharmacon) were transfected into 10^4 HeLa cells per well of a 96-well plate by reverse transfection using Lipofectamine 2000 (Invitrogen) according to the supplier's instructions. Transfection mixes were replaced by normal growth medium after 6 h. Luciferase activities were measured 28 h after transfection with the Dual-Glo Luciferase Assay System (Promega). Renilla luciferase signals were normalized to firefly luciferase and for each 3′UTR construct set to 100% for the cotransfection with the control mimic. Each transfection was repeated at least six times. Growth medium was DMEM, 10% FCS, 1% PSG (Gibco).

Immunoprecipitation-Western blotting

Liver pieces of 122ASO and control mice, ZT0 and ZT12 (triplicates) were homogenized in three volumes of RIPA (150 mM NaCl, 1% NP40, 0.5% Na-deoxycholate, 0.1% SDS, 50 mM Tris-HCl at pH 8.0, protease inhibitors) using a motorized hand tool (Xenox). Insoluble material was removed by centrifugation (15 min, 20,000*g*, 0°C). Supernatants were kept at −80°C. For the immunoprecipitation, extracts were further diluted to 5 vol of RIPA per volume of liver and adjusted to 0.2% SDS. After another spin (as above), equal amounts of protein extract from the triplicates of the same experimental condition were pooled (~600 μg protein/liver). An aliquot was kept for the input sample, and immunoprecipitation was performed from the remaining pool using standard protocols with a rabbit polyclonal antibody to PPARβ/δ (ab8937, Abcam) and protein A-agarose (Roche). Immunoprecipitated complexes and inputs were analyzed by SDS-PAGE/Western blotting using antibodies to PPARβ/δ and U2AF65 (Sigma, U4758). Semiquantitative analysis of Western blots was performed using Quantity One Software (Bio-Rad).

FFA analysis

Liver homogenates in MeOH (0.1% BHT) were spiked with heptadecatrienoate, TAG (17:0/17:0/17:0) and heptadecanoic acid (FFA C17:0) and extracted by chloroform after addition of 0.9% sodium chloride. The lower organic phase was separated, evaporated under nitrogen flow, and dissolved into petroleum ether (bp 40°C–60°C). The samples were transesterified with sodium methoxide (NaOMe, 0.5 M in MeOH), acidified (15% $NaHSO_4$ in H_2O) and extracted with petroleum ether. The organic phase containing fatty acid methyl esters (FAME) from bound fatty acids and FFAs was separated, evaporated under nitrogen flow, and redissolved into hexane. Two-microliter aliquots were used for GC injection (splitless 1 min) at 280°C and the analyses were performed on an FFAP fused silica capillary column (25 m, i.d. 0.32 mm) by using helium as the carrier gas (pressure program). The oven temperature was increased from 70°C to 240°C at 7°C per minute, and the fatty acids were detected by flame ionization detector (FID, 300°C). Identification was based on retention times and GC/MS spectra of reference substances.

Acknowledgments

We thank Suvendra Bhattacharyya and Witek Filipowicz for plasmids, staff at the NCCR Genomics Platform for help with microarray/qPCR experiments, Tuulikki Seppänen-Laakso for help with FFA measurements, Nicolas Leuenberger and Walter Wahli for discussions and communication of unpublished data, members of the Schibler laboratory for comments on this manuscript, and Nicolas Roggli for help with artwork. This research was supported by the Swiss National Science Foundation (through an individual research grant to U.S., and the National Center of Competence in Research Program Frontiers in Genetics, and grant SNSF PDFM33-118375 to E.Z.), the State of Geneva, the Louis Jeantet Foundation of Medicine, the Bonizzi-Theler-Stiftung, and the 6th European Framework Project EUCLOCK. D. Gatfield received and gratefully acknowledges long-term fellowships from The Federation of European Biochemical Societies (FEBS) and The International Human Frontier Science Program Organization (HFSP).

References

Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP. 2008. The impact of microRNAs on protein output. *Nature* **455:** 64–71.

Beitzinger M, Peters L, Zhu JY, Kremmer E, Meister G. 2007. Identification of human microRNA targets from isolated argonaute protein complexes. *RNA Biol* **4:** 76–84.

Bhattacharyya SN, Habermacher R, Martine U, Closs EI, Filipowicz W. 2006. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell* **125:** 1111–1124.

Bushati N, Cohen SM. 2007. microRNA functions. *Annu Rev Cell Dev Biol* **23:** 175–205.

Chang J, Nicolas E, Marks D, Sander C, Lerro A, Buendia MA, Xu C, Mason WS, Moloshok T, Bort R, et al. 2004. miR-122, a mammalian liver-specific microRNA, is processed from hcr mRNA and may downregulate the high affinity cationic amino acid transporter CAT-1. *RNA Biol* **1:** 106–113.

Cheng HY, Papp JW, Varlamova O, Dziema H, Russell B, Curfman JP, Nakazawa T, Shimizu K, Okamura H, Impey S, et al. 2007. microRNA modulation of circadian-clock period and entrainment. *Neuron* **54:** 813–829.

Cohen SM, Brennecke J, Stark A. 2006. Denoising feedback loops by thresholding—A new role for microRNAs. *Genes & Dev* **20:** 2769–2772.

Cusick JK, Xu LG, Bin LH, Han KJ, Shu HB. 2006. Identification of RELT homologues that associate with RELT and are phosphorylated by OSR1. *Biochem Biophys Res Commun* **340:** 535–543.

Elmen J, Lindow M, Schutz S, Lawrence M, Petri A, Obad S, Lindholm M, Hedtjarn M, Hansen HF, Berger U, et al. 2008a. LNA-mediated microRNA silencing in non-human primates. *Nature* **452:** 896–899.

Elmen J, Lindow M, Silahtaroglu A, Bak M, Christensen M, Lind-Thomsen A, Hedtjarn M, Hansen JB, Hansen HF, Straarup EM, et al. 2008b. Antagonism of microRNA-122 in mice by systemically administered LNA-antimiR leads to up-regulation of a large set of predicted target mRNAs in the liver. *Nucleic Acids Res* **36:** 1153–1162.

Esau C, Davis S, Murray SF, Yu XX, Pandey SK, Pear M, Watts L, Booten SL, Graham M, McKay R, et al. 2006. miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. *Cell Metab* **3:** 87–98.

Eulalio A, Huntzinger E, Nishihara T, Rehwinkel J, Fauser M, Izaurralde E. 2009. Deadenylation is a widespread effect of miRNA regulation. *RNA* **15:** 21–32.

Fyffe SA, Alphey MS, Buetow L, Smith TK, Ferguson MA, Sorensen MD, Bjorkling F, Hunter WN. 2006. Recombinant human PPAR-β/δ ligand-binding domain is locked in an activated conformation by endogenous fatty acids. *J Mol Biol* **356:** 1005–1013.

Gachon F, Nagoshi E, Brown SA, Ripperger J, Schibler U. 2004. The mammalian circadian timing system: From gene expression to physiology. *Chromosoma* **113:** 103–112.

Gallego M, Virshup DM. 2007. Post-translational modifications regulate the ticking of the circadian clock. *Nat Rev Mol Cell Biol* **8:** 139–148.

Gerlach D, Kriventseva EV, Rahman N, Vejnar CE, Zdobnov EM. 2009. miROrtho: Computational survey of microRNA genes. *Nucleic Acids Res* **37:** D111–D117. doi: 10.1093/nar/gkn707.

Green CB, Douris N, Kojima S, Strayer CA, Fogerty J, Lourim D, Keller SR, Besharse JC. 2007. Loss of Nocturnin, a circadian deadenylase, confers resistance to hepatic steatosis and diet-induced obesity. *Proc Natl Acad Sci* **104:** 9888–9893.

Hofacker IL. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res* **31:** 3429–3431.

Hubbell E, Liu WM, Mei R. 2002. Robust estimators for expression analysis. *Bioinformatics* **18:** 1585–1592.

Karginov FV, Conaco C, Xuan Z, Schmidt BH, Parker JS, Mandel G, Hannon GJ. 2007. A biochemical approach to identifying microRNA targets. *Proc Natl Acad Sci* **104:** 19291–19296.

Katoh T, Sakaguchi Y, Miyauchi K, Suzuki T, Kashiwabara S, Baba T, Suzuki T. 2009. Selective stabilization of mammalian microRNAs by 3′ adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. *Genes & Dev* **23:** 433–438.

Kornmann B, Schaad O, Bujard H, Takahashi JS, Schibler U. 2007a. System-driven and oscillator-dependent circadian transcription in mice with a conditionally active liver clock. *PLoS Biol* **5:** e34. doi: 10.1371/journal.pbio.0050034.

Kornmann B, Schaad O, Reinke H, Saini C, Schibler U. 2007b. Regulation of circadian gene expression in liver by systemic signals and hepatocyte oscillators. *Cold Spring Harb Symp Quant Biol* **72:** 319–330.

Krutzfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, Stoffel M. 2005. Silencing of microRNAs in vivo with 'antagomirs.' *Nature* **438:** 685–689.

Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T. 2002. Identification of tissue-specific microRNAs from mouse. *Curr Biol* **12:** 735–739.

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, et al. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425:** 415–419.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.

Li Y, Wang F, Lee JA, Gao FB. 2006. MicroRNA-9a ensures the precise specification of sensory organ precursors in *Drosophila*. *Genes & Dev* **20:** 2793–2805.

Li S, Liu C, Li N, Hao T, Han T, Hill DE, Vidal M, Lin JD. 2008. Genome-wide coactivation analysis of PGC-1α identifies BAF60a as a regulator of hepatic lipid metabolism. *Cell Metab* **8:** 105–117.

Li X, Cassidy JJ, Reinke CA, Fischboeck S, Carthew RW. 2009. A microRNA imparts robustness against environmental fluctuation during development. *Cell* **137:** 273–282.

Liu J. 2008. Control of protein synthesis and mRNA degradation by microRNAs. *Curr Opin Cell Biol* **20:** 214–221.

Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, et al. 2002. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* **18:** 1593–1599.

Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U. 2004. Nuclear export of microRNA precursors. *Science* **303:** 95–98.

Miller BH, McDearmon EL, Panda S, Hayes KR, Zhang J, Andrews JL, Antoch MP, Walker JR, Esser KA, Hogenesch JB, et al. 2007. Circadian and CLOCK-controlled regulation of the mouse transcriptome and cell proliferation. *Proc Natl Acad Sci* **104:** 3342–3347.

Narkar VA, Downes M, Yu RT, Embler E, Wang YX, Banayo E, Mihaylova MM, Nelson MC, Zou Y, Juguilon H, et al. 2008. AMPK and PPARδ agonists are exercise mimetics. *Cell* **134:** 405–415.

Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, Schultz PG, Kay SA, Takahashi JS, Hogenesch JB. 2002. Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell* **109:** 307–320.

Pillai RS, Bhattacharyya SN, Artus CG, Zoller T, Cougot N, Basyuk E, Bertrand E, Filipowicz W. 2005. Inhibition of translational initiation by Let-7 MicroRNA in human cells. *Science* **309:** 1573–1576.

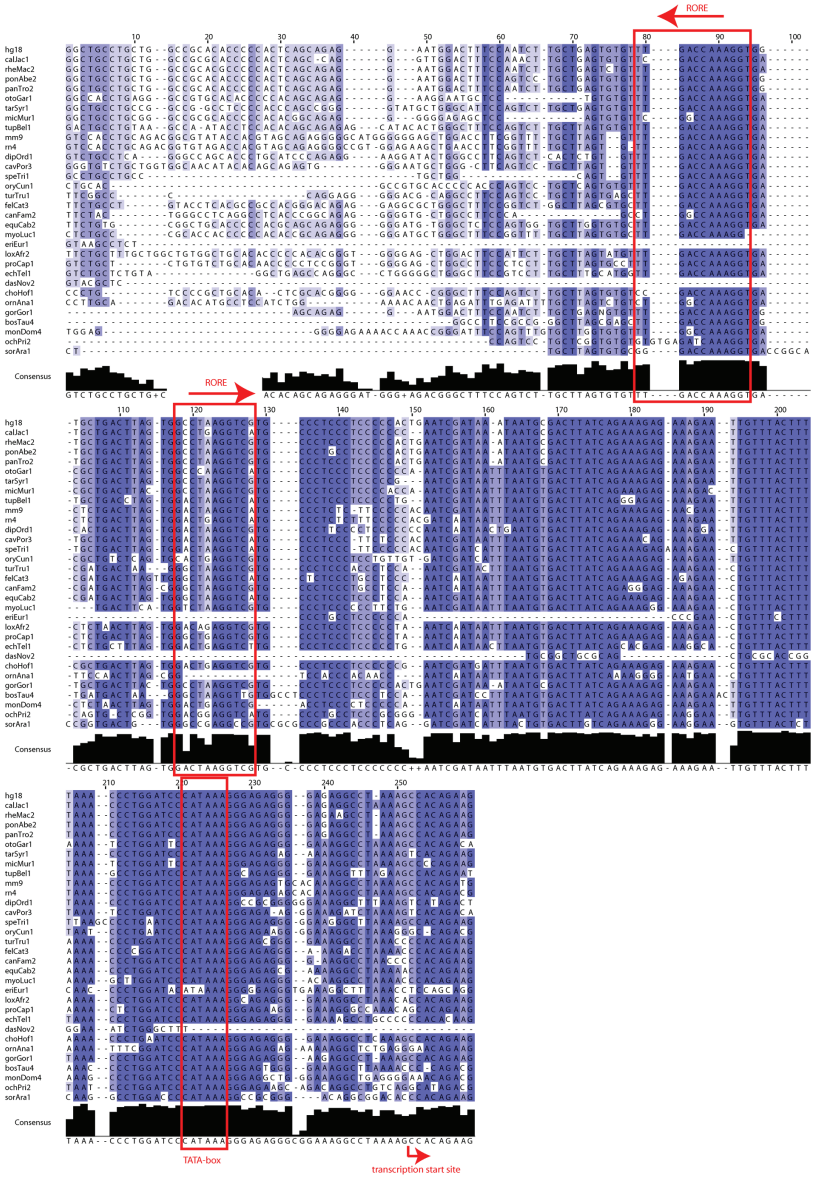
Preitner N, Damiola F, Lopez-Molina L, Zakany J, Duboule D, Albrecht U, Schibler U. 2002. The orphan nuclear receptor REV-ERBα controls circadian transcription within the positive limb of the mammalian circadian oscillator. *Cell* **110:** 251–260.

Reddy AB, Karp NA, Maywood ES, Sage EA, Deery M, O'Neill JS, Wong GK, Chesham J, Odell M, Lilley KS, et al. 2006.

A.1.3.3 *Supplementary information*

An extract of the supplementary material for the paper "Integration of microRNA miR-122 in hepatic circadian gene expression" covering my work is presented on pages 173–174. The original source and can be found here:
`http://genesdev.cshlp.org/content/suppl/2009/05/20/23.11.1313.DC1/GatfieldSuppMat.pdf`

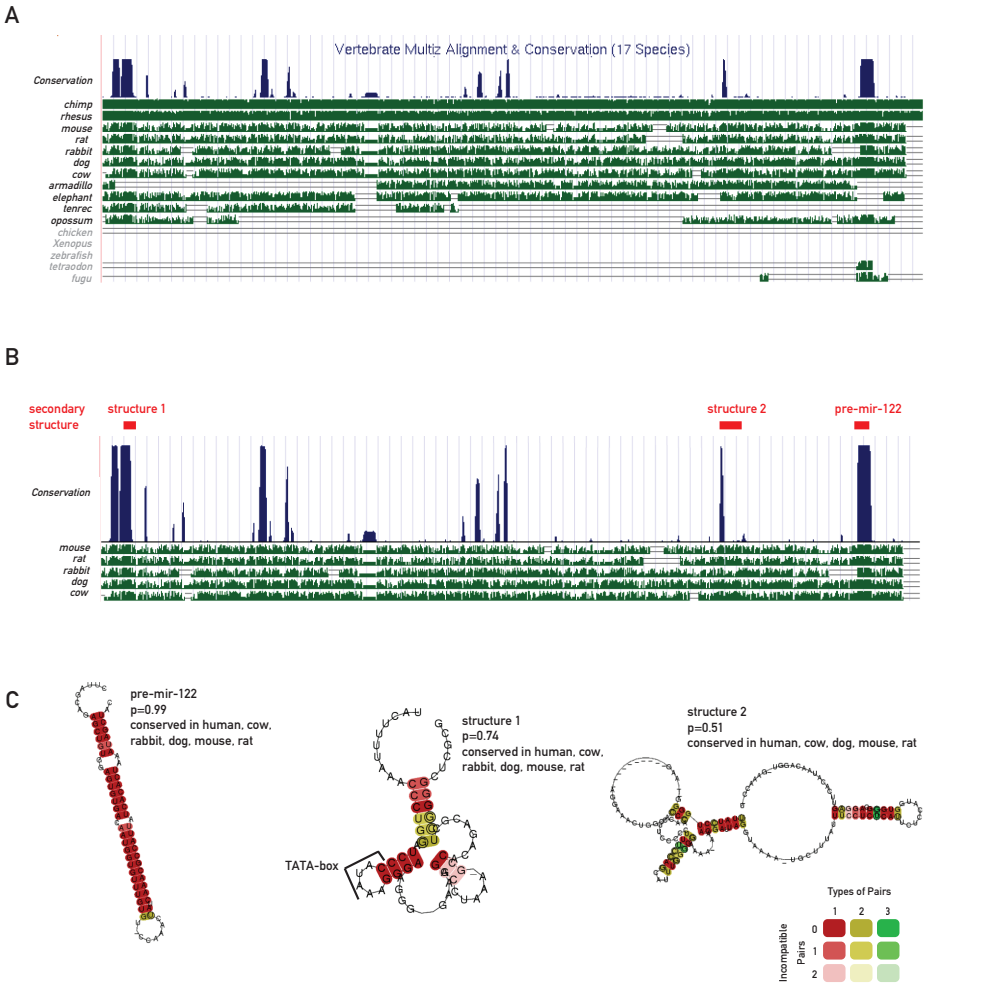
Gatfield_Suppl. Fig2



Supplemental Figure 2. Conservation of the pri-mir-122 promoter.

cons44way UCSC alignment of the mir-122 promotor region of the genomic sequence upstream of the predicted transcriptional start site of pri-mir-122 in 32 mammalian species (species abbreviation: http://hgdownload.cse.ucsc.edu/downloads.html). The predicted ROREs, the TATA-box and transcriptional start site are indicated in red. Both the first and, to a slightly lesser extent, also the second RORE, as well as the overall promoter structure show a high degree of conservation among mammals, down to Platypus. miR-122, however, is conserved even further in evolution down to fish (see Gerlach et al. 2009: http://cegg.unige.ch/mirortho/results?searchterm=mir-122&submit=Search).

Gatfield_Suppl. Fig3



Supplemental Figure 3. Evolutionary conservation of the mir-122 locus and conserved secondary structures.

(A) multiz17way UCSC alignment of the miR-122 locus (http://genome.ucsc.edu; genomic coordinates: hg18, chr18:54,264,258-54,269,700). Note the two high conservation peaks at the 5' and 3' ends representing the promotor region and pre-mir-122, respectively. **(B)** The conserved regions in human, mouse, rat, dog, cow, and rabbit were analyzed using RNAz (Washietl et al., 2005). The software scanned the input alignment using a sliding window of 100 nt in steps of 10 nt. The alignment was color-coded according to sequence conservation and visualized using Jalview (Clamp et al., 2004). **(C)** Three conserved statistically significant structural elements ("RNA class probability" P>0.5) were found within the pri-mir-122 locus. Of these, only pre-mir-122 showed the typical "saddle"-like conservation profile known for conserved miRNA genes and was conserved to a statistically significant extent ("RNA class probability" P=0.99). The two other potential structures were either located immediately upstream of the transcriptional start site (structure 1; P=0.74), or were just above the threshold for significance (structure 2; P=0.51). Moreover, they did not resemble any known functional structures. The three hits found by RNAz were color-coded according to consistent and compensatory mutations using RNAalifold (Hofacker et al., 2002).

A.1.4 *Functional and evolutionary insights from the genomes of three parasitoid Nasonia species*

Nasonia Genome Working Group (MicroRNAs and tRNAs analysis: Anzola JM, Behura SK, Elsik CG, Gerlach D, Hagen DE, Munoz-Torres MC, and Zdobnov EM). Functional and Evolutionary Insights from the Genomes of Three Parasitoid Nasonia Species. *Science* (2010) 327:343–348.

A.1.4.1 *Contributions*

The Nasonia Genome Working Group presents the complete genomic sequences of three parasitoid Nasonia species.

I contributed to the work with the a set of miRNA gene predictions produced via the `miROrtho` pipeline. My predictions can be found here: `http://genomes.arc.georgetown.edu/nasonia/nasonia_genome_consortium/data/Table_S5_comp_miRNAs.xls`. A composed set of miRNA predictions merging my efforts with those of other groups can be found here: `http://genomes.arc.georgetown.edu/nasonia/nasonia_genome_consortium/data/Table_S6_all_miRNAs.pdf`.

A.1.4.2 *Main paper*

See pages 176–181 or at:
`http://www.sciencemag.org/cgi/content/full/327/5963/343`

but enhanced by GTP-γS and AlF_4^- (Fig. 2, C and D), indicating that the cytoplasmic domains of β_3 and β_1 can directly interact with $G\alpha_{13}$ and that GTP enhances the interaction. The $G\alpha_{13}$-β_3 interaction was enhanced in platelets adherent to fibrinogen, and by thrombin, which stimulates GTP binding to $G\alpha_{13}$ via GPCR (Fig. 2E). Hence, the interaction is regulated by both integrin occupancy and GPCR signaling.

To map the β_3 binding site in $G\alpha_{13}$, we incubated cell lysates containing Flag-tagged wild type or truncation mutants of $G\alpha_{13}$ (fig. S5) with GST-β_3CD beads. GST-β_3CD associated with wild-type $G\alpha_{13}$ and the $G\alpha_{13}$ 1 to 212 fragment containing α-helical region and switch region I (SRI), but not with the $G\alpha_{13}$ fragment containing residues 1 to 196 lacking SRI (Fig. 2F). Thus, SRI appears to be critical for β_3 binding. To further determine the importance of SRI, $G\alpha_{13}$-β_3 binding was assessed in the presence of a myristoylated synthetic peptide, Myr-LLARRPTKGIHEY (mSRI), corresponding to the SRI sequence of $G\alpha_{13}$ (197 to 209) (*21*, *22*). The mSRI peptide, but not a myristoylated scrambled peptide, inhibited $G\alpha_{13}$ binding to β_3 (Fig. 2G), indicating that mSRI is an effective inhibitor of β_3-$G\alpha_{13}$ interaction. Therefore, we further examined whether mSRI might inhibit integrin signaling. Treatment of platelets with mSRI inhibited integrin-dependent phosphorylation of c-Src Tyr^{416} and accelerated RhoA activation (Fig. 3A). The effect of mSRI is unlikely to result from its inhibitory effect on the binding of RhoGEFs to $G\alpha_{13}$ SRI because $G\alpha_{13}$ binding to RhoGEFs stimulates RhoA activation, which should be inhibited rather than promoted by mSRI (*22*). Thus, these data suggest that β_3-$G\alpha_{13}$ interaction mediates activation of c-Src and inhibition of RhoA. Furthermore, mSRI inhibited integrin-mediated platelet spreading (Fig. 3B), and this inhibitory effect was reversed by C3 toxin (which catalyzes the ADP ribosylation of RhoA) or Y27632, confirming the importance of $G\alpha_{13}$-dependent inhibition of RhoA in platelet spreading. Thrombin promotes platelet spreading, which requires cdc42/Rac pathways (*23*). However, thrombin-promoted platelet spreading was also abolished by mSRI (Fig. 3B), indicating the importance of $G\alpha_{13}$-β_3 interaction. Thus, $G\alpha_{13}$-integrin interaction appears to be a mechanism that mediates integrin signaling to c-Src and RhoA, thus regulating cell spreading.

To further determine whether $G\alpha_{13}$ mediates inhibition of integrin-induced RhoA-dependent contractile signaling, we investigated the effects of mSRI and depletion of $G\alpha_{13}$ on platelet-dependent clot retraction (shrinking and consolidation of a blood clot requires integrin-dependent retraction of platelets from within) (*7*, *8*). Clot retraction was accelerated by mSRI and depletion of $G\alpha_{13}$ (Fig. 4, A and B, and fig. S6), indicating that $G\alpha_{13}$ negatively regulates RhoA-dependent platelet retraction and coordinates cell spreading and retraction. The coordinated cell spreading-retraction process is also important in wound healing, cell migration, and proliferation (*24*).

The function of $G\alpha_{13}$ in mediating the integrin-dependent inhibition of RhoA contrasts with the traditional role of $G\alpha_{13}$, which is to mediate GPCR-induced activation of RhoA. However, GPCR-mediated activation of RhoA is transient, peaking at 1 min after exposure of platelets to thrombin, indicating the presence of a negative regulatory signal (Fig. 4, D and F). Furthermore, thrombin-stimulated activation of RhoA occurs during platelet shape change before substantial ligand binding to integrins (Fig. 4, C, D, and F). In contrast, after thrombin stimulation, β_3 binding to $G\alpha_{13}$ was diminished at 1 min when $G\alpha_{13}$-dependent activation of RhoA occurs, but increased after the occurrence of integrin-dependent platelet aggregation (Fig. 4, E and F). Thrombin-stimulated binding of $G\alpha_{13}$ to $\alpha_{IIb}\beta_3$ and simultaneous RhoA inhibition both require ligand occupancy of $\alpha_{IIb}\beta_3$ and are inhibited by the integrin inhibitor Arg-Gly-Asp-Ser (RGDS) (Fig. 4, D to F). Thus, our study demonstrates not only a function of integrin $\alpha_{IIb}\beta_3$ as a noncanonical $G\alpha_{13}$-coupled receptor but also a new concept of $G\alpha_{13}$-dependent dynamic regulation of RhoA, in which $G\alpha_{13}$ mediates initial GPCR-induced RhoA activation and subsequent integrin-dependent RhoA inhibition (Fig. 4G). These findings are important for our understanding of how cells spread, retract, migrate, and proliferate, which is fundamental to development, cancer, immunity, wound healing, hemostasis, and thrombosis.

References and Notes

1. R. O. Hynes, *Cell* **110**, 673 (2002).
2. M. H. Ginsberg, A. Partridge, S. J. Shattil, *Curr. Opin. Cell Biol.* **17**, 509 (2005).
3. Y. Q. Ma, J. Qin, E. F. Plow, *J. Thromb. Haemost.* **5**, 1345 (2007).
4. S. J. Shattil, *Trends Cell Biol.* **15**, 399 (2005).
5. A. Obergfell *et al.*, *J. Cell Biol.* **157**, 265 (2002).
6. E. G. Arias-Salgado *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13298 (2003).
7. P. Flevaris *et al.*, *J. Cell Biol.* **179**, 553 (2007).
8. P. Flevaris *et al.*, *Blood* **113**, 893 (2009).
9. N. A. Riobo, D. R. Manning, *Trends Pharmacol. Sci.* **26**, 146 (2005).
10. L. F. Brass, D. R. Manning, S. J. Shattil, *Prog. Hemost. Thromb.* **10**, 127 (1991).
11. A. Moers *et al.*, *Nat. Med.* **9**, 1418 (2003).
12. T. Kozasa *et al.*, *Science* **280**, 2109 (1998).
13. M. J. Hart *et al.*, *Science* **280**, 2112 (1998).
14. B. Klages, U. Brandt, M. I. Simon, G. Schultz, S. Offermanns, *J. Cell Biol.* **144**, 745 (1999).
15. V. Senyuk *et al.*, *Cancer Res.* **69**, 262 (2009).
16. B. S. Coller, *Blood* **55**, 169 (1980).
17. Z. Li, G. Zhang, R. Feil, J. Han, X. Du, *Blood* **107**, 965 (2006).
18. M. Gu, X. Xi, G. D. Englund, M. C. Berndt, X. Du, *J. Cell Biol.* **147**, 1085 (1999).
19. W. T. Arthur, L. A. Petch, K. Burridge, *Curr. Biol.* **10**, 719 (2000).
20. S. Tanabe, B. Kreutz, N. Suzuki, T. Kozasa, *Methods Enzymol.* **390**, 285 (2004).
21. Single-letter abbreviations for amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
22. J. S. Huang, L. Dong, T. Kozasa, G. C. Le Breton, *J. Biol. Chem.* **282**, 10210 (2007).
23. C. Vidal, B. Geny, J. Melle, M. Jandrot-Perrus, M. Fontenay-Roupie, *Blood* **100**, 4462 (2002).
24. K. Moissoglu, M. A. Schwartz, *Biol. Cell* **98**, 547 (2006).
25. X. P. Du *et al.*, *Cell* **65**, 409 (1991).
26. X. D. Ren, M. A. Schwartz, *Methods Enzymol.* **325**, 264 (2000).

27. This work was supported by grants HL080264, HL062350, and HL068819 from the National Heart, Lung, and Blood Institute (X.D.) and GM061454 and GM074001 from the National Institute of General Medical Sciences (T.K.). We thank G. Nucifora for help with bone marrow transplantation and K. O'Brien and M. K. Delaney for proofreading.

Supporting Online Material

www.sciencemag.org/cgi/content/full/327/5963/340/DC1
Materials and Methods
Figs. S1 to S6
References

9 April 2009; accepted 4 December 2009
10.1126/science.1174779

Functional and Evolutionary Insights from the Genomes of Three Parasitoid *Nasonia* Species

The *Nasonia* Genome Working Group*†

We report here genome sequences and comparative analyses of three closely related parasitoid wasps: *Nasonia vitripennis*, *N. giraulti*, and *N. longicornis*. Parasitoids are important regulators of arthropod populations, including major agricultural pests and disease vectors, and *Nasonia* is an emerging genetic model, particularly for evolutionary and developmental genetics. Key findings include the identification of a functional DNA methylation tool kit; hymenopteran-specific genes including diverse venoms; lateral gene transfers among Pox viruses, *Wolbachia*, and *Nasonia*; and the rapid evolution of genes involved in nuclear-mitochondrial interactions that are implicated in speciation. Newly developed genome resources advance *Nasonia* for genetic research, accelerate mapping and cloning of quantitative trait loci, and will ultimately provide tools and knowledge for further increasing the utility of parasitoids as pest insect-control agents.

Parasitoid wasps are insects whose larvae parasitize various life stages of other arthropods (for example, insects, ticks, and mites). Female wasps sting, inject venom, and lay eggs on or in the host, where the developing offspring consume and eventually kill it. Parasitoids

are widely used in the biological control of insect pests, and they are very diverse, with estimates of over 600,000 species (*1*, *2*). *Nasonia* is the second genus of Hymenoptera to have whole-genome sequencing, after *Apis mellifera* (Fig. 1), and *Nasonia* comprises four closely related parasitoid species: *N. vitripennis*, *N. giraulti*, *N. longicornis*, and *N. oneida* (*3*, *4*). *Nasonia* are genetically tractable organisms with short generation time (~2 weeks), large family size, ease of laboratory rearing, and cross-fertile species. Like other hymenopterans, haploid males develop from unfertilized eggs, and diploid females develop from fertilized eggs. Cross-fertile species facilitate the mapping and cloning of genes that are involved in species differences. Haploid genetics assist efficient genotyping, mutational screening (*5*), and evaluation of gene interactions (epistasis) without the added complexity of genetic dominance. As a result, *Nasonia* are now emerging as genetic model organisms, particularly for complex trait analysis, developmental genetics, and evolutionary genetics (*4*).

We sequenced, assembled, annotated, and analyzed the genome of *N. vitripennis* from sixfold Sanger sequence genome coverage by using a highly inbred line of *N. vitripennis* (*6*). The draft genome assembly comprises 26,605 contigs [total length of 239.8 Mb, with half of the bases residing in contigs larger than 18.5 kb (N_{50}), 40.6% guanine plus cytosine content (GC)]. Contigs were placed with mate-pair information into 6,181 scaffolds (total size 295 Mb, N_{50} =709 kb). We assessed the *N. vitripennis* assembly for completeness and accuracy by comparing it with 19 finished bacterial artificial chromosome (BAC) sequences and 18,000 expressed sequence tags (ESTs). The genome assembly contained 98% of the BAC and 97% of the EST sequences, with an error rate of 5.9×10^{-4}. Thus the assembly is a high-quality representation of both genomic and transcribed *N. vitripennis* sequences.

Highly inbred lines of the two sibling species *N. giraulti* and *N. longicornis* (Fig. 1B) were sequenced with onefold Sanger and 12-fold, 45–base pair (bp) Illumina genome coverage. Assembled by alignment to the *N. vitripennis* reference using stringent criteria (*6*), these reads cover 62% and 62.6% of the *N. vitripennis* assembly, and 84.7% and 86.3% of protein coding regions, respectively. These were used for genome comparisons and provided resources [for example, single nucleotide polymorphisms (SNPs) and microsatellites] for scaffold, gene, and quantitative trait loci (QTL) mapping. Sequence error rates for the *N. giraulti* alignment are estimated to be 3.8×10^{-3} for the entire alignment and 1.47×10^{-4} for coding sequences on the basis of comparison to three finished *N. giraulti* BACs (*6*). Sequences of 25 coding genes in both species perfectly matched their respective aligned sequences.

Normally, the intracellular bacteria *Wolbachia* prevent the formation of interspecies hybrids; however, antibiotically cured strains are cross-fertile (*7*). Hybrid crosses (Fig. 1C) (*6*) were used to map scaffolds and visible mutations onto the five chromosomes of *Nasonia* (Fig. 2). Several interspecies QTL have already been mapped using genetic/genomic resources, including wing size (*8*, *9*), host preference (*10*), female mate preference (*11*), and in this study, sex-ratio control and male courtship (*6*). Linkage analysis has revealed that the genome-wide recombination rate in *Nasonia* is 1.4 to 1.5 centimorgans (cM)/Mb, which is lower than that of honeybees (*12*, *13*), and shows a 100-fold difference in rate between high- and low-recombination regions of the genome (Fig. 2) (*6*).

An official gene set (OGS v1.1) was generated from comparisons to *A. mellifera*, *Tribolium castaneum*, *Drosophila melanogaster*, *Pediculus humanus*, *Daphnia pulex*, and *Homo sapiens* [details are given in (*6*)]. Overall, *Nasonia* encodes a typical insect gene repertoire (Fig. 3) (*6*), of which 60% of genes have a human ortholog, 18% are arthropod-specific, and 2.4% appear to be hymenoptera-specific, showing high conservation between *Nasonia* and *Apis* and low conservation or absence in other taxa. An additional 12% are either *Nasonia*-specific or without clear orthology. Many (63%) single-copy orthologs shared between *Nasonia* and *Apis* occur in microsynteny blocks, which is similar to the amount of microsynteny blocks in *Aedes aegypti*/*Anopheles gambiae* and *H. sapiens*/*Gallus gallus* (*14*). Four hundred and forty-five orthologs between *Nasonia* and humans lack a candidate homolog in *D. melanogaster* (table S1), including the human transcription factors E2F7 and E2F8, which are involved in cell-cycle regulation. Further refinement of the gene set resulted in OGS v1.2 (*15*), which totals 17,279 genes, of which 74% have tiling microarray or EST support (*6*).

Nasonia is abundant in transposable elements (TEs) and other repetitive DNA (table S2 and fig. S1). This contrasts with a paucity of TEs in *A. mellifera* (*16*). TE diversity in *Nasonia* is 30% higher (2.9 TE types/Mb) than the next most diverse insect (*Bombyx mori*, 2.1 TE types/Mb), and is 10-fold higher than the average dipteran (*6*, *17*). *Nasonia* also contains an unusual abundance of nuclear-mitochondrial insertions and a higher density of microsatellites (10.9 kb/Mb) than most other arthropod species (*18*, *19*), suggesting that the accumulation of repetitive DNA is a feature of these insects.

The *Nasonia* genome encodes a full DNA methylation tool kit, including all three DNA cytosine-5-methyltransferase (Dnmt) types (Fig. 1A).

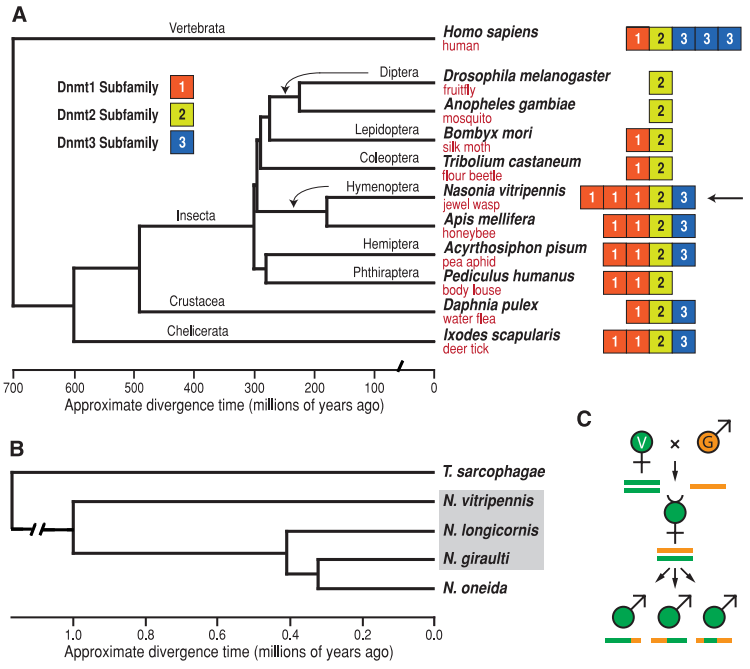


Fig. 1. Phylogenetic relationships of *Nasonia* and the DNA methylation tool kit. (**A**) *Nasonia* relationships to other sequenced genomes (*6*). Right: DNA methyltransferase subfamilies (Dnmt1, Dnmt2, Dnmt3) in these taxa. (**B**) Relationships among the three sequenced *Nasonia* genomes. (**C**) Crossing scheme used for mapping scaffolds on the *Nasonia* chromosomes and for studies of nuclear-cytoplasmic incompatibility.

*All authors with their affiliations appear at the end of this paper.
†To whom correspondence should be addressed. E-mail: werr@mail.rochester.edu (J.H.W.); stephenr@bcm.tmc.edu (S.R.)

In vertebrates, Dnmt3 establishes DNA methylation patterns, Dnmt1 maintains these patterns, and Dnmt2 is involved in tRNA methylation (*20*). The *Nasonia* genome encodes three Dnmt1 genes, one Dnmt2, and one Dnmt3, in contrast with *D. melanogaster*, which has only Dnmt2. The presence of all three subfamilies in both *Nasonia* and *Apis* (Fig. 1) raises the question of whether methylation has similar regulatory functions in Hymenoptera as it does in vertebrates. DNA methylation is important in *Apis* caste development (*21*) and is suggested for *Nasonia* sex determination (*22*). Coding exons of both *Nasonia* and *Apis* show bimodal distributions in observed/expected CpG (fig. S2) (*6*, *23*), which is consistent with mutational biases due to DNA methylation of hyper- and hypomethylated genes. We confirmed methylated CpG dinucleotides in five examined *N. vitripennis* genes by bisulfite sequencing (fig. S3). These results suggest that epigenetic modifications by DNA methylation may be important in Hymenoptera. *Nasonia* also has the largest number of ankyrin (ANK) repeat–containing proteins (over 200) so far found in any insect (table S3) (*6*), suggesting a regulatory importance through protein-protein interactions (*24*).

Systemic RNA interference (RNAi) in *Nasonia* allows for gene expression knockdowns (*4*, *25*). The *Nasonia* genome encodes homologs for the majority of genes implicated in small RNA processes (table S4). However, as in *Tribolium* and *Apis*, *Nasonia* lacks an RNA-dependent RNA polymerase (RdRp) ortholog, indicating a different systemic RNAi mechanism than in *Caenorhabditis*. Using various computational approaches (*6*), we identified 52 putative micro RNAs (miRNAs) with homologies to known miRNAs (*26*), nine that were previously unknown, and 11 additional Hymenoptera-specific miRNAs (table S5). Small-RNA library sequencing confirmed 39 predicted and identified 59 additional miRNAs (table S6).

Nasonia shares a long germ-band mode of embryonic development with *Drosophila*, but exhibits significant differences in the genetic mechanisms involved (*5*, *27*, *28*) (see fig. S4). All major components of the dorso-ventral patterning system are present, with many *Nasonia*-specific gene duplications in the Toll pathway. Orthologs of vertebrate genes absent from *Drosophila* include the transforming growth factor–β (TGFβ) ligands *ADMP* and *myostatin*, and the bone morphogenesis protein (BMP) inhibitors *BAMBI* and *DAN*, but their functions in *Nasonia* are not yet known. *A. mellifera* shows an expansion of the *yellow/major royal jelly* (*yellow/MRJP*) genes that are linked to caste formation and sociality (*29*). *Nasonia* has the largest number of *yellow/MRJP* genes so far found in any insect, including an independent amplification of MRJP-like proteins (fig. S5) (*6*, *29*). Although their function in *Nasonia* is unknown, these genes are expressed broadly in different tissues and life stages (table S7). The insect sex peptide/receptor system, which causes female re-mating refractoriness (*30*), is highly conserved in insects but is absent in *Nasonia* and *Apis* (table S8) (*6*). Instead, *Nasonia* males inhibit female re-mating behaviorally with a special "post-copulatory display" (*31*). Additional features analyzed (*6*) include those related to sex determination (fig. S6), pathogens and immunity (fig. S7), neuropeptides (tables S9 and S10), cuticular proteins (table S11), xenobiotics (fig. S8), and diapause (table S12).

We investigated genome microevolution, including rapidly evolving genes that are potentially involved in species differences and speciation, by using the genomes of the three closely related *Nasonia* species. Synonymous divergence between *N. vitripennis* and its sibling species *N. giraulti* and *N. longicornis* is 0.031 ± 0.0002 SE and 0.030 ± 0.0002 SE, respectively, and between *N. giraulti* and *N. longicornis* is 0.014 ± 0.0001 SE (*6*), which is comparable to those among *Drosophila* sibling species (*32*). We compared the ratio of synonymous-to-nonsynonymous substitutions (*dN*/*dS*) between *Nasonia* species pairs with respect to gene ontology (GO) term categories, using genes with high-quality alignments and 1:1 orthologs between *Nasonia* and *Drosophila*. Nuclear genes that interact with mitochondria revealed significantly elevated *dN*/*dS* [by comparison of *dN*/*dS* distributions for each GO term to resampled distributions, see (*6*) and table S13], specifically those encoding mitochondrial ribosomes ($P < 0.003$ for all species pairs) and oxidative phosphorylation complex I ($P < 0.03$ for *N. vitripennis*/*N. giraulti* and *N. vitripennis*/*N. longicornis*) and complex V ($P < 0.04$ for all species pairs). This finding is consistent with the rapid evolutionary rate of *Nasonia* mitochondria (*33*) and studies implicating nuclear-mitochondrial incompatibilities in F2 hybrid breakdown (*7*, *31*). For example, reciprocal crosses between *N. giraulti* × *N. vitripennis* have identical F1 nuclear genotypes, but their mitochondrial haplotypes differ. Yet, microarray hybridization (Fig. 2) (*6*) of DNA from pooled surviving adult F2 haploid males shows distortion in the recovery of particular regions of the genome, which is dependent upon their mitochondrial haplotype (*giraulti* versus *vitripennis*). Because hybrid mortality is post-embryonic (*7*) and embryo ratios are Mendelian (*33*), these distortions reflect larval to adult mortality. In particular, F2 males with *N. vitripennis* alleles on the left arm of chro-

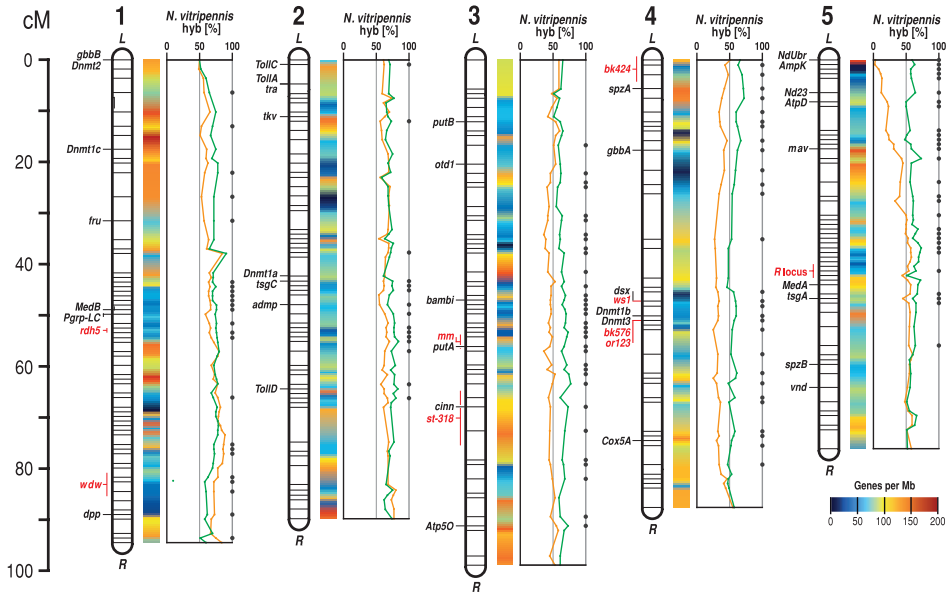


Fig. 2. A high-resolution recombination map of the five *Nasonia* chromosomes is shown (*6*), with estimated gene density and locations of visible markers, landmark genes, and QTL. The hybridization percentage to *N. vitripennis* alleles is shown among surviving adult *N. vitripennis* × *N. giraulti* F2 hybrid males with either *N. vitripennis* (green curve) or *N. giraulti* (orange curve) mitochondria. Dots specify genome regions with significant differences in the hybridization ratio between the reciprocal crosses ($P < 0.01$).

mosome 5 and *N. giraulti* mitochondria suffer nearly 100% mortality (Fig. 2). This region contains three genes encoding mitochondrial interacting proteins, *atpD*, *ampK*, and *nadh-ubiquinone oxireductase* (Fig. 2). Coevolution of nuclear and mitochondrial genomes can accelerate evolution (*34*, *35*), and these findings indicate that such interactions contribute to reproductive incompatibility and speciation in *Nasonia*.

Sequences of 25 gene regions from multiple strains for the three *Nasonia* species (*6*) show low levels of intraspecific variation (table S14) with synonymous site variation ranging from 0.0005 in *N. giraulti* to 0.0026 in *N. vitripennis*, which are much lower than in *Drosophila* species and more akin to levels observed in humans (*36*). This low nuclear variation could be explained by founder events, purging of deleterious mutations in haploid males, or inbreeding.

Recent lateral gene transfers from the bacterial endosymbiont *Wolbachia* into the genomes of *Nasonia* and other arthropods have been identified (*37*). Detecting ancient lateral transfers is more problematic. By examining protein domain arrangements in *Nasonia* relative to other organisms, we uncovered an ancient lateral gene transfer involving Pox viruses, *Wolbachia*, and *Nasonia*. Thirteen ANK repeat–bearing proteins encoded in the *N. vitripennis* genome also contain C-terminal PRANC (Pox proteins repeats of ankyrin–C terminal) domains. This domain was previously only described in Pox viruses, where it is associated with ANK repeats and inhibits the nuclear factor κB (NF-κB) pathway in mammalian hosts (*38*). A computational screen revealed ANK-PRANC–bearing genes in some *Wolbachia* and a related Rickettsiales (Fig. 4). Screening additional *Wolbachia* confirmed the presence of ANK-PRANC genes in diverse *Wolbachia*. The *Nasonia* PRANC genes are clearly integrated in the genome (*6*) and are expressed in different life stages (table S15). Phylogenetic analysis of the PRANC-domain sequences suggests that the *Nasonia* lineage acquired one or more of these proteins from *Wolbachia*, with subsequent amplification and divergence (Fig. 4). Such lateral gene transfers between bacteria and animals could be an important source of evolutionary innovation (*37*).

Nasonia is a carnivore, feeding on an amino acid–rich diet both as larva and adult (*4*). Mapping of *Nasonia* genes onto metabolic pathways (*39*) revealed loss or rearrangement in some amino acid metabolic pathways, including tryptophan and aminosugar metabolism (fig. S9) (*6*). The changes may reflect its specialized carnivorous diet and can inform efforts to produce artificial diets for more economical parasitoid rearing.

The venom of parasitoids, injected into a host before oviposition, serves to condition the host for successful development of wasp progeny (*1*, *2*). Unlike the defensive *Apis* venom that inflicts pain and damage, parasitoid venoms have diverse physiological effects on hosts, including developmental arrest; alteration in growth and physiology; suppression of immune responses; induction of paralysis, oncosis, or apoptosis; and alteration of host behavior (*40*). The identification of *Nasonia*

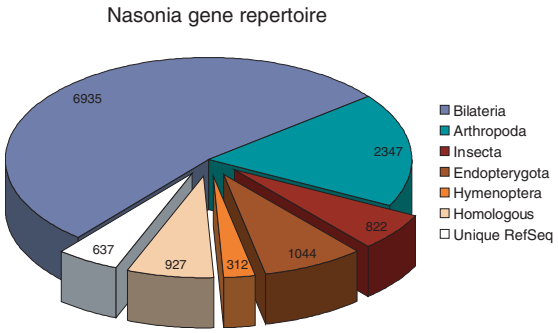


Fig. 3. Distribution of recognizable *Nasonia* orthologs and *Nasonia*-specific genes among gene models with expression sequencing support (*6*).

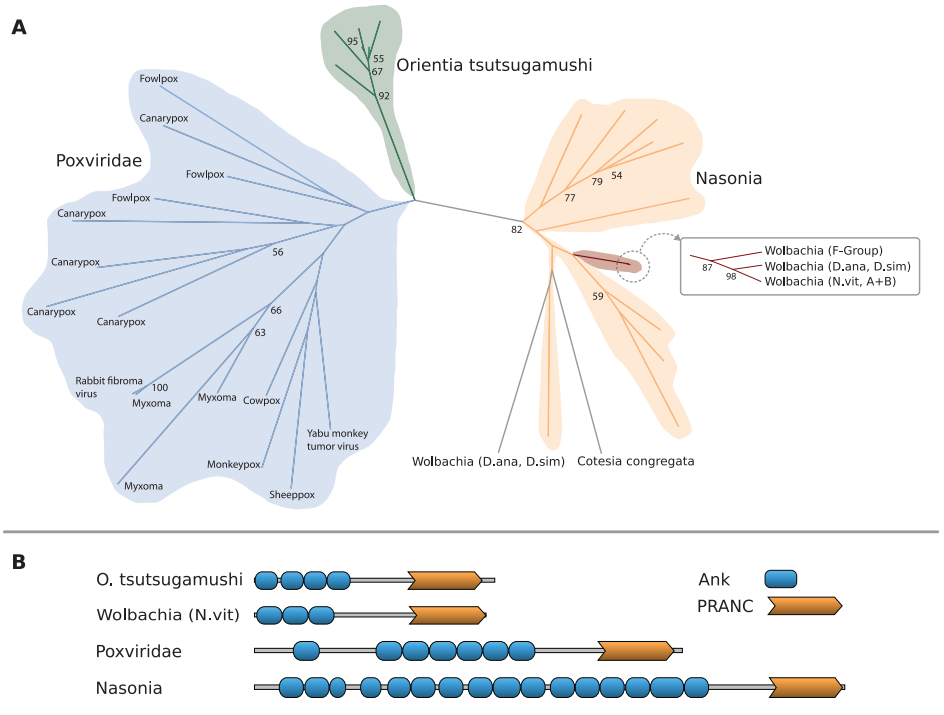


Fig. 4. PRANC domain proteins in *Nasonia*, Pox viruses, and *Wolbachia*. (**A**) Maximum-likelihood tree of PRANC-domain sequences found in Pox viruses, rickettsia (*Wolbachia* and *Orientia*), and parasitoids (*N. vitripennis* and *Cotesia congregata*). The tree was estimated using RaxML with 1000 bootstrap replicates and model settings estimated by ProtTest [see (*6*); alignment deposited in Treebase with ID SN4709]. Bootstrap values above 50% are shown by the corresponding nodes. The phylogenetic relationships suggest lateral transfer from *Wolbachia* to the *Nasonia* lineage. (**B**) Representative domain arrangements for ANK-PRANC proteins.

genes with venom features and proteomic analyses of venom reservoir tissues have uncovered a rich assemblage of 79 candidate venom proteins (table S16) (*41*). Some *Nasonia* venom reservoir proteins belong to previously known insect venom families such as serine proteases; however, nearly half were not related to any known insect venoms. As expected, many of these venom candidates show highly elevated expression in the female reproductive tract, which includes the venom glands and reservoirs. Venom genes also showed significantly higher *dN*/*dS* ratios between *N. vitripennis* and *N. giraulti* than nonvenom genes did (Mann-Whitney U test, $P < 2 \times 10^{-6}$), suggesting that changes in host use between the species may be accompanied by rapid evolution of venom proteins. The large venom protein set found in *Nasonia* with diverse physiological effects (*40*) and abundance of parasitoid species (*1*, *2*) suggests that parasitoids may contain a rich venom pharmacopeia of potential new drugs.

N. vitripennis is a generalist parasitoid with a wide host utilization of many fly species, whereas the other *Nasonia* species are specialists (*4*, *10*). Using genomic tools, a major host preference locus has been mapped to a region of ~2 cM (*10*). Other genes in the *Nasonia* genome that are potentially involved in host finding include odorant binding proteins (table S17) and chemoreceptors (*42*), which show expansions, contractions, and pseudogenization, indicative of rapid turnover.

A suite of genetic tools and resources is available or under development for the *Nasonia* system (*4*, *6*, *11*, *28*), and the genome resources presented here can be used for fine-scale mapping (*6*, *9*-*11*) and positional cloning (*8*) of QTLs. By combining haploid genetics, ease of rearing, short generation time, systemic RNAi, interfertile species, and new genome resources for three species, *Nasonia* shows promise as a genetic model system for evolutionary and developmental genetics. Genome resources described here and our resulting enhanced understanding of parasitoid biology will also open avenues for improving parasitoid utility in biological control of pests of agricultural and medical importance.

References and Notes

1. D. L. J. Quicke, *Parasitic Wasps* (Chapman & Hall, London, 1997).
2. J. Heraty, in *Insect Biodiversity: Science and Society*, R. Tootti and P. Alder, Eds. (Wiley-Blackwell, Hoboken, NJ, 2009), pp. 445–462.
3. R. Raychoudhury *et al.*, *Heredity* 10.1038/hdy.2009.147 (2010).
4. J. H. Werren, D. Loehlin, *Cold Spring Harb. Protocols* 10.1101/pdb.emo134 (2009).
5. M. A. Pultz *et al.*, *Genetics* **154**, 1213 (2000).
6. Materials and methods and supplementary text are available as supporting material on *Science* Online.
7. J. A. J. Breeuwer, J. H. Werren, *Evolution* **49**, 705 (1995).
8. D. W. Loehlin *et al.*, *PLoS Genet.* 10.1371/journal.pgen.1000821 (2010).
9. D. W. Loehlin, L. S. Enders, J. H. Werren, *Heredity* 10.1038/hdy.2009.146 (2010).
10. C. A. Desjardins, F. Perfecti, J. D. Bartos, L. S. Enders, J. H. Werren, *Heredity* 10.1038/hdy.2009.145 (2010).
11. B. J. Velthuis, W. Yang, T. van Opijnen, J. H. Werren, *Anim. Behav.* **69**, 1107 (2005).
12. O. Niehuis *et al.*, *PLoS ONE* 10.1371/journal.pone.0008597 (2010).
13. L. Wilfert, J. Gadau, P. Schmid-Hempel, *Heredity* **98**, 189 (2007).
14. E. M. Zdobnov, P. Bork, *Trends Genet.* **23**, 16 (2007).
15. The official gene set OGS v1.2 is available at http://nasoniabase.org/nasonia_genome_consortium/datasets.html.
16. Honeybee Genome Sequencing Consortium, *Nature* **443**, 931 (2006).
17. C. D. Smith *et al.*, *Gene* **389**, 1 (2007).
18. L. Viljakainen, D. C. S. G. Oliveira, J. H. Werren, S. K. Behura, *Insect Mol. Biol.* **19**, 27 (2010).
19. B. A. Pannebakker, O. Niehuis, A. Hedley, J. Gadau, D. M. Shuker, *Insect Mol. Biol.* **19**, 91 (2010).
20. T. P. Jurkowski *et al.*, *RNA* **14**, 1663 (2008).
21. R. Kucharski, J. Maleszka, S. Foret, R. Maleszka, *Science* **319**, 1827 (2008).
22. L. W. Beukeboom, A. Kamping, L. van de Zande, *Semin. Cell Dev. Biol.* **18**, 371 (2007).
23. N. Elango, B. G. Hunt, M. A. D. Goodisman, S. V. Yi, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11206 (2009).
24. L. K. Mosavi, T. J. Cammett, D. C. Desrosiers, Z. Y. Peng, *Protein Sci.* **13**, 1435 (2004).
25. J. A. Lynch, C. Desplan, *Nat. Protoc.* **1**, 486 (2006).
26. D. Gerlach, E. V. Kriventseva, N. Rahman, C. E. Vejnar, E. M. Zdobnov, *Nucleic Acids Res.* **37**, D111 (2009).
27. J. A. Lynch, A. E. Brent, D. S. Leaf, M. A. Pultz, C. Desplan, *Nature* **439**, 728 (2006).
28. A. E. Brent, G. Yucel, S. Small, C. Desplan, *Science* **315**, 1841 (2007).
29. M. D. Drapeau, S. Albert, R. Kucharski, C. Prusko, R. Maleszka, *Genome Res.* **16**, 1385 (2006).
30. N. Yapici, Y. J. Kim, C. Ribeiro, B. J. Dickson, *Nature* **451**, 33 (2008).
31. O. Niehuis, A. K. Judson, J. Gadau, *Genetics* **178**, 413 (2008).
32. J. M. Flowers *et al.*, *Mol. Biol. Evol.* **24**, 1347 (2007).
33. J. van den Assem, J. Visser, *Biol. Comport.* **1**, 37 (1976).
34. D. C. S. G. Oliveira, R. Raychoudhury, D. V. Lavrov, J. H. Werren, *Mol. Biol. Evol.* **25**, 2167 (2008).
35. D. M. Rand, R. A. Haney, A. J. Fry, *Trends Ecol. Evol.* **19**, 645 (2004).
36. C. F. Aquadro, V. B. Dumont, F. A. Reed, *Curr. Opin. Genet. Dev.* **11**, 627 (2001).
37. J. C. Dunning-Hotopp *et al.*, *Science* **317**, 1753 (2007).
38. S. J. Chang *et al.*, *J. Virol.* **83**, 4140 (2009).
39. M. Kanehisa *et al.*, *Nucleic Acids Res.* **36**, D480 (2008).
40. D. B. Rivers, Y. A. Yoder, L. Ruggiero, *Trends Entomol.* **2**, 1 (1999).
41. D. C. de Graaf *et al.*, *Insect Mol. Biol.* **19**, 11 (2010).
42. H. M. Robertson, J. Gadau, K. W. Wanner, *Insect Mol. Biol.* **19**, 121 (2010).

43. Genome sequencing, assembly and annotation were funded by the National Human Genome Research Institute (NHGRI U54 HG003273). The whole-genome shotgun project has been deposited at the DNA Databank of Japan (DDBJ)/European Molecular Biology Laboratory (EMBL)/GenBank under accession numbers AAZX00000000 (*N. vitripennis*), ADAO00000000 (*N. giraulti*), and ADAP00000000 (*N. longicornis*). Additional support, acknowledgments, and accession numbers are provided in the supporting online material.

Author List: The *Nasonia* Genome Working Group

John H. Werren,[1]*† Stephen Richards,[2]*† Christopher A. Desjardins,[1] Oliver Niehuis,[3]‡ Jürgen Gadau,[3] John K. Colbourne,[4] Leo W. Beukeboom,[5] Claude Desplan,[6] Christine G. Elsik,[7] Cornelis J. P. Grimmelikhuijzen,[8] Paul Kitts,[9] Jeremy A. Lynch,[10] Terence Murphy,[9] Deodoro C. S. G. Oliveira,[1]§ Christopher D. Smith,[11,12] Louis van de Zande,[5] Kim C. Worley,[2] Evgeny M. Zdobnov,[13,14,15] Maarten Aerts,[16] Stefan Albert,[17] Victor H. Anaya,[18] Juan M. Anzola,[19] Angel R. Barchuk,[20] Susanta K. Behura,[21] Agata N. Bera,[22] May R. Berenbaum,[23] Rinaldo C. Bertossa,[24] Márcia M. G. Bitondi,[25] Seth R. Bordenstein,[26,27] Peer Bork,[28] Erich Bornberg-Bauer,[29] Marleen Brunain,[30] Giuseppe Cazzamali,[8] Lesley Chaboub,[2] Joseph Chacko,[2] Dean Chavez,[2] Christopher P. Childers,[7] Jeong-Hyeon Choi,[4] Michael E. Clark,[1] Charles Claudianos,[31] Rochelle A. Clinton,[32] Andrew G. Cree,[2] Alexandre S. Cristino,[31,33] Phat M. Dang,[34] Alistair C. Darby,[35] Dirk C. de Graaf,[30] Bart Devreese,[16] Huyen H. Dinh,[2] Rachel Edwards,[1] Navin Elango,[36] Eran Elhaik,[37] Olga Ermolaeva,[9] Jay D. Evans,[38] Sylvain Foret,[39] Gerald R. Fowler,[2] Daniel Gerlach,[13,14] Joshua D. Gibson,[3] Donald G. Gilbert,[40] Dan Graur,[37] Stefan Gründer,[41] Darren E. Hagen,[7] Yi Han,[2] Frank Hauser,[8] Da Hultmark,[42] Henry C. Hunter IV,[11] Gregory D. D. Hurst,[35] Shalini N. Jhangian,[2] Huaiyang Jiang,[2] Reed M. Johnson,[43] Andrew K. Jones,[22] Thomas Junier,[13] Tatsuhiko Kadowaki,[44] Albert Kamping,[5] Yuri Kapustin,[9] Bobak Kechavarzi,[45] Jaebum Kim,[46] Jay Kim,[11] Boris Kiryutin,[9] Tosca Koevoets,[5] Christie L. Kovar,[2] Evgenia V. Kriventseva,[47] Robert Kucharski,[48] Heewook Lee,[45] Sandra L. Lee,[2] Kristin Lees,[22] Lora R. Lewis,[2] David W. Loehlin,[1] John M. Logsdon Jr.,[49] Jacqueline A. Lopez,[4] Ryan J. Lozado,[2] Donna Maglott,[9] Ryszard Maleszka,[48] Anoop Mayampurath,[45] Danielle J. Mazur,[49] Marcella A. McClure,[32] Andrew D. Moore,[29] Margaret B. Morgan,[2] Jean Muller,[28] Monica C. Munoz-Torres,[7,50] Donna M. Muzny,[2] Lynne V. Nazareth,[2] Susanne Neupert,[51] Ngoc B. Nguyen,[2] Francis M. F. Nunes,[25,52] John G. Oakeshott,[53] Geoffrey O. Okwuonu,[2] Bart A. Pannebakker,[5,54] Vikas R. Pejaver,[45] Zuogang Peng,[36] Stephen C. Pratt,[3] Reinhard Predel,[51] Ling-Ling Pu,[2] Hilary Ranson,[55] Rhitoban Raychoudhury,[1] Andreas Rechtsteiner,[4,56] Justin T. Reese,[7,57] Jeffrey G. Reid,[2] Megan Riddle,[58]‖ Hugh M. Robertson,[23] Jeanne Romero-Severson,[59] Miriam Rosenberg,[6] Timothy B. Sackton,[60] David B. Sattelle,[22] Helge Schlüns,[61] Thomas Schmitt,[62] Martina Schneider,[8] Andreas Schüler,[29] Andrew M. Schurko,[49] David M. Shuker,[63] Zilá L. P. Simões,[25] Saurabh Sinha,[46] Zachary Smith,[4] Victor Solovyev,[64] Alexandre Souvorov,[9] Andreas Springauf,[41] Elisabeth Stafflinger,[8] Deborah E. Stage,[1] Mario Stanke,[65] Yoshiaki Tanaka,[66] Arndt Telschow,[29] Carol Trent,[58] Selina Vattathil,[2]¶ Eveline C. Verhulst,[5] Lumi Viljakainen,[67] Kevin W. Wanner,[68] Robert M. Waterhouse,[15] James B. Whitfield,[23] Timothy E. Wilkes,[35] Michael Williamson,[8] Judith H. Willis,[69] Florian Wolschin,[70,3] Stefan Wyder,[13] Takuji Yamada,[28] Soojin V. Yi,[36] Courtney N. Zecher,[27] Lan Zhang,[2] Richard A. Gibbs[2]

[1]Department of Biology, University of Rochester, Rochester, NY 14627, USA. [2]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA. [3]School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA. [4]The Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405, USA. [5]Evolutionary Genetics–Centre for Ecological and Evolutionary Studies, University of Groningen, 9750 AA Haren, Netherlands. [6]Department of Biology, New York University, New York, NY 10003, USA. [7]Department of Biology, Georgetown University, Washington, DC 20057, USA. [8]Center for Comparative and Functional Insect Genomics, Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark. [9]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. [10]Institut für Entwicklungsbiologie, Universität zu Köln, 50923 Köln, Germany. [11]Department of Biology, San Francisco State University, San Francisco, CA 94132, USA. [12]Drosophila Heterochromatin Genome Project, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [13]Department of Genetic Medicine and Development, University of Geneva Medical School, CH-1211 Geneva, Switzerland. [14]Swiss Institute of Bioinformatics, CH-1211 Geneva, Switzerland. [15]Imperial College London, London, SW7 2AZ, UK. [16]Laboratory of Protein Biochemistry and Biomolecular Engineering, Ghent University, B-9000 Ghent, Belgium. [17]BEEgroup and Institute of Pharmaceutical Biology, University of Würzburg, 97082 Würzburg, Germany. [18]Institute for Theoretical Biology, Humboldt University Berlin, 10115 Berlin, Germany. [19]Animal Science and Biology, Texas A&M University, College Station, TX 77843, USA. [20]Departamento de Ciências Biomédicas, Universidade Federal de Alfenas, Alfenas, Minas Gerais, 37130-000, Brazil. [21]Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA. [22]Medical Research Council Functional Genomics Unit, Department of Physiology Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK. [23]Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [24]Chronobiology–Centre for Behavior and Neurosciences, University of Groningen, 9750 AA Haren, Netherlands. [25]Faculdade de Filosofia,

Ciências e Letras de Ribeirão Preto, Departamento de Biologia, Universidade de São Paulo, Ribeirão Preto, São Paolo 14040-901, Brazil. [26]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA. [27]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02536, USA. [28]European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [29]Institute for Evolution and Biodiversity, University of Münster, 48143 Münster, Germany. [30]Laboratory of Zoophysiology, Ghent University, B-9000 Ghent, Belgium. [31]The Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia. [32]Department of Microbiology and the Center for Computational Biology, Montana State University, Bozeman, MT 59715, USA. [33]Instituto de Física de São Carlos, Departamento de Física e Informática, Universidade de São Paulo, São Carlos, São Paolo 13560-970, Brazil. [34]Subtropical Insects Research Unit, United States Department of Agriculture–Agricultural Research Service (USDA-ARS), U.S. Horticultural Research Lab, Fort Pierce, FL 34945, USA. [35]School of Biological Sciences, University of Liverpool, Liverpool L69 7ZB, UK. [36]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA. [37]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA. [38]Bee Research Lab, USDA-ARS, Beltsville, MD, 20705, USA. [39]Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland 4811, Australia. [40]Department of Biology, Indiana University, Bloomington, IN 47405, USA. [41]Institute of Physiology, Rheinisch-Westfaelische Technische Hochschule (RWTH) Aachen University, D-52074 Aachen, Germany. [42]Department of Molecular Biology, Umeå University, S-901 87 Umeå, Sweden. [43]Department of Entomology, University of Nebraska, Lincoln, NE 68583, USA. [44]Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya 464-8601, Japan. [45]School of Informatics, Indiana University, Bloomington, IN 47405, USA. [46]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [47]Department of Structural Biology and Bioinformatics, University of Geneva Medical School, CH-1211 Geneva, Switzerland. [48]Research School of Biology, Australian National University, Canberra, Australian Capital Territory 2601, Australia. [49]Roy J. Carver Center for Comparative Genomics and Department of Biology, University of Iowa, Iowa City, IA 52242, USA. [50]Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA. [51]Institute of General Zoology, University of Jena, D-7743 Jena, Germany. [52]Faculdade de Medicina de Ribeirão Preto, Departamento de Genética, Universidade de São Paulo, Ribeirão Preto, São Paolo 14049-900, Brazil. [53]Department of Entomology, Commonwealth Scientific and Industrial Research Organisation, Canberra, Australian Capital Territory 2601, Australia. [54]Institute of Evolutionary Biology–School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK. [55]Vector Group, Liverpool School of Tropical Medicine, Liverpool L3 5QA, UK. [56]Department of Molecular, Cell, and Developmental Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA. [57]Reese Consulting, 157/10 Tambon Ban Deau, Amphur Muang, Nong Khai, 43000, Thailand. [58]Department of Biology, Western Washington University, Bellingham, WA 98225, USA. [59]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA. [60]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [61]School of Marine and Tropical Biology and Centre for Comparative Genomics, James Cook University, Townsville, Queensland 4811, Australia. [62]Department of Evolutionary Biology and Animal Ecology, University of Freiburg, 79104 Freiburg, Germany. [63]School of Biology, University of St Andrews, St Andrews KY16 9TH, UK. [64]Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK. [65]Institut für Mikrobiologie und Genetik, Universität Göttingen, 37077 Göttingen, Germany. [66]Division of Insect Sciences, National Institute of Agrobiological Science, Tsukuba, Ibaraki 305-8634, Japan. [67]Department of Biology and Biocenter Oulu, University of Oulu, 90014 Oulu, Finland. [68]Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, MT 59717, USA. [69]Department of Cellular Biology, University of Georgia, Athens, GA 30602, USA. [70]Department of Biotechnology, Chemistry, and Food Science, Norwegian University of Life Sciences, N-1432 Ås, Norway.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: werr@mail.rochester.edu (J.H.W.); stephenr@bcm.tmc.edu (S.R.)
‡Current address: Verhaltensbiologie, Universität Osnabrück, 49076 Osnabrück, Germany.
§Current address: Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 8193 Bellaterra, Spain.
‖Current address: Weill Cornell Medical College, New York, NY 10065, USA.
¶Current address: Department of Epidemiology, University of Texas, M.D. Anderson Cancer Center, Houston, TX 77030, USA.

Supporting Online Material
www.sciencemag.org/cgi/content/full/327/5963/343/DC1
Materials and Methods
SOM Text
Figs. S1 to S25
Tables S1 to S57
References

22 June 2009; accepted 24 November 2009
10.1126/science.1178028

Zebrafish Behavioral Profiling Links Drugs to Biological Targets and Rest/Wake Regulation

Jason Rihel,[1]*† David A. Prober,[1]*‡ Anthony Arvanites,[2] Kelvin Lam,[2] Steven Zimmerman,[1] Sumin Jang,[1] Stephen J. Haggarty,[3,4,5] David Kokel,[6] Lee L. Rubin,[2] Randall T. Peterson,[3,6,7] Alexander F. Schier[1,2,3,8,9]†

A major obstacle for the discovery of psychoactive drugs is the inability to predict how small molecules will alter complex behaviors. We report the development and application of a high-throughput, quantitative screen for drugs that alter the behavior of larval zebrafish. We found that the multidimensional nature of observed phenotypes enabled the hierarchical clustering of molecules according to shared behaviors. Behavioral profiling revealed conserved functions of psychotropic molecules and predicted the mechanisms of action of poorly characterized compounds. In addition, behavioral profiling implicated new factors such as ether-a-go-go–related gene (ERG) potassium channels and immunomodulators in the control of rest and locomotor activity. These results demonstrate the power of high-throughput behavioral profiling in zebrafish to discover and characterize psychotropic drugs and to dissect the pharmacology of complex behaviors.

Most current drug discovery efforts focus on simple in vitro screening assays. Although such screens can be successful, they cannot recreate the complex network interactions of whole organisms. These limitations are particularly acute for psychotropic drugs because brain activity cannot be modeled in vitro (*1*–*3*). Motivated by recent small-molecule screens that probed zebrafish developmental processes (*4*–*7*), we developed a whole organism, high-throughput screen for small molecules that alter larval zebrafish locomotor behavior. We used an automated rest/wake behavioral assay (*3*, *8*) to monitor the activity of larvae exposed to small molecules at 10 to 30 μM for 3 days (Fig. 1A) (*3*). Multiple behavioral parameters were measured, including the number and duration of rest bouts, rest latency, and waking activity (i.e., activity not including time spent at rest) (Fig. 1B) (*3*). We screened 5648 compounds representing 3968 unique structures and 1680 duplicates and recorded more than 60,000 behavioral profiles. Of these, 547 compounds representing 463 unique structures significantly altered behavior relative to controls, according to a stringent statistical cutoff (*3*).

Because the alterations in behavior were multidimensional and quantitative, we assigned a behavioral fingerprint to each compound and applied clustering algorithms to organize molecules according to their fingerprints (Fig. 2A and figs. S1 to S3). This analysis organized the data set broadly into arousing and sedating compounds and identified multiple clusters corresponding to specific phenotypes (Fig. 2, B to F; Fig. 3, A to C; Fig. 4, B and C; and figs. S1 to S4). Clustering allowed us to address three questions: (i) Do structural, functional, and behavioral profiles overlap? (ii) Does the data set predict links between known and unknown small molecules and their mechanisms of action? (iii) Does the data set identify unexpected

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA. [2]Harvard Stem Cell Institute, Harvard University, Cambridge, MA 02138, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [4]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [5]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA. [6]Developmental Biology Laboratory, Cardiovascular Research Center, Massachusetts General Hospital, Charlestown, MA 02129, USA. [7]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA. [8]Division of Sleep Medicine, Harvard Medical School, Boston, MA 02215, USA. [9]Center for Brain Science, Harvard University, Cambridge, MA 02138, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: schier@fas.harvard.edu (A.F.S.); rihel@fas.harvard.edu (J.R.)
‡Present address: Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA.

A.1.4.3 *Supplementary information*

An extract of the supplementary material for the paper "Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species" covering my work is presented on pages 183–192. The original source can be found here:
`http://www.sciencemag.org/cgi/data/327/5963/343/DC1/1`

Analysis of transcript-form diversity: The tiling array experiments allowed detection of up to ten transcript forms from single gene models (maximum of one form per sample-type hybridized on the arrays). We counted a total of 37,009 transcript variants from 14,924 genes represented on the arrays. To create a validated set of transcript forms, 114,885 EST sequences were aligned via BLAST ($p < 10^{-100}$) against the predicted cDNA sequence for each detected transcript. Transcripts were unambiguously supported if one or more ESTs aligned uniquely to a variant. Similarly, if one or more ESTs aligned to more than one transcript-form of a gene, it was reported as ambiguously supported.

We compared the number of transcript-forms for *Nasonia* and *Drosophila melanogaster* by first obtaining a set of previously computed orthologs identified by OrthoDB (*S28*). 7,719 *Nasonia* genes were confidently assigned to 9,800 *D. melanogaster* orthologs within 6,650 ortholog groups. For our analysis, these were classified into sets of 1) one-to-one orthologs that were single copy genes in both species; 2) one-to-many orthologs that were single copy genes in *Nasonia*, yet duplicates were located in *Drosophila*; 3) many-to-one othologs that specified *Nasonia* paralogs related to single copy *Drosophila* genes; and 4) many-to-many orthologs that related multi-copy genes in both species. Transcription-form data for *D. melanogaster* were obtained from FlyBase (Release dmel_r5.16 mRNA) (http://flybase.org/).

4. MicroRNA Prediction

Four different sets of computational microRNA predictions were generated by two different groups. Both groups created homology-based sets (Sets 1a and 2a) using slightly different strategies. The two groups also produced miRNA predictions based on sequence conservation: Set 2a was based on comparison with 40 animal species, and set 2b was based on comparison between *N. vitripennis* and *A. mellifera*. An additional set (set 3) was generated using bioinformatics analysis of 454 sequencing reads from small RNA libraries of *N. vitripennis*.

a. miRNA prediction set 1a: First homology-based prediction set

All known metazoan miRNAs from miRBase 12.0 (*S29*) were aligned to the *Nasonia vitripennis* genome assembly v1.0 using BLASTN of the WU-BLAST package [BLASTN 2.0MP-WashU (http://blast.wustl.edu)] with the following parameters adapted for cross species comparison: -M 1 -N -1 -Q 3 -R 2 -W 9 -filter dust -mformat 2 -hspsepSmax 40 -e 1e-3. BLASTN matches longer than 20 bp were extended at both ends to match the length of the query sequence. In a following step the extended blast hits were aligned to their query sequence using MAFFT (*S30*) with the following parameters: --maxiterate 1000 --localpair -quiet. To remove unstable or spurious hits a set of features were calculated for each hit and evaluated: 1) total sequence length > 40 bp, 2) 100% conserved seed (nt 2-8 of the putative mature part) region in regard to the query sequence, 3) more than 90% sequence identity for the mature part, 4) sequence conservation of the total precursor sequence larger than 60%, 5) no more than two gaps in the mature

part, 6) minimum free folding energy smaller -15 kcal/mol, 7) more than 40% of the bases should be paired, 8) mature regions should not overlap a multi-branch loop, 9) RandFold p-value smaller 0.05 if precursor conservation smaller 95% to any known miRNA. RandFold (*S31*) estimates the stability of the folding compared to dinucleotide shuffled folded sequences (100 randomizations). As the query set was redundant (e.g. containing dme-bantam, ame-bantam) the final predictions were clustered according to their locus on the *Nasonia* genome using GALAXY (*S32*). From a single locus the hit with the highest conservation of the mature miRNA and the highest overall percent alignment identity over the entire putative pre-miRNA was used as a single representative sequence. Using this approach, we predicted 51 microRNAs in *Nasonia* (Table S5 available from supporting data sets and online at http://nasoniabase.org/nasonia_genome_consortium/datasets.html).

b. miRNA prediction set 1b: Comparative sequence approach based on SVM

We used a comparative approach on the basis of a Support Vector Machine (SVM) model of hairpin-like structures followed by an orthology assignment step. This method allows prediction of novel miRNAs that do not show sequence homology to known miRNAs. The complete method is described in (*S33*) and the results are available from http://cegg.unige.ch/nasonia_genome. What follows is a brief outline of the basic principles: First, an ab initio SVM model was created to score stem-loop like sequences extracted from the genomic sequence with RNAfold (*S34*). Second, an orthology assignment pipeline grouped putative precursors from over 40 animal species, then precursors within groups were aligned. In a third step the orthologous groups were again subjected to an SVM model designed to distinguish alignments of orthologous miRNA sequences from other ncRNA alignments or false positive predictions, taking into account typical conservation patterns in pre-miRNA sequence alignments. Those predictions were put forward as the miRNA prediction set 1b if they had an ortholog in at least the bee genome and were supported by a 454 sequence read (described under miRNA prediction set 3 below). In total this method yielded three new *Nasonia* miRNAs which are conserved in the bee genome and supported by a high-throughput read, but not present in set 1a.

Using a 10 Kb window 17 out of the 54 *Nasonia* miRNA from combined miRNA prediction sets 1 and 2 group in cluster containing two or more miRNAs. Four clusters contain two miRNAs, one cluster three and the biggest cluster contains 6 miRNAs within a 1.9 Kb long region (nvit-mir-71; nvit-mir-2a, nvit-mir-13a; nvit-mir-13b; nvit-mir-2b; nvit-mir-2c). All *Nasonia* miRNA in this cluster are on the forward strand, whereas in the honey bee the miRNAs of this cluster are all on the reverse strand. Although the synteny is conserved, in honey bee the miRNAs span a larger chromosomal region (2.7 Kb) than in *Nasonia*. Except the six miRNA in this region, an alignment of the honey bee region vs. the *Nasonia* region does not show any significant sequence conservation (Fig. S13).

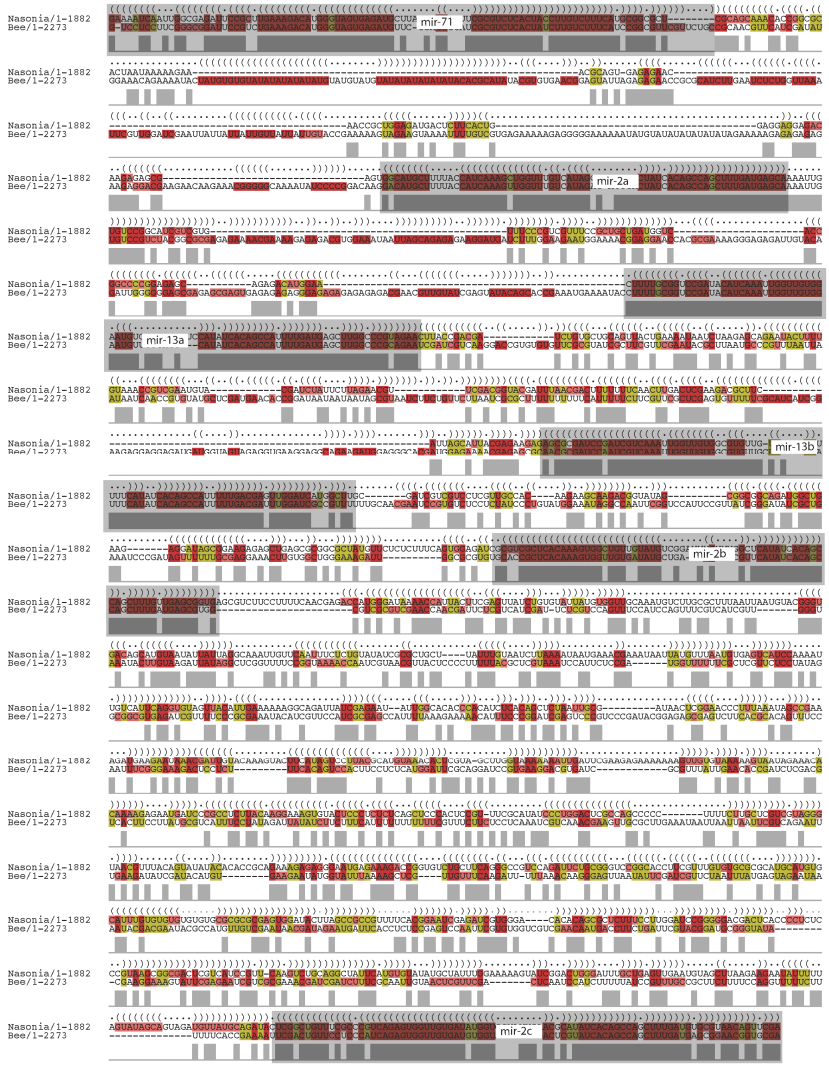


Figure S13: Genomic cluster of microRNA genes conserved between *Nasonia* and honeybee. A conserved cluster of 6 miRNAs (nvit-mir-71; nvit-mir-2a, nvit-mir-13a; nvit-mir-13b; nvit-mir-2b; nvit-mir-2c) spans a 1.9 kb in Nasonia, and 2.7 kb in honeybee. Apart from the miRNA encoding regions the alignment shows no significant sequence conservation between these two Hymenoptera species along the syntenic block. The region was color coded to visualize consistent and compensatory base changes supporting the common RNA secondary structure by folding with RNAalifold software.

References

1. J. A. J. Breeuwer, J. H. Werren, *Evolution* **49**, 705 (1995).
2. J. H. Werren, *Annu. Rev. Entomol.* **42**, 587 (1997).
3. S. R. Bordenstein, F. P. O'Hara, J. H. Werren, *Nature* **409**, 707 (Feb 8, 2001).
4. P. Havlak *et al.*, *Genome Res* **14**, 721 (Apr, 2004).
5. A. Coghlan *et al.*, *BMC Bioinformatics* **9**, 549 (2008).
6. Y. Kapustin, A. Souvorov, T. Tatusova, paper presented at the RECOMB 2004 - Currents in Computational Molecular Biology., 2004.
7. B. Kiryutin, A. Souvorov, paper presented at the ISMB 2005, 2005.
8. A. Souvorov, T. Tatusova, D. Lipman, paper presented at the ISMB 2004, 2004.
9. B. J. Haas *et al.*, *Nucleic Acids Res* **31**, 5654 (Oct 1, 2003).
10. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* **35**, D61 (Jan, 2007).
11. D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova, *Nucleic Acids Res.* **35**, D26 (Jan, 2007).
12. M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, *Bioinformatics* **24**, 637 (Mar 1, 2008).
13. V. Solovyev, P. Kosarev, I. Seledsov, D. Vorobyev, *Genome Biol.* **7 Suppl 1**, S10 1 (2006).
14. C. G. Elsik *et al.*, *Genome Biol* **8**, R13 (2007).
15. A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (Apr, 2000).
16. V. Solovyev, in *Handbook of Statistical Genetics,* D. J. Balding, M. Bishop, C. Cannings, Eds. (John Wiley & Sons, Chichester, England; Hoboken, NJ, 2007), pp. 97-159.
17. M. Stanke, S. Waack, *Bioinformatics* **19 Suppl 2**, ii215 (Oct, 2003).
18. A. Bairoch, B. Boeckmann, S. Ferro, E. Gasteiger, *Brief. Bioinform.* **5**, 39 (2004).
19. G. S. Slater, E. Birney, *BMC Bioinformatics* **6**, 31 (2005).
20. T. D. Wu, C. K. Watanabe, *Bioinformatics* **21**, 1859 (May 1, 2005).
21. W. R. Pearson, D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444 (Apr, 1988).
22. S. E. Lewis *et al.*, *Genome Biol* **3**, RESEARCH0082 (2002).
23. W. B. Hunter *et al.*, *J Insect Sci* **3**, 23 (2003).
24. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (Oct 5, 1990).
25. Z. Tang *et al.*, *BMC Genomics* **10**, 174 (2009).
26. W. J. Kent, *Genome Res* **12**, 656 (Apr, 2002).
27. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, *Bioinformatics* **19**, 185 (Jan 22, 2003).
28. E. V. Kriventseva, N. Rahman, O. Espinosa, E. M. Zdobnov, *Nucleic Acids Res* **36**, D271 (Jan, 2008).

29. S. Griffiths-Jones, H. K. Saini, S. van Dongen, A. J. Enright, *Nucleic Acids Res* **36**, D154 (Jan, 2008).
30. K. Katoh, H. Toh, *BMC Bioinformatics* **9**, 212 (2008).
31. E. Bonnet, J. Wuyts, P. Rouze, Y. Van de Peer, *Bioinformatics* **20**, 2911 (Nov 22, 2004).
32. B. Giardine *et al.*, *Genome Res* **15**, 1451 (Oct, 2005).
33. D. Gerlach, E. V. Kriventseva, N. Rahman, C. E. Vejnar, E. M. Zdobnov, *Nucleic Acids Res* **37**, D111 (Jan, 2009).
34. I. L. Hofacker, *Curr Protoc Bioinformatics* **Chapter 12**, Unit 12 2 (Feb, 2004).
35. C. Notredame, D. G. Higgins, J. Heringa, *J. Mol. Biol.* **302**, 205 (Sep 8, 2000).
36. W. R. Pearson, *Methods Enzymol.* **183**, 63 (1990).
37. Z. Zhang, S. Schwartz, L. Wagner, W. Miller, *J Comput. Biol.* **7**, 203 (Feb-Apr, 2000).
38. O. Niehuis *et al.*, *PLoS ONE* DOI: 10.1371/journal.pone.0008597 (Jan, 2010).
39. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res* **8**, 175 (Mar, 1998).
40. X. Wang, B. Seed, *Bioinformatics* **19**, 796 (May 1, 2003).
41. D. W. Loehlin, L. S. Enders, J. H. Werren, *Heredity* DOI: 10.1038/hdy.2009.146 (Jan, 2010).
42. R. C. Edgar, E. W. Myers, *Bioinformatics* **21 Suppl 1**, i152 (Jun, 2005).
43. R. C. Edgar, *BMC Bioinformatics* **5**, 113 (Aug 19, 2004).
44. A. F. A. Smit, R. Hubley, P. Green. (1996-2004), pp. RepeatMasker Open-3.0.
45. G. Benson, *Nucleic Acids Res.* **27**, 573 (Jan 15, 1999).
46. R. D. Finn *et al.*, *Nucleic Acids Res* **36**, D281 (Jan, 2008).
47. D. G. Eickbush, T. H. Eickbush, J. H. Werren, *Chromosoma* **101**, 575 (Aug, 1992).
48. J. Jurka *et al.*, *Cytogenet Genome Res* **110**, 462 (2005).
49. C. D. Smith *et al.*, *Gene* **389**, 1 (Mar 1, 2007).
50. M. A. McClure *et al.*, *Genomics* **85**, 512 (Apr, 2005).
51. W. Gish, *http://blast.wustl.edu*, (1996-2004).
52. Z. Xu, H. Wang, *Nucleic acids research* **35**, W265 (Jul 1, 2007).
53. C. D. Smith, S. Shu, C. J. Mungall, G. H. Karpen, *Science* **316**, 1586 (Jun 15, 2007).
54. H. M. Robertson, K. H. Gordon, *Genome Res.* **16**, 1345 (Nov, 2006).
55. M. Osanai, K. K. Kojima, R. Futahashi, S. Yaguchi, H. Fujiwara, *Gene* **376**, 281 (Jul 19, 2006).
56. R. Frydrychova, P. Grossmann, P. Trubac, M. Vitkova, F. Marec, *Genome* **47**, 163 (Feb, 2004).
57. M. Vitkova, J. Kral, W. Traut, J. Zrzavy, F. Marec, *Chromosome Res.* **13**, 145 (2005).
58. H. Fujiwara, M. Osanai, T. Matsumoto, K. K. Kojima, *Chromosome Res* **13**, 455 (2005).

59. S. Richards *et al.*, *Nature* **452**, 949 (Apr 24, 2008).
60. M. G. Goll, T. H. Bestor, *Annu Rev Biochem* **74**, 481 (2005).
61. N. Kunert, J. Marhold, J. Stanke, D. Stach, F. Lyko, *Development* **130**, 5083 (Nov, 2003).
62. J. Marhold *et al.*, *Insect Mol. Biol.* **13**, 117 (Apr, 2004).
63. A. Dong *et al.*, *Nucleic Acids Res* **29**, 439 (Jan 15, 2001).
64. R. Maleszka, *Epigenetics* **3**, 188 (Jul-Aug, 2008).
65. R. Albalat, *Dev. Genes Evol.* **218**, 691 (Dec, 2008).
66. C. P. Ponting, *Nature Rev. Genet.* **9**, 689 (Sep, 2008).
67. H. G. S. Consortium, *Nature* **443**, 931 (Oct 26, 2006).
68. M. J. Sharkey, *Zootaxa* **1668**, 521 (2007).
69. D. Grimaldi, M. S. Engel, *Evolution of the Insects*. (Cambridge University Press, Cambridge, U.K., 2005).
70. D. L. J. Quicke, H. H. Basibuyuk, M. G. Fitton, A. P. Rasnitsyn, *Zoologica Scripta* **28**, 175 (1999).
71. J. B. Whitfield, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7508 (May 28, 2002).
72. J. Savard *et al.*, *Genome Res.* **16**, 1334 (Nov, 2006).
73. V. Krauss *et al.*, *Mol. Biol. Evol.* **25**, 821 (May, 2008).
74. B. M. Wiegmann *et al.*, *BMC Biol* **7**, 34 (2009).
75. J. B. Whitfield, K. M. Kjer, *Annu. Rev. Entomol.* **53**, 449 (2008).
76. J. C. Regier *et al.*, *Syst. Biol.* **57**, 920 (2008).
77. J. C. Regier, J. W. Shultz, R. E. Kambic, *Proc. R. Soc. London Ser. B* **272**, 395 (Feb 22, 2005).
78. S. J. Bourlat, C. Nielsen, A. D. Economou, M. J. Telford, *Mol. Phylogenet. Evol.* **49**, 23 (Oct, 2008).
79. P. Bernaola-Galván, R. Román-Roldán, J. L. Oliver, *Physical Review E* **53**, 5181 (1996).
80. N. Cohen, T. Dagan, L. Stone, D. Graur, *Mol. Biol. Evol.* **22**, 1260 (May, 2005).
81. M. Weber *et al.*, *Nature Genet.* **39**, 457 (Apr, 2007).
82. S. Suzuki *et al.*, *PLoS Genet* **3**, e55 (Apr 13, 2007).
83. S. R. Eddy, *Bioinformatics* **14**, 755 (1998).
84. A. Heger, C. A. Wilton, A. Sivakumar, L. Holm, *Nucleic Acids Res* **33**, D188 (Jan 1, 2005).
85. Y. Wurm *et al.*, *BMC Genomics* **10**, 5 (2009).
86. I. M. Wallace, O. O'Sullivan, D. G. Higgins, C. Notredame, *Nucleic Acids Res.* **34**, 1692 (2006).
87. A. Stamatakis, *Bioinformatics* **22**, 2688 (Nov 1, 2006).
88. F. Abascal, R. Zardoya, D. Posada, *Bioinformatics* **21**, 2104 (May 1, 2005).
89. D. T. Jones, W. R. Taylor, J. M. Thornton, *Comput Appl Biosci* **8**, 275 (Jun, 1992).
90. G. Talavera, J. Castresana, *Syst Biol* **56**, 564 (Aug, 2007).
91. S. Guindon, O. Gascuel, *Syst Biol* **52**, 696 (Oct, 2003).
92. A. Subramanian, H. Kuehn, J. Gould, P. Tamayo, J. P. Mesirov, *Bioinformatics* **23**, 3251 (Dec 1, 2007).

93. R. Durbin, S. Eddy, A. Krogh, G. Mitchison, in *Biological sequence analysis: probablistic models of proteins and nucleic acids* (Cambridge University Press, Cambridge, 1998).
94. K. Rutherford *et al.*, *Bioinformatics* **16**, 944 (Oct, 2000).
95. K. Tamura, J. Dudley, M. Nei, S. Kumar, *Mol Biol Evol* **24**, 1596 (Aug, 2007).
96. F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* **19**, 1572 (Aug 12, 2003).
97. J. C. Wilgenbusch, D. Swofford, *Curr Protoc Bioinformatics* **Chapter 6**, Unit 6 4 (Feb, 2003).
98. D. Posada, K. A. Crandall, *Bioinformatics* **14**, 817 (1998).
99. R. Chenna *et al.*, *Nucleic Acids Res.* **31**, 3497 (Jul 1, 2003).
100. J. Brennecke *et al.*, *Cell* **128**, 1089 (Mar 23, 2007).
101. N. Elango, S. V. Yi, *Mol Biol Evol* **25**, 1602 (Aug, 2008).
102. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (Sep 1, 1997).
103. S. Foret, R. Maleszka, *Genome Res* **16**, 1404 (Nov, 2006).
104. Z. Yang, *Mol. Biol. Evol.* **24**, 1586 (Aug, 2007).
105. J. D. Storey, *J. R. Stat. Soc., Ser. B* **64**, 479 (2002).
106. R. Raychoudhury, L. Baldo, D. C. Oliveira, J. H. Werren, *Evolution* **63**, 165 (Jan, 2009).
107. T. A. Hall, *Nucleic Acids Symposium Series* **41**, 95 (1999).
108. J. Rozas, J. C. Sanchez-DelBarrio, X. Messeguer, R. Rozas, *Bioinformatics* **19**, 2496 (Dec 12, 2003).
109. F. Wolschin, G. Gadau, *PLoS ONE* **4: e6394**, (2009).
110. L. Y. Geer *et al.*, *J Proteome Res* **3**, 958 (Sep-Oct, 2004).
111. N. Peiren *et al.*, *FEBS Lett* **580**, 4895 (Sep 4, 2006).
112. F. Vanrobaeys, R. Van Coster, G. Dhondt, B. Devreese, J. Van Beeumen, *J Proteome Res* **4**, 2283 (Nov-Dec, 2005).
113. K. A. Reidegeld *et al.*, *Proteomics* **8**, 1129 (Mar, 2008).
114. R. M. Waterhouse *et al.*, *Science* **316**, 1738 (Jun 22, 2007).
115. A. Bateman *et al.*, *Nucleic Acids Res.* **27**, 260 (Jan 1, 1999).
116. J. D. Evans *et al.*, *Insect Mol Biol* **15**, 645 (Oct, 2006).
117. M. W. Pfaffl, *Nucleic Acids Res* **29**, e45 (May 1, 2001).
118. J. van de Assem, in *Insect Parasitoids, 13th Symposium of the Royal Entomological Society of London,* J. K. Waage, D. J. Greathead, Eds. (Academic Press, London, 1986), pp. 137-167.
119. J. van den Assem, J. H. Werren, *J. Insect Behav.* **7**, 53 (1994).
120. J. W. Van Ooijen, M. P. Boer, C. Jansen, C. Maliepaard. (Plant Research International, Wageningen, the Netherlands, 2002).
121. R. Maleszka, R. Kucharski, *Biochem. Biophys. Res. Commun.* **270**, 773 (Apr 21, 2000).
122. M. D. Drapeau, S. Albert, R. Kucharski, C. Prusko, R. Maleszka, *Genome Res* **16**, 1385 (Nov, 2006).
123. J. E. Rebers, J. H. Willis, *Insect Biochem Mol Biol* **31**, 1083 (Oct, 2001).
124. R. S. Cornman *et al.*, *BMC Genomics* **9**, 22 (2008).
125. R. S. Cornman, J. H. Willis, *Insect Biochem Mol Biol* **38**, 661 (Jun, 2008).
126. M. V. Karouzou *et al.*, *Insect Biochem Mol Biol* **37**, 754 (Aug, 2007).

127. R. Futahashi *et al.*, *Insect Biochem Mol Biol* **38**, 1138 (Dec, 2008).
128. N. He *et al.*, *Insect Biochem Mol Biol* **37**, 135 (Feb, 2007).
129. R. Kucharski, J. Maleszka, R. Maleszka, *Insect Biochem Mol Biol* **37**, 128 (Feb, 2007).
130. T. Togawa, W. Augustine Dunn, A. C. Emmons, J. H. Willis, *Insect Biochem Mol Biol* **37**, 675 (Jul, 2007).
131. X. Guan, B. W. Middlebrooks, S. Alexander, S. A. Wasserman, *Proc Natl Acad Sci U S A* **103**, 16794 (Nov 7, 2006).
132. E. Danty *et al.*, *J Neurosci* **19**, 7468 (Sep 1, 1999).
133. J. Maleszka, S. Foret, R. Saint, R. Maleszka, *Dev Genes Evol* **217**, 189 (Mar, 2007).
134. E. M. Rasch, J. D. Cassidy, R. C. King, *Chromosoma* **59**, 323 (Feb 23, 1977).
135. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (Sep, 2001).
136. D. E. Stage, T. H. Eickbush, *Insect Mol. Biol.* **19**, 37 (Jan, 2010).
137. T. P. Jurkowski *et al.*, *RNA* **14**, 1663 (Aug, 2008).
138. S. Kumar *et al.*, *Nucleic Acids Res* **22**, 1 (Jan 11, 1994).
139. X. Cheng, R. M. Blumenthal, *Structure* **16**, 341 (Mar, 2008).
140. T. Yokomine, K. Hata, M. Tsudzuki, H. Sasaki, *Cytogenet. Genome Res.* **113**, 75 (2006).
141. N. Elango, B. B. Hunt, M. A. D. Goodisman, S. V. Yi, *Proc Natl Acad Sci U S A* **106**, 11206 (Jul, 2009).
142. M. M. Suzuki, A. R. Kerr, D. De Sousa, A. Bird, *Genome Res* **17**, 625 (May, 2007).
143. G. B. Saul, *Genetics Maps*. S. J. O'Brien, Ed., (Cold Spring Harbor Press, Cold Spring Harbor, NY, 1993).
144. S. L. Ryan, G. B. Saul, 2nd, G. W. Conner, *J Hered* **78**, 273 (Jul-Aug, 1987).
145. F. Perfectti, J. H. Werren, *Evolution* **55**, 1069 (May, 2001).
146. M. J. Perrot-Minnot, J. H. Werren, *Heredity* **87**, 8 (Jul, 2001).
147. A. Bhutkar, S. M. Russo, T. F. Smith, W. M. Gelbart, *Genome Res.* **17**, 1880 (Dec, 2007).
148. M. Ashburner *et al.*, *Nat Genet* **25**, 25 (May, 2000).
149. G. Bernardi, *Gene* **241**, 3 (2000).
150. E. S. Lander *et al.*, *Nature* **409**, 860 (Feb 15, 2001).
151. W. Li, *Gene* **300**, 129 (Oct 30, 2002).
152. W. Li, P. Bernaola-Galvan, P. Carpena, J. L. Oliver, *Comput Biol Chem* **27**, 5 (Feb, 2003).
153. O. Clay, G. Bernardi, *Mol. Biol. Evol.* **22**, 2315 (Dec, 2005).
154. E. M. Zdobnov, P. Bork, *Trends Genet.* **23**, 16 (Jan, 2007).
155. S. Wyder, E. V. Kriventseva, R. Schroder, T. Kadowaki, E. M. Zdobnov, *Genome Biol* **8**, R242 (2007).
156. L. K. Mosavi, T. J. Cammett, D. C. Desrosiers, Z. Y. Peng, *Protein Sci.* **13**, 1435 (Jun, 2004).
157. I. Letunic, T. Yamada, M. Kanehisa, P. Bork, *Trends Biochem. Sci.* **33**, 101 (Mar, 2008).

158. M. Kanehisa *et al.*, *Nucleic Acids Research* **36**, D480 (Jan, 2008).
159. L. R. Serbus, C. Casper-Lindley, F. Landmann, W. Sullivan, *Annu. Rev. Genet.* **42**, 683 (2008).
160. S. R. Bordenstein, M. L. Marshall, A. J. Fry, U. Kim, J. J. Wernegreen, *PLoS Pathog.* **2**, e43 (May, 2006).
161. U. Tram, P. M. Ferree, W. Sullivan, *Microbes Infect.* **5**, 999 (Sep, 2003).
162. S. J. Chang *et al.*, *J. Virol.* **83**, 4140 (May, 2009).
163. S. Sonnberg, B. T. Seet, T. Pawson, S. B. Fleming, A. A. Mercer, *Proc Natl Acad Sci U S A* **105**, 10955 (Aug 5, 2008).
164. T. Walker *et al.*, *BMC Biol* **5**, 39 (2007).
165. N. H. Cho *et al.*, *Proc Natl Acad Sci U S A* **104**, 7981 (May 8, 2007).
166. J. C. Dunning-Hotopp *et al.*, *Science* **317**, 1753 (Sep 21, 2007).
167. M. S. Dushay, B. Asling, D. Hultmark, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10343 (Sep 17, 1996).
168. Y. C. Zhu, A. K. Dowdy, J. E. Baker, *Pesticide Science* **55**, 398 (1999).
169. J. H. Werren, S. W. Skinner, A. M. Huger, *Science* **231**, 990 (Feb 28, 1986).
170. O. Duron *et al.*, *BMC Biol* **6**, 27 (2008).
171. T. Wilkes *et al.*, *Insect Mol. Biol.* **19**, 59 (Jan, 2010).
172. M. M. Pearson *et al.*, *J Bacteriol* **190**, 4027 (Jun, 2008).
173. G. R. Erdmann, *Parasitol Today* **3**, 214 (Jul, 1987).
174. M. I. Rosenberg, J. A. Lynch, C. Desplan, *Biochim. Biophys. Acta*, **1789**, 333-342 (April, 2009).
175. J. A. Lynch, A. E. Brent, D. S. Leaf, M. A. Pultz, C. Desplan, *Nature* **439**, 728 (Feb 9, 2006).
176. A. E. Brent, G. Yucel, S. Small, C. Desplan, *Science* **315**, 1841 (Mar 30, 2007).
177. E. C. Olesnicky *et al.*, *Development* **133**, 3973 (Oct, 2006).
178. P. Z. Liu, T. C. Kaufman, *Evol. Dev.* **7**, 629 (Nov-Dec, 2005).
179. J. A. Lynch, E. C. Olesnicky, C. Desplan, *Dev. Genes Evol.* **216**, 493 (Jul-Aug, 2006).
180. M. A. Pultz *et al.*, *Genetics* **154**, 1213 (Mar, 2000).
181. J. M. Cook, *Heredity* **71**, (1993).
182. M. Beye, *Bioessays* **26**, 1131 (Oct, 2004).
183. S. W. Skinner, J. H. Werren, *Genetics* **94**, 98 (1980).
184. D. C. S. G. Oliveira *et al.*, *Insect Mol. Biol.*, **19**, 99 (Jan, 2010).
185. M. L. Hedley, T. Maniatis, *Cell* **65**, 579 (May 17, 1991).
186. V. Heinrichs, L. C. Ryner, B. S. Baker, *Mol. Cell. Biol.* **18**, 450 (Jan, 1998).
187. R. C. Bertossa, L. van de Zande, L. W. Beukeboom, *Mol Biol Evol* **26**, 1557 (Jul, 2009).
188. M. Hasselmann *et al.*, *Nature* **454**, 519 (Jul 24, 2008).
189. B. A. Pannebakker *et al.*, *Evolution* **62**, 1921 (Aug, 2008).
190. F. Hauser *et al.*, *Front Neuroendocrinol* **29**, 142 (Jan, 2008).
191. A. K. Jones, A. N. Bera, K. Lees, D. B. Sattelle, *Heredity* DOI: 10.1038/hdy.2009.97 (Jan, 2010).

192. R. Acher, J. Chauvet, M. T. Chauvet, *Adv. Exp. Med. Biol.* **395**, 615 (1995).
193. E. Stafflinger *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 3262 (Mar 4, 2008).
194. H. Lin, K. J. Mann, E. Starostina, R. D. Kinser, C. W. Pikielny, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12831 (Sep 6, 2005).
195. L. Liu *et al.*, *Nature* **450**, 294 (Nov 8, 2007).
196. L. R. Bell, E. M. Maine, P. Schedl, T. W. Cline, *Cell* **55**, 1037 (Dec 23, 1988).
197. J. M. Belote, M. McKeown, R. T. Boggs, R. Ohkawa, B. A. Sosnowski, *Dev Genet* **10**, 143 (1989).
198. K. C. Burtis, B. S. Baker, *Cell* **56**, 997 (Mar 24, 1989).
199. J. W. Erickson, J. J. Quintero, *PLoS Biol* **5**, e332 (Dec, 2007).
200. A. Dubendorfer, M. Hediger, G. Burghardt, D. Bopp, *Int J Dev Biol* **46**, 75 (Jan, 2002).
201. G. Saccone *et al.*, First research coordination meeting, IAEA/FAO, Vienna (1996).
202. A. Pane, M. Salvemini, P. Delli Bovi, C. Polito, G. Saccone, *Development* **129**, 3715 (Aug, 2002).
203. D. Lagos, M. F. Ruiz, L. Sanchez, K. Komitopoulou, *Gene* **348**, 111 (Mar 28, 2005).
204. D. Lagos, M. Koukidou, C. Savakis, K. Komitopoulou, *Insect Mol Biol* **16**, 221 (Apr, 2007).
205. M. F. Ruiz *et al.*, *Genetics* **171**, 849 (Oct, 2005).
206. M. F. Ruiz *et al.*, *PLoS ONE* **2**, e1239 (2007).
207. M. Beye, M. Hasselmann, M. K. Fondrk, R. E. Page, S. W. Omholt, *Cell* **114**, 419 (Aug 22, 2003).
208. S. Cho, Z. Y. Huang, J. Zhang, *Genetics* **177**, 1733 (Nov, 2007).
209. J. D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, *J Mol Biol* **340**, 783 (Jul 16, 2004).
210. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, *J Mol Biol* **305**, 567 (Jan 19, 2001).

A.1.5 *Genome Sequences of the Human Body Louse and its Primary Endosymbiont: Insights into the Permanent Parasitic Lifestyle*

Body Louse Genome Working Group (MicroRNA analysis: Gerlach D, and Zdobnov EM). Genome Sequences of the Human Body Louse and its Primary Endosymbiont: Insights into the Permanent Parasitic Lifestyle. *Proc. Natl. Acad. Sci. U.S.A.* (2010) In Press.

A.1.5.1 *Contributions*

The Body Louse Genome Working Group presents the complete genomic sequences of the human body louse *Pediculus humanus humanus*.

I contributed to the work in predicting a set of miRNA genes using the `miROrtho` pipeline and describing important insect miRNAs that seem to be lost in the body louse genome.

A.1.5.2 *Main paper*

See pages 194–201.

Genome sequences of the human body louse and its primary endosymbiont: insights into the permanent parasitic lifestyle

Ewen F. Kirkness [1], Brian J. Haas [1], Weilin Sun [2], Henk R. Braig [3], M. Alejandra Perotti [4], John M. Clark [5], Si Hyeock Lee [6], Hugh M. Robertson [2], Ryan C. Kennedy [7], Eran Elhaik [8], Daniel Gerlach [9], Evgenia V. Kriventseva [9], Christine G. Elsik [10], Dan Graur [8], Catherine A. Hill [11], Jan A. Veenstra [12], Brian Walenz [1], Jose Manuel C. Tubío [13], Jose M.C. Ribeiro [14], Julio Rozas [15], J. Spencer Johnston [16], Justin T. Reese [10], Aleksandar Popadic [17], Yoshi Tomoyasu [4], Marta Tojo [18], Didier Raoult [19], David L. Reed [20], Emily Krause [21], Omprakash Mittapalli [22], Venu M. Margam [11], Hong-Mei Li [2], Jason M. Meyer [11], Reed Johnson [2], Jeanne Romero-Severson [7], Janice Pagel VanZee [11], David Alvarez-Ponce [15], Filipe G. Vieira [15], Montserrat Aguadé [15], Sara Guirao-Rico [15], Juan M. Anzola [10], Kyong Sup Yoon [5], Joseph P. Strycharz [5], Maria F. Unger [7], Scott Christley5, Marta Tojo [13], Neil F. Lobo [7], Manfredo J. Seufferheld [23], NaiKuan Wang [24], Gregory A. Dasch [25], Claudio J. Struchiner26Laboratory of Malaria and Vector Research, 12735 Twinbrook Parkway, Room 2E-32 Twinbrook III Building NIAID, NIH, MSC 8132, Bethesda, MD 20892-8132, USA, Greg Madey [7], Linda I. Hannick [1], Shelby Bidwell [1], Vinita Joardar [1], Elisabeth Caler [1], Renfu Shao [26], Stephen Barker [26], Stephen Cameron [27], Robert V. Bruggner [7], Allison Regier [7], Justin Johnson [1], Lakshmi Viswanathan [1], Terry R. Utterback [1], Granger G. Sutton [1], Daniel Lawson [28], Robert M. Waterhouse [9]. Craig Venter [1], Robert L. Strausberg [1], May Berenbaum [2], Frank H. Collins [7], Evgeny M. Zdobnov [9], and Barry R. Pittendrigh [2]

[1]Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA,[2]Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA,[28]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK,[9]Department of Genetic Medicine and Development, University of Geneva Medical School, 1 rue Michel-Servet, Geneva 1211, Switzerland,[7]Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN 46556, USA,[10]Department of Animal Science, Texas A&M University, College Station, TX 77843, USA,[16]Department of Entomology, Texas A&M University, College Station, TX 77843, USA,[26]School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, Brisbane, Queensland 4072, Australia,[27]Australian National Insect Collection & CSIRO Entomology, Canberra ACT 2601, Australia,[25]Centers for Disease Control and Prevention, 1600 Clifton Rd. NE, MS G-13, Atlanta, GA 30333, USA,[20]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA,[3]School of Biological Sciences, Bangor University, Bangor, LL57 2UW, Wales, UK,[11]Department of Entomology, Purdue University, West Lafayette, Indiana 47907, USA,[6]Department of Agricultural Biotechnology, Seoul National University, Seoul, South Korea,[5]Department of Veterinary and Animal Science, University of Massachusetts, Amherst, MA 01003, USA,[8]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA,[22]Department of Entomology, OARDC/ The Ohio State University, Wooster, OH 44691, USA,[17]Department of Biological Sciences, Wayne State University, Detroit, MI 48202, USA,[4]School of Biological Sciences, University of Reading, Reading, RG6 6AS, UK,[18]Servicio de Anatomía Patolóxica, Complexo Hospitalario Universitario de Santiago de Compostela, Santiago de Compostela 15706, Spain,[13]Servicio de Hematoloxía, Complexo Hospitalario Universitario de Santiago de Compostela, Santiago de Compostela 15706, Spain,[14]Departamento de Xenética, CIBUS, Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain,[15]Departament de Genètica, Universitat de Barcelona, Av. Diagonal 645, 08028-Barcelona, Spain,[19]Unité des Rickettsies, 27, boulevard Jean Moulin, 13385 Marseille Cedex 05, France,[12]School of Biological Sciences, Bangor University, Bangor, LL57 2UW, Wales, UK,[21]Department of Entomology Kansas State University, Manhattan, KS 66502,[23]Department of Crop Sciences, 311 Edward R. Madigan Laboratory, West Gregory Drive, Urbana, IL 61801, and [24]Chung Hwa College of Medical Technology, 89, Wen-Hwa 1st Street, Jen-Te Hsiang, Tainan, Taiwan 700

Submitted to Proceedings of the National Academy of Sciences of the United States of America

As an obligatory parasite of humans, the body louse (*Pediculus humanus humanus*) is an important vector for human diseases, including epidemic typhus, relapsing fever, and trench fever. Here we present genome sequences of the body louse and its primary bacterial endosymbiont *Candidatus Riesia pediculicola.* The body louse has the smallest known insect genome, spanning 108 Mb. Despite its status as an obligate parasite, it retains a remarkably complete "basal insect" repertoire of 10,773 protein-coding genes and 57 microRNAs. Representing hemimetabolous insects the genome of the body louse thus provides the most complete reference for studies of holometabolous insects. Compared with other insect genomes, the body louse genome contains significantly fewer genes associated with environmental sensing and response, including odorant and gustatory receptors and detoxifying enzymes. The unique architecture of the 18 mini-circular body louse mitochondrial chromosomes may be linked to the loss of the gene encoding the mitochondrial single-stranded DNA binding protein. The genome of the obligatory louse endosymbiont *Candidatus Riesia pediculicola* encodes fewer than 600 genes on a short, linear chromosome and a circular plasmid. The plasmid harbors a unique arrangement of genes required for the synthesis of pantothenate, an essential vitamin deficient in the louse diet. The human body louse, its primary endosymbiont, and the bacterial pathogens it vectors all possess genomes reduced in size in comparison to their free-living close relatives. The body louse genome project thus offers unique information and tools to use in advancing understanding of coevolution among vectors, symbionts, and pathogens.

Like their primate relatives, humans have had a long evolutionary association with parasitic sucking lice. Contact between sucking lice and primate hosts dates back at least 25 million years (1). Chimpanzee lice (*Pediculus schaeffi*) and human lice (*Pediculus humanus*) diverged from their common ancestors, as did chimpanzees (*Pan troglodytes*) and humans (*Homo sapiens*), five to seven million years ago (2, 3). The two subspecies—the human body louse (*Pediculus humanus humanus L.*) and the head louse (*P. h. capitis DG.*)—are

Reserved for Publication Footnotes

closely related obligate parasites that feed exclusively on human blood. Body lice likely evolved from head louse ancestors as humans began to wear clothing, which is required for egg deposition in body lice (4).

P. h. humanus has been of tremendous medical and social importance throughout human history. Of the two forms, only the body louse has been implicated as a vector of human disease and is the principal vector of epidemic typhus (*Rickettsia prowazekii*), relapsing fever (*Borrelia recurrentis*), and trench fever (*Bartonella quintana*) (5–9). In the United States, as in the rest of the world, body lice are primarily a concern in transient homeless populations, whereas head lice tend to infest populations of elementary school-age children. Historically, epidemic typhus has been responsible for massive mortality in wartime (9); in contemporary times, major outbreaks of epidemic typhus are found primarily among refugees, e.g., in Burundi in 1996 (8), but sporadic cases have also been observed in general populations in Russia (10), Peru, Algeria, and France (11).

Like all hematophagous lice, body lice depend on obligate endosymbionts to supplement their nutritionally deficient blood diet (12). The primary endosymbiont of *P. h. humanus* has been given the provisional name *Candidatus Riesia pediculicola* (13) (hereafter Riesia). The body louse maintains organs, mycetomes, that house the primary endosymbiont, except during passage to the ovaries for transovarial transmission (14). The tripartite interdependency of this bacterial endosymbiont, its body louse host, and the human host of the body louse appears to have coevolved over several million years (15).

Here, we present the genome sequence of the body louse and its coevolved primary endosymbiont. This genome, the smallest known insect genome, encodes a remarkably complete gene repertoire and thus provides a robust phylogenetic outgroup for understanding the evolution of holometabolous insects. The striking reduction in genome size is particularly notable in gene families associated with environmental sensing and response, as befits a monophagous permanent parasite with a substantially reduced need to seek out food sources and detect and avoid enemies relative to free-living species.

Results and Discussion

Genome Features.

Genome sequencing, assembly and annotation

The genome of the body louse was sequenced to 8.5x average coverage using a whole genome shotgun approach, with 1.3 million paired-end reads from plasmid libraries. The assembled contigs and scaffolds, spanning 108 Mb and 110 Mb respectively, confirmed previous estimates based on flow cytometry data (103–109 Mb) that the body louse has the smallest known genome size of any insect (16, 17). The 300 longest scaffolds span more than 95% of the assembled genome sequence (scaffold N50 size of 488 kb). A range of automated and manual methods (18) yielded 10 tentative superscaffolds of up to 9 Mb each, spanning a total of 49 Mb. This effort provided large chromosomal segments that were close to continuous, with only a few remaining clone gaps, usually involving simple-sequence gene deserts.

The remarkable compactness of the genome greatly facilitated accurate gene annotation. Predictions using multiple gene modeling approaches resulted in consensus annotation (Table 1) of 10,773 protein-coding genes, 161 tRNAs for all 20 amino acids and 57 microRNAs (Table S1). Comparing predicted gene lengths with their Drosophila melanogaster orthologs (the best experimentally studied insect that drives comparative gene annotation) revealed greater consistency with body louse genes (concordance 0.91; identical with Anopheles gambiae) than with the honey bee, Apis mellifera (concordance 0.89), or the red flour beetle, Tribolium castaneum (concordance 0.88), despite greater evolutionary divergence (Fig. S1).

GC content

Compared to other sequenced insect genomes, the body louse genome has the highest abundance of small homogeneous GC-content domains (7–30 Kb with GC-content between 18 and 63%). The average GC-content of the *P. h. humanus* genome is 28%, similar to that of the *A. mellifera* genome (33%), making these two genomes unusually AT-rich. However, the *A. mellifera* genome harbors more extremes, with only 77% of homogeneous domains having a GC-content between 20 and 60% in *A. mellifera*, compared to 94% in *P. h. humanus*, which is more similar in this respect to the genome of *T. castaneum* (99%) (Fig. S2).

Telomeres

Unlike *A. mellifera* telomeres (19), none of the body louse telomeres appeared to be assembled completely at the ends of long superscaffolds. Therefore, we sought candidate telomere sequences using the strategy employed for *T. castaneum* (20). The body louse is diploid and has a haploid complement of five metacentric chromosomes and one telocentric chromosome, for a total of 11 putative telomeres (21). Although we were unable to reconstruct an entire telomere because of its highly repetitive nature, we identified a long subtelomeric repeat region that was partially assembled on at least nine of the 11 putative telomeres between unique flanking DNA and telomeric TTAGG repeats. This subtelomeric region consists of various satellite-like repeats in addition to pseudogenes and simple sequences and varies considerably in length. The TTAGG repeats commonly contain SART-like retrotransposons, which are also characteristic of the telomeres from *T. castaneum* and *Bombyx mori* (domestic silkworm). This combination might represent the basal insect situation. If so, the simple TTAGG telomeres of *A. mellifera* would represent an evolved condition in which most retrotransposons have been lost, rather than the ancestral condition (19). Alternatively, insect telomeres may have repeatedly been invaded as a "safe harbor" by non-LTR retrotransposons of the R-element family that belong to the SART group (20).

Transposable elements

Both class I and class II mobile elements are present in the genome of *P. h. humanus*, yet they represent only 1% of the genome (Table S2), which is markedly lower than any sequenced insect genome. Interestingly, the body louse genome size is near the hypothesized 100 MB critical threshold at which transposable elements can be established in eukaryote genomes (22).

Mitochondrial genome

The mitochondrial genome of *P. h. humanus* contains the full complement of 37 genes organized in an unusual architecture of 18 mini-circular chromosomes (23). It is possible that multiple mini-circular chromosomes promote recombination between genes on different chromosomes. Indeed, there is evidence in the genome sequence data for at least two novel

mini-circular chromosomes that have arisen from such recombination (Fig. S3).

Of 305 mitochondrial-targeted, nuclear-encoded genes known in *D. melanogaster*, 282 have louse orthologs. This finding suggests that the basic mitochondrial functions (e.g., OXPHOS, membrane transport and protein synthesis) are unimpeded by the reorganized mitochondrial genome. The body louse genome revealed the apparent loss of the mitochondrial single-stranded binding protein (mtSSB), a factor required for optimal initiation and processivity during mitochondrial genome replication in both insects and mammals (24, 25). In the absence of mtSSB, complete replication of a full-sized mitochondrial genome may not be possible (25), as loss of mtSSB function in *D. melanogaster* is lethal at the late-third instar/pupal stages due to a loss of mtDNA content (26). It is not yet known if the mtSSB function can be replaced by an endosymbiont homolog, or the multiple minicircles render the mtSSB unnecessary.

Endosymbiont Genome.

Genome sequencing, assembly and annotation

Like many other sucking lice (*Anoplura*, *Rhyncophthirina*), the body louse has mycetomes that harbor the primary endosymbiotic bacteria (p-endosymbionts). The genome of the *Pediculus* symbiont, *Riesia*, was sequenced to an average coverage of 50x and is composed of a single linear chromosome of at least 574,526 bp with palindromic termini and a single circular plasmid of 7,628 bp. The chromosome contains 557 open reading frames (ORFs) and 33 transfer RNAs, six ribosomal RNAs and one other structural RNA.

Comparisons with other endosymbionts

We compared the genome of *Riesia* with the genomes of other endosymbionts and the infectious plague pathogen *Yersinia pestis* (Fig S4A, B, C, D & E). This genome-wide sequence comparison revealed a core of 237 genes common to all bacteria examined, with only 24 genes unique to *Riesia* and 30 genes present in all except *Riesia* (Table S3A & B). Several genes unique to *Riesia* are transport and binding proteins, along with enzymes involved in lipopolysaccharide biosynthesis. Conversely, the enzymes missing from *Riesia* are mainly exonucleases, which are required for conjugation, and enzymes involved in energy metabolism. The Riesia-specific transport and binding proteins and the lack of energy metabolism genes may reflect the dependence of *Riesia* on its louse host for nutrients. Lipopolysaccharides might be important for cell wall stability when *Riesia* migrate extracellularly through the louse to reach filial mycetomes in the ovaries (14) (Table S3B).

Riesia is required by lice for the production of pantothenic acid (vitamin B5). Without *Riesia*, nymphs die during their first molt (27). Surprisingly, the genes for three key enzymes in the synthesis of pantothenic acid, panB, panC and panE, are missing from the linear chromosome of *Riesia*. These genes are instead found together on the plasmid, an arrangement not found in other bacteria. Similar cases are known from evolutionarily more ancient endosymbionts, e.g. Buchnera, in which essential genes are also found on a plasmid (28). This arrangement could represent a mechanism that reduces the risk of genome degradation and consequently synthesis of pantothenic acid at required levels. Interestingly, there is preliminary evidence that endosymbiont replacement may be commonplace in sucking lice (29), possibly facilitated by the acquisition of plasmids that harbor genes essential to the host. It is also possible that having the genes on a multi-copy plasmid increases expression levels.

Nakabachi et al. (30) proposed that integration of essential genes from the p-endosymbiont into the host genome might be an important mechanism for the host to overcome the consequences of genome degradation of its endosymbiont. *Riesia* and the human, like the pea aphid (*Acyrthosiphon pisum*), represent a system where the genomes of both symbiotic partners are available to test this hypothesis. The body louse genome does not appear to contain any genes of prokaryotic origin, suggesting absence of transfers from *Riesia* to lice. The fact that *Riesia* is in an extracellular environment inside the mycetome, the stomach disc, might contribute to the lack of transfer.

The dramatic reduction in genome size and high AT bias suggest a long association between *Riesia* and its host insect, and, like some other ancient gammaproteobacterial symbiotic associations, the *Riesia* genome is free of mobile elements. Yet Riesia's association with its host is only 13–25 million years old, making *Riesia* one of the youngest endosymbionts known (31).

Comparative Genomics.

The hemimetabolous genome

Hemimetabolous genome is therefore an important outgroup reference for comparative analyses of sequenced holometabolous insects (Fig. 1A). The complete metamorphosis of holometabolous insects is a highly successful evolutionary strategy, whereby larvae and adults can take adv outgroup: The human body louse is among the first sequenced representatives of hemimetabolous insects (32), a group distinguished by progressive intermediate development as nymphal instars rather than larva-pupa-adult transformations. The louse antage of different ecological niches. The molecular innovations that have contributed to the success of holometabolous insects can now be viewed in the context of a hemimetabolous outgroup genome sequence that is largely complete.

In addition to being the smallest genome of any insect studied to date, the body louse genome is, as far as can be determined, functionally complete. Of the 10,773 body louse protein-coding genes, 90% share homology to genes known in other species, enabling orthology delineation for 80% of louse genes (33). This level is comparable to results from initial analyses from *A. mellifera* (34) and *T. castaneum* (20). The phylogenetic tree reconstructed using single-copy orthologs (Fig. 1A) confirms the basal position of Hemimetabola compared to Holometabola within Insecta and suggests an average rate of molecular evolution in the lineage of lice comparable to that of Hymenoptera and Coleoptera.

Micro-synteny analysis (35) between genomes of the body louse and the hymenopteran honey bee *A. mellifera* or *Nasonia* parasitoid wasp species suggest that about 20% of single-copy orthologs are retained in their ancestral arrangements (Table S4). This percentage is similar to micro-synteny conservation levels between *A. mellifera* (Hymenoptera) and *T. castaneum* (Coleoptera) and substantially greater than their conservation with dipterans ($< 15\%$) (36), highlighting the derived state of Diptera.

Ancestral insect gene repertoire

Contrary to the expectations of reductive evolution common in obligate parasites, the body louse has retained a remarkably complete repertoire of both protein-coding and non-

protein-coding genes (Table 1). The distribution of orthologous genes across four representative insect species (Fig. 1B & C) shows that Hymenoptera and Coleoptera share more orthologs with the body louse than they do with the fruit fly *D. melanogaster*. Relative to the well-studied *D. melanogaster* model, the louse genome may be utilized as a robust outgroup to Holometabola.

Examining microRNA gene families shared among crustaceans and insects revealed that mir-315, mir-283, mir-33, and mir-29 were lost from the body louse genome (Table S1, Fig S5; mir-iab-4 and mir-46 have been found in the trace archive). As all true lice are wingless, it is intriguing to note that mir-315 has been identified as a potent activator of Wingless signaling in *D. melanogaster* (37).

Evolution of gene families in relation to the life history of the body louse. The body louse has maintained many genes important for basic physiological processes, losing only a few of these roles to its endosymbiont *Riesia*. As the expansion and contraction of gene families may indicate functional adaptation and evolution, we compared the body louse gene repertoire with those of the honey bee and the red flour beetle. Comparisons were made both at the level of protein families, which could be generally defined using InterPro domain signatures (Table S5A, B & C), and at a finer scale at the level of orthologous groups of genes (Fig. S6). On both scales, the body louse genome appears to have several gene families with fewer members than those found in other invertebrates.

Fewer genes are associated with environmental sensing and response.

G protein-coupled receptors (GPCRs)

With 104 non-sensory GPCRs and three opsins (visual receptors; Table S6A,B), *P. h. humanus* has the smallest repertoire of GPCRs identified in any sequenced insect genome to date (20, 34, 38–40). The louse genome has orthologs for approximately 80% of non-sensory GPCRs identified in *D. melanogaster*. These GPCRs seemingly represent a minimal suite of receptors needed to maintain conserved GPCR-mediated signaling pathways common to diverse insect taxa (41). The relatively small number of louse opsins likely reflects its simple visual system. Moreover, the body louse lacks a putative long-wavelength sensitive opsin typically found in other insects (42), a feature that might have evolved during its adaptation to the obligate parasitic lifestyle.

Odorant, gustatory and chemosensory related genes

The genome sequence revealed just 10 odorant receptor (Or) genes, fewer than any other insects examined to date by almost an order of magnitude. The gustatory receptor (Gr) family is comparably small, with just six loci encoding eight proteins through alternative splicing of the N-terminus of one locus. There are no orthologs of the otherwise highly conserved carbon dioxide heterodimer Gr receptors (39, 40, 43, 44) or the putative sugar receptors (45, 46). *P. h. humanus* contains respectively five and seven putative functional odorant-binding proteins (OBPs) and chemosensory proteins (CSPs) (Table S7), which is dramatically fewer than those found in other insects (47). These aforementioned sensory genes and their resultant proteins are presumably not necessary for host location and selection. Furthermore, lice do not need to avoid the many bitter xenobiotic toxins to which most insect Grs appear to be tuned (45).

Insulin/TOR pathway genes

The insulin/TOR signal transduction pathway plays a central role in multiple and critical biological processes, including organismal growth, anabolic metabolism, cell survival, fertility, and lifespan determination (48, 49). This pathway has been well characterized in multiple organisms, including *D. melanogaster* (50). Both the structure of the pathway and the molecular function of its components are well conserved across metazoans. The body louse genome encodes a complete insulin/TOR signaling pathway. However, these genes are reduced in number in the body louse in contrast with *D. melanogaster*, which possesses multiple copies (Table S8). Remarkably, the louse has a single insulin-like peptide (ilp) gene. Given that there is some evidence for differential expression of ilp genes under different dietary conditions in insects (51, 52), the presence of a single ilp gene in the body louse genome might reflect its restricted and homogeneous diet.

Detoxification enzymes

The louse genome encodes the smallest number of detoxification enzymes observed in any insect, reflecting its obligate parasite lifestyle in which it is sheltered from xenobiotic challenges faced by free-living insects (e.g., plant secondary compounds). There are notably few cytochrome P450s, with only 12 genes within the CYP3 clade that is most closely associated with xenobiotic metabolism. In contrast, *D. melanogaster* and *A. mellifera* have 36 and 28 CYP3 clade genes, respectively. Among the 13 glutathione-S-transferases (GST)(Table S9), none belong to the Epsilon class that has been demonstrated to contribute to insect adaptation to environmental selection pressures (53). The Epsilon class was also missing in the pea aphid genome. In contrast, the relative abundance of Delta class GSTs (more than *A. mellifera*) suggests that *P. h. humanus* still possesses some capacity for detoxification of xenobiotics, including insecticides (54).

Body louse coevolution and allopatric speciation. With their characteristic extreme host specificity, pediculid lice provide dramatic examples of host-parasite coevolution and allopatric speciation (55). One consequence of this specificity is the difficulty encountered when adapting human lice to experimental animals (8). Body lice have reduced genomes and harbor specific bacterial symbionts and pathogens that also exhibit genome reduction (56–63). These combined observations support the hypothesis that the body louse is an example of a unique and extreme form of allopatric speciation that hosts bacteria exhibiting the same level of speciation. Such extreme specializations, associated with dramatic genome reductions, may have resulted from a lack of gene exchange following allopatric speciation. This association of an insect host, its symbionts and its bacterial pathogens coevolving and showing congruent reductive genome evolution provides a dramatic example of the evolutionary consequences of genome interactions and interdependency over time.

Conclusions

The body louse genome provides a unique repository of data of considerable basic and practical significance. The availability of sequence data will facilitate molecular studies of a vector for diseases that continue to afflict human populations around the world. The louse relies on *Riesia*, an obligatory louse bacterial endosymbiont that lacks antibiotic resistance genes for survival; the development of novel louse control strategies targeting this symbiont may thus be possible. With respect to understanding the evolution of multigene families mediating

responses to environmental selective forces, the body louse genome, with its drastically reduced inventories in the context of its exceptionally homogeneous environment, provides extraordinary prospects for characterizing the functionalities of these rapidly evolving proteins. As well, further studies focusing on the smaller repertoire of detoxification genes and olfactory receptors in the body louse may guide the development of novel pediculicides and repellents with negligible impacts on human hosts. Moreover, the remarkable completeness of this genome, despite its small size, will serve as a key evolutionary reference point for studies of all sequenced insect species in characterizing the fundamental prerequisites for insect growth and development. Finally, the body louse genome will provide an opportunity for the scientific community to gain greater insights into host-parasite-symbiont tripartite coevolution and speciation.

Materials and Methods

Genome sequencing, assembly and annotation. Lice were obtained from an inbred colony derived from the Culpepper strain (64), which has been maintained on rabbits since 1999 at the University of Massachusetts, Amherst. Total DNA was extracted from approximately 100 first instar nymphs before their first blood meal and was used to construct libraries in the plasmid, pHOS2 (3–4 kb and 10–12 kb inserts), or the fosmid, pCCFOS1 (35–40 kb inserts). End sequencing of clones from each library was conducted using a standard capillary platform (ABI 3730), and yielded 1.30 million good traces (96% paired) with a mean clear read length of 656 bases. All traces were deposited in the NCBI trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?). The reads were assembled with Celera Assembler (http://wgs-assembler.sourceforge.net) (38, 65, 66) and deposited with NCBI (Accession AAZO00000000). The details of the assembly and annotation are given in SI. Additional analyses of other aspects of the body louse genome are given in the SI.

ACKNOWLEDGMENTS. Details of funding support can be found in the supplemental information Excel document which also outlines the details the contributions of each of the co-authors.

References

1. Reed DL, Light JE, Allen JM, Kirchman JJ (2007) Pair of lice lost or parasites regained: the evolutionary history of anthropoid primate lice. *BMC Biol.* 5:7.
2. Reed DL, Smith VS, Hammond SL, Rogers AR, Clayton DH (2004) Genetic analysis of lice supports direct contact between modern and archaic humans. *PLoS Biol.* 2:e340.
3. Light JE, Reed DL (2009) Multigene analysis of phylogenetic relationships and divergence times of primate sucking lice (Phthiraptera: Anoplura). *Mol. Phylogenet. Evol.* 50:376–390.
4. Kittler R, Kayser M, Stoneking M (2003) Molecular evolution of Pediculus humanus and the origin of clothing. *Curr. Biol.* 13:1414–1417.
5. Eremeeva ME, Madan A, Shaw CD, Tang K, Dasch GA (2005) New perspectives on rickettsial evolution from new genome sequences of rickettsia, particularly R. canadensis, and Orientia tsutsugamushi. *Ann. N. Y. Acad. Sci.* 1063:47–63.
6. Rotz LD, Khan AS, Lillibridge SR, Ostroff SM, Hughes JM (2002) Public health assessment of potential biological terrorism agents. *Emerging Infect. Dis.* 8:225–230.
7. Andersson JO, Andersson SG, Nicolle C (2000) A century of typhus, lice and Rickettsia. *Res. Microbiol.* 151:143–150.
8. Raoult D, Roux V (1999) The body louse as a vector of reemerging human diseases. *Clin. Infect. Dis.* 29:888–911.
9. Raoult D, et al. (2006) Evidence for louse-transmitted diseases in soldiers of Napoleon's Grand Army in Vilnius. *J. Infect. Dis.* 193:112–120.
10. Tarasevich I, Rydkina E, Raoult D (1998) Outbreak of epidemic typhus in Russia. *Lancet* 352:1151.
11. Bechah Y, Capo C, Mege JL, Raoult D (2008) Epidemic typhus. *Lancet Infect Dis* 8:417–426.
12. Buchner P (1965) *Endosymbiosis of Animals with Plant Microorganisms* (Interscience Publishers, New York), p 909.
13. Sasaki-Fukatsu K, et al. (2006) Symbiotic bacteria associated with stomach discs of human lice. *Appl. Environ. Microbiol.* 72:7349–7352.
14. Perotti MA, Allen JM, Reed DL, Braig HR (2007) Host-symbiont interactions of the primary endosymbiont of human head and body lice. *FASEB J.* 21:1058–1066.
15. Allen JM, Reed DL, Perotti MA, Braig HR (2007) Evolutionary relationships of "Candidatus Riesia spp.," endosymbiotic enterobacteriaceae living within hematophagous primate lice. *Appl. Environ. Microbiol.* 73:1659–1664.
16. Pittendrigh BR, et al. (2006) Sequencing of a new target genome: the Pediculus humanus humanus (Phthiraptera: Pediculidae) genome project. *J. Med. Entomol.* 43:1103–1111.
17. Johnston JS, Yoon KS, Strycharz JP, Pittendrigh BR, Clark JM (2007) Body lice and head lice (Anoplura: Pediculidae) have the smallest genomes of any hemimetabolous insect reported to date. *J. Med. Entomol.* 44:1009–1012.
18. Robertson HM, et al. (2007) Manual superscaffolding of honey bee (Apis mellifera) chromosomes 12-16: implications for the draft genome assembly version 4, gene annotation, and chromosome structure. *Insect Mol. Biol.* 16:401–410.
19. Robertson HM, Gordon KH (2006) Canonical TTAGG-repeat telomeres and telomerase in the honey bee, Apis mellifera. *Genome Res.* 16:1345–1351.
20. Richards S, et al. (2008) The genome of the model beetle and pest Tribolium castaneum. *Nature* 452:949–955.
21. Hindle E, Pontecorvo G (1942) Mitotic divisions following meiosis in Pediculus corporis males. *Nature* 149:668.
22. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302:1401–1404.
23. Shao R, Kirkness EF, Barker SC (2009) The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, Pediculus humanus. *Genome Res.* 19:904–912.
24. Farr CL, Matsushima Y, Lagina AT, Luo N, Kaguni LS (2004) Physiological and biochemical defects in functional interactions of mitochondrial DNA polymerase and DNA-binding mutants of single-stranded DNA-binding protein. *J. Biol. Chem.* 279:17047–17053.
25. Korhonen JA, Pham XH, Pellegrini M, Falkenberg M (2004) Reconstitution of a minimal mtDNA replisome in vitro. *EMBO J.* 23:2423–2429.
26. Maier D, et al. (2001) Mitochondrial single-stranded DNA-binding protein is required for mitochondrial DNA replication and development in Drosophila melanogaster. *Mol. Biol. Cell* 12:821–830.
27. Perotti MA, Kirkness EF, Reed DL, Braig HR (2009) in *Insect Symbiosis 3*, eds Bourtzis KM, Taylor F, Boca R pp 205–220.
28. Ding H, Hynes MF (2009) Plasmid transfer systems in the rhizobia. *Can. J. Microbiol.* 55:917–927.
29. Hypsa V, Krízek J (2007) Molecular evidence for polyphyletic origin of the primary symbionts of sucking lice (phthiraptera, anoplura). *Microb. Ecol.* 54:242–251.
30. Nakabachi A, et al. (2006) The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science* 314:267.

31. Allen JM, Light JE, Perotti MA, Braig HR, Reed DL (2009) Mutational meltdown in primary endosymbionts: selection limits Muller's ratchet. *PLoS ONE* 4:e4969.
32. Richards S, et al. (2010) Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS Biol.* 8:e1000313.
33. Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* 36:D271–275.
34. Weinstock GM, et al. (2006) Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* 443:931–949.
35. Zdobnov EM, et al. (2002) Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science* 298:149–159.
36. Zdobnov EM, Bork P (2007) Quantification of insect genome divergence. *Trends Genet.* 23:16–20.
37. Klingensmith J, Nusse R (1994) Signaling by wingless in Drosophila. *Dev. Biol.* 166:396–414.
38. Myers EW, et al. (2000) A whole-genome assembly of Drosophila. *Science* 287:2196–2204.
39. Robertson HM, Warr CG, Carlson JR (2003) Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U.S.A.* 100 Suppl 2:14537–14542.
40. Benton R, Sachse S, Michnick SW, Vosshall LB (2006) Atypical membrane topology and heteromeric function of Drosophila odorant receptors in vivo. *PLoS Biol.* 4:e20.
41. Wistrand M, Käll L, Sonnhammer EL (2006) A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Sci.* 15:509–521.
42. Briscoe AD, Chittka L (2001) The evolution of color vision in insects. *Annu. Rev. Entomol.* 46:471–510.
43. Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB (2007) Two chemosensory receptors together mediate carbon dioxide detection in Drosophila. *Nature* 445:86–90.
44. Frommer WB (2010) Biochemistry. CO2mmon sense. *Science* 327:275–276.
45. Marella S, et al. (2006) Imaging taste responses in the fly brain reveals a functional map of taste category and behavior. *Neuron* 49:285–295.
46. Chyb S, Dahanukar A, Wickens A, Carlson JR (2003) Drosophila Gr5a encodes a taste receptor tuned to trehalose. *Proc. Natl. Acad. Sci. U.S.A.* 100 Suppl 2:14526–14530.
47. Sánchez-Gracia A, Vieira FG, Rozas J (2009) Molecular evolution of the major chemosensory gene families in insects. *Heredity* 103:208–216.
48. Goberdhan DC, Wilson C (2003) The functions of insulin signaling: size isn't everything, even in Drosophila. *Differentiation* 71:375–397.
49. Oldham S, Hafen E (2003) Insulin/IGF and target of rapamycin signaling: a TOR de force in growth control. *Trends Cell Biol.* 13:79–85.
50. Alvarez-Ponce D, Aguadé M, Rozas J (2009) Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. *Genome Res.* 19:234–242.
51. Wheeler DE, Buck N, Evans JD (2006) Expression of insulin pathway genes during the period of caste determination in the honey bee, Apis mellifera. *Insect Mol. Biol.* 15:597–602.
52. Arsic D, Guerin PM (2008) Nutrient content of diet affects the signaling activity of the insulin/target of rapamycin/p70 S6 kinase pathway in the African malaria mosquito Anopheles gambiae. *J. Insect Physiol.* 54:1226–1235.
53. Ranson H, et al. (2002) Evolution of supergene families associated with insecticide resistance. *Science* 298:179–181.
54. Enayati AA, Ranson H, Hemingway J (2005) Insect glutathione transferases and insecticide resistance. *Insect Mol. Biol.* 14:3–8.
55. Page RD, Lee PL, Becher SA, Griffiths R, Clayton DH (1998) A different tempo of mitochondrial DNA evolution in birds and their parasitic lice. *Mol. Phylogenet. Evol.* 9:276–293.
56. Blanc G, et al. (2007) Reductive genome evolution from the mother of Rickettsia. *PLoS Genet.* 3:e14.
57. Andersson SG, et al. (1998) The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* 396:133–140.
58. Ogata H, et al. (2001) Mechanisms of evolution in Rickettsia conorii and R. prowazekii. *Science* 293:2093–2098.
59. Lescot M, et al. (2008) The genome of Borrelia recurrentis, the agent of deadly louse-borne relapsing fever, is a degraded subset of tick-borne Borrelia duttonii. *PLoS Genet.* 4:e1000185.
60. Alsmark CM, et al. (2004) The louse-borne human pathogen Bartonella quintana is a genomic derivative of the zoonotic agent Bartonella henselae. *Proc. Natl. Acad. Sci. U.S.A.* 101:9716–9721.
61. Fournier PE, Suhre K, Fournous G, Raoult D (2006) Estimation of prokaryote genomic DNA G+C content by sequencing universally conserved genes. *Int. J. Syst. Evol. Microbiol.* 56:1025–1029.
62. Fournier PE, et al. (2006) Comparative genomics of multidrug resistance in Acinetobacter baumannii. *PLoS Genet.* 2:e7.
63. Vallenet D, et al. (2008) Comparative analysis of Acinetobacters: three genomes for three lifestyles. *PLoS ONE* 3:e1805.
64. Culpepper GH (1944) The rearing and maintenance of a laboratory colony of the body louse. *Am J Trop Med Hyg* 24:327–329.
65. Levy S, et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol.* 5:e254.
66. Venter JC, et al. (2001) The sequence of the human genome. *Science* 291:1304–1351.

Table 1: Summary of the genome features of *Pediculus humanus humanus* as compared to *Drosophila melanogaster*[1]. The more numerous body louse exons and introns suggest intron loss in *D. melanogaster*, yet with an increase of their sizes.

Genome feature	Count	Nucleotides (Mb)	Genome Fraction (%)
P. h. humanus	6 chromosomes	111	100
(*D. melanogaster*)	(4 chromosomes)	(169)	(100)
Gene rich clusters (containing 95% of genes[2])	1,110	55	50
	(1,130)	(70)	(41)
Protein-Coding Genes			
Total [Multi-Exon]	10,773 [10,424]	33.8	31
	(13,794 [11,458])	(82.6)	(49)
Coding Exons	69,261	16.6	15
	(54,606)	(22.3)	(13)
Introns	58,522	17.2	15
	(44,698)	(48.6)	(29)
Non-Protein-Coding Genes			
tRNAs	161	0.012	<1
	(292)	(0.022)	<1
miRNAs	57	0.005	<1
	(90)	(0.008)	<1
Transposable Elements	3,558	1.1	1
	(9,409)	(11.6)	(7)
Tandem Repeats[3]	130,608	6.9	6
	(25,904)	(6.1)	(4)

[1] D. melanogaster values obtained from FlyBase release 5.23 using the same parameters used to obtain, parse, and count the P. h. humanus genome.
[2&3] Supporting documentation is given in Figures S6 and S7.

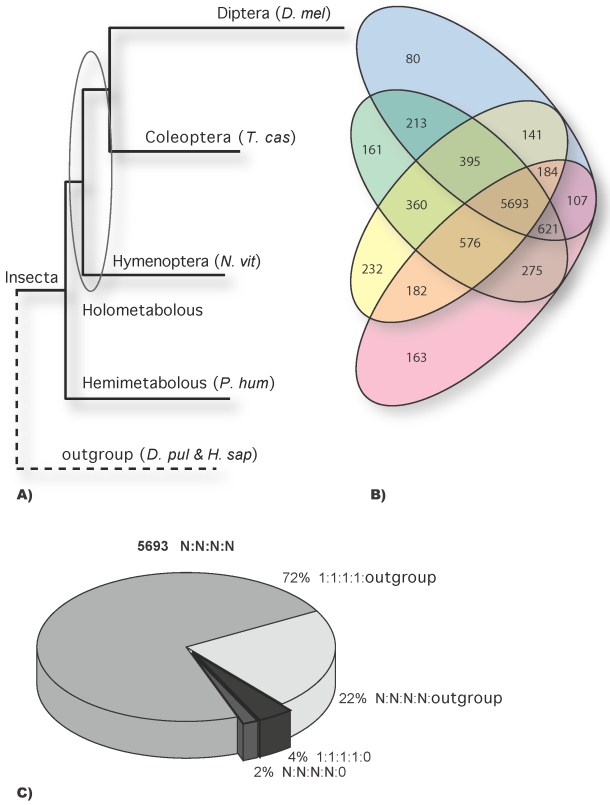


Fig. 1: The *Pediculus humanus humanus* (P. hum) genome reveals a basal insect gene repertoire. The encoded P. hum proteome is compared to sequenced representatives of the orders Diptera, Coleoptera and Hymenoptera, and outgroup species beyond Insecta. D. mel, *Drosophila melanogaster*; T. cas, *Tribolium castaneum*; N. vit, *Nasonia vitripennis*; D. pul, *Daphnia pulex*; H. sap, *Homo sapiens*. (A) The Maximum-Likelihood phylogenetic tree was reconstructed using the superalignment of protein sequences of universal single-copy orthologs. The obtained tree confirms the basal position of Hemimetabola compared to Holometabola within Insecta. The branch lengths are proportional to the accumulated number of substitutions suggesting an average rate of molecular evolution in lice that is comparable to those in Hymenoptera and Coleoptera. (B) The Venn diagram shows the numbers of orthologous groups of genes shared among the four insects (a lower estimate of the ancestral number of genes). It depicts the phylogenetic distribution of orthologs, highlighting the completeness of the gene repertoire encoded in the body louse genome, which does not exhibit excessive losses. P. hum, pink; N. vit, yellow; T. cas green; D. mel, blue. (C) The pie-chart partitions the largest fraction of core body louse proteins with orthologs in three holometabolous insect orders and the outgroup species beyond Insecta with respect to single (1:1:1:1) and multiple (N:N:N:N) -copy orthologs. Of the 5693 groups of single and multiple-copy orthologs common across Insecta, 94% are shared across Bilateria as single-copy (72%) or multi-copy (22%) orthologs, and only 6% appear as insect-specific orthologous groups (4% as single copies and 2% as multi-copy).

A.1.5.3 *Supplementary figures*

An extract of the supplementary material for the paper "Genome Sequences of the Human Body Louse and its Primary Endosymbiont: Insights into the Permanent Parasitic Lifestyle" covering my work is presented on pages 203–207.

distributions show the fraction of genes (thick lines) or of the entire genome (thin lines) occurring in GC-content domains < X %GC.

Figure S3 A model of non-homologous end-joining (NHEJ) between mitochondrial minichromosomes that generated chimeric mitochondrial chromosomes in *Pediculus. h. humanus*. Coding regions of minichromosomes are in yellow and green and non-coding regions are in blue. Black arrows in coding regions indicate the orientation of gene transcription. Broken lines indicate sites of double-strand breakages where the two minichromosomes that recombine share homologous sequences. Of the 37,144 sequence reads that contained mitochondrial genes, a small number (1.5%) aligned only partially with the 18 abundant mini-circular chromosomes. Almost all (98%) of these 529 reads could be assembled into two novel chromosomes, each a chimeric derivative of two known chromosomes that appear to have recombined by NHEJ via a common "microhomologous" sequence of 12 bp (A) or 19 bp (B). The protein-coding genes of the chimeric chromosomes have only fragments of the full-length cox2, cox3, nad1 and atp6 genes. However, the two tRNA genes, trnA and trnY, were the same length as their counterparts in the known minichromosomes, and therefore potentially functional. Interestingly, the genic regions of all mitochondrial chromosomes have a common upstream motif (CAAAYCTCAACTCGTTTCAT), and all except one have the same orientation relative to the conserved noncoding region (23). The exceptional chromosome (encoding nad1) shares a 56 bp segment with rrnL that may have arisen from a similar NHEJ event between the ancestral nad1 and rrnL minichromsomes.(C)

Figure S4 A genome-wide comparison of *Candidatus* Riesia pediculicola (Rp) with the primary endosymbionts, *Wigglesworthia glossinida* (Wg, tsetste flies), *Blochmannia floridanus* (not shown), *B. pennsylvanicus* (Bp, carpenter ants), the automonous *Buchnera aphidicola* (Ba-ap, aphids) strains APS and BBp, Sg (not shown), *Baumannia cicadellinicola* (Bc, leafhoppers, sharpshooters) and the pathogens *Photorhabdus luminescens* subsp. laumondii TTO (Pf) and *Yersinia pestis* str. CO92 (Yp), revealed a core of 237 genes in all aforementioned bacteria, with only 27 genes unique to Riesia (Table S3B) and 30 genes present in all bacteria except Riesia (Table S3A). In this comparison, Riesia shares the most orthologues with *P. luminescens*.

Figure S5. Multiple alignment of microRNA genes well represented in insect genomes and found in at least few more basal lineages (shown in bold) that we failed to identify in both the assembled body louse genome and raw sequencing reads, suggesting evolutionary loss of these genes (A) miR-29, (B) miR-33, (C) miR-283, (D) miR-315. The species as are follows: *Drosophila melanogaster* (Dmel), *D. simulans* (Dsim), *D.*

sechellia (Dsec), *D. yakuba* (Dyak), *D. erecta* (Dere), *D. ananassae* (Dana), *D. pseudoobscura* (Dpse), *D. persimilis* (Dpse), *D. willistoni* (Dwil), *D. mojavensis* (Dmoj), *D. virilis* (Dvir), *D. grimshawi* (Dgri), *Culex pipiens quinquefasciatus* (Cqui), *Aedes aegypti* (Aaeg), *Anopheles gambiae* (Agam), *Bombyx mori* (Bmor), *Tribolium castaneum* (Tcas), *Nasonia vitripennis* (Nvit), *Apis mellifera (Apis), Daphnia pulex* (*Dpul*), and *Ixodes scapularis* (Isca).

Figure S6 Orthologous Group Expansions The *Pediculus humanus* (*Phum*) proteome was compared to the insects *Drosophila melanogaster* (*Dmel*), *Tribolium castaneum* (*Tcas*), *Nasonia vitripennis* (Nvit), with the outgroup species *Daphnia pulex* (*Dpul*) and *Homo sapiens* (*Hsap*) to delineate groups of orthologous protein-coding genes (shown in Figure 1, main text). Examining 633 expanded groups with members in all four insects reveals the lower number of expansions and the significantly smaller proportions of *Phum* proteins in these expanded orthologous groups. The examined groups were required to have at least one member from each of the four insect species and a minimum of six proteins in total. These expanded groups therefore exhibit a minimally duplication in two species or a triplication in one species. Fewer than half the groups show an expansion in *Phum* (47% >1 member), whereas the other species exhibit more expansions (*Nvit*, 59%; *Tcas*, 70%; *Dmel*, 64%). *Phum* also shows lower mean and median values for the proportions of orthologous group members as shown in the figure boxplots, and paired Wilcoxon signed-rank tests show these differences to be statistically significant.

Figure S7: Gene-rich portions of the *Pediculus humanus* (Louse) and *Drosophila melanogaster* (Fly) genomes. General Feature Format (GFF) files for Louse (VectorBase PhumU1.2) and Fly (FlyBase Dmel5.23) gene sets were interrogated to calculate gene spans and intergenic distances defined by protein-coding gene start and stop codons. The transcript with the longest coding sequence span (CDS) was used for genes with alternative transcripts. Merging of overlapping or intronic genes ensured that each genomic region was only counted once in the sum of genomic spans. The numbers of genes and their total genomic spans (gene plus intergenic) were summed for intergenic thresholds in 200bp steps up to 20Kb

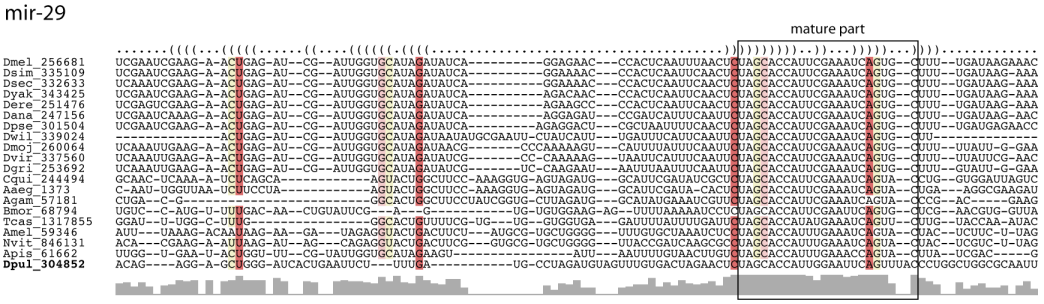


Figure S5A

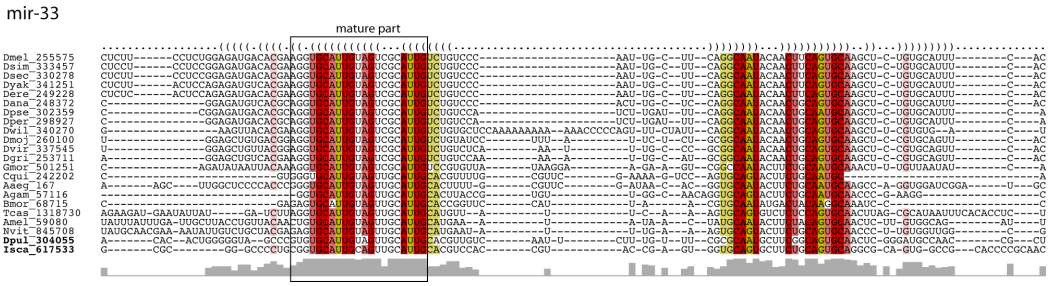


Figure S5B

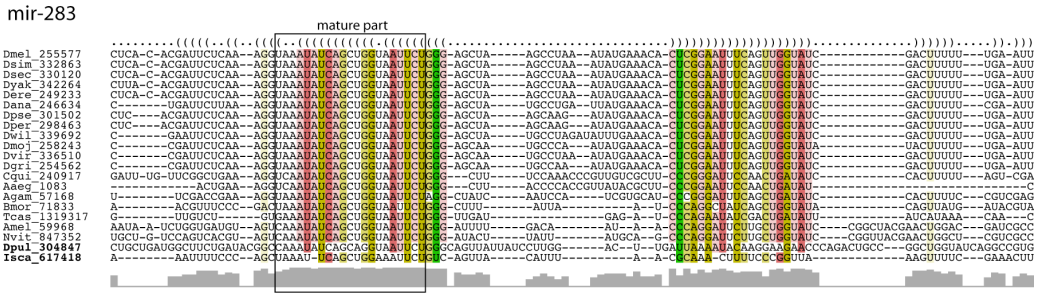


Figure S5C

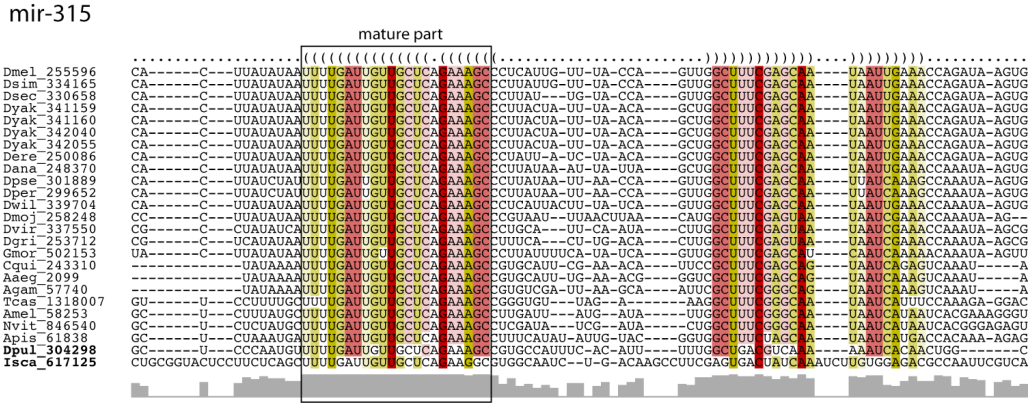


Figure S5D

A.1.5.4 *Supplementary tables*

An extract of the supplementary material for the paper "Genome Sequences of the Human Body Louse and its Primary Endosymbiont: Insights into the Permanent Parasitic Lifestyle" covering my work is presented on pages 209–214.

Supplemental Tables

Table S1 – Predicted microRNAs in the genome of *Pediculus humanus humanus (1).* In summary, 57 pre-microRNA genes, 6 of which encodes 2 mature microRNAs. The mir-281 found in 4 different reads (but not in the genome) is probably only 1 gene as the sequences are identical.

Name	Family	Contig/ trace_file	Start	End	Strand	mature_sequence	mature_start	mature_end
phum-bantam	bantam	AAZO01005717.1	66948	67038	1	AAAAAGGAAAACGAAACUGGUUUUCACAAUGAUUUGCCAGAUAGUUUUUGUUAUUCUGAGAUCAUUGUGAAAGCUGAUUUUGUUAUGGAAC	57	79
phum-let-7	let-7	AAZO01005184.1	23359	23437	1	AGCAGGGUGAGGUAGUAGACUGUAUAGUAAAGAAUUACAUCAUUUGGAGGUACUGUACAAUCUGCUAACUUUCCCUGGU	8	28
phum-mir-1	mir-1	AAZO01007144.1	49271	49355	1	UGCCUCUUACUAGGUCCGUGCUUCCUUACUUCCCAUAUAGUGUUUGUUUUAUGGAAUGUAAAGAAGUAUGGAGCCUGUACGGGCG	52	73
phum-mir-10	mir-10	AAZO01003084.1	338	437	1	GCCAUUUUAUGCUUUACAUCUACCCUGUAGAUCCGAAUUUGUUUGAUAUCAAACGACAAAUUCGGUUCUAGAGAGGUUUGUGUGGGGCAUACAAAUCAUU	22	42
phum-mir-1000	mir-1000	AAZO01004199.1	7978	8077	1	UGGCGACGCCUGAGAUAUUGUCCUGUCACAGCAGUACUGUAUUUGAGAUACUGCUGUAUCGGGACAUAGCCUCAGACGUUCGUUCGUUCUUUCAUACCGA	15	35
phum-mir-12	mir-12	AAZO01007275.1	4670	4769	-1	GAAAAAAAUGACUUUGCUCGUGAGUAUUGCAUCAGGUACUGGUUAUUAUUAUUCUUAUCAGUAUCUGGGUAAAACUCAUGAGCCGGGUUAAAUCAACAGA	21	43
phum-mir-124	mir-124	AAZO01006175.1	169059	169136	-1	UGCCUGCUCCUCGUGUUCACUGUUGGCCUUGAUGUUUUAAUGAUCAUAAGGCACGCGGUGAAUGCCAAGAGCUAGCUC	47	69

phum-mir-125	mir-125	AAZO01005184.1	24423	24510	1	CCCCGGCUCCCCAUUCCCUGAGACCCUAACUUGUGAUGGUUAUGUUAAAUGUCUCACAGGUUGGAUUCUCGGGCCUGGGUGGUCGGGU	15	36
phum-mir-125as	mir-125as	AAZO01005184.1	24431	24518	-1	AUAAUACAACCCGACCACCCAGGCCCGAGAAUCCAACCUGUGAGACAUUUAACAUAACCAUCACAAGUUAGGGUCUCAGGGAAUGGGG	61	82
phum-mir-133	mir-133	AAZO01007144.1	88843	88941	1	GACUAAUGCUCUUCUUUAGCUGGUUGACGCCGGGCCAAGUUUAGCCGAUUCGAGUCAUUUGGUCCCCUUCAACCAGCUGUAGUUGGCAUUUAGAGCAAC	59	80
phum-mir-137	mir-137	AAZO01004444.1	55299	55390	-1	UUUAUAACUUGGUUGGUCACGCGUAUUCUUGGGGAAUUAACACACCAAAAGAGAUGUUAUUGCUUGAGAAUACACGUAGCUGACAAGUUUGA	57	78
phum-mir-14	mir-14	AAZO01007080.1	16499	16583	-1	GUAAAUGUGGUAGGCGAGAGUGAUAGACUUGGGCUGAGAUUUGUUACCAGUCAGUCUUUUUCUCUCUCCUAUUCUUUUUGCCAGU	51	71
phum-mir-184	mir-184	AAZO01006369.1	2406	2506	1	UUUAUUCAUUCUACUACGAUUCCUUGUCAUUCUUCUGUCCGGCGAGCUUUCAUUAAAUGUUGGACGGAGAACUGAUAAGGGCUCGUGGAUGAAAAAAAUCU	16	35
phum-mir-190	mir-190	AAZO01006982.1	10457	10554	-1	AAAAUGAUUUCGUUUCAUCUCUGAGAUAUGUUUGAUAUUCUUGGUUGUUUUAUUUAAUUUCAACCAAAUAUCGAACAUAUUUCAAGGCUGAAAUCCCA	24	47
phum-mir-13a	mir-2	AAZO01007079.1	40320	40391	1	UCUCGGGUCACAAAGUGUAUGUGAAAUGUGGCGUCUUUGAAUUCAUAUCACAGCCACUUUGAUGACCUCGGA	46	67
phum-mir-13b	mir-2	AAZO01007079.1	40967	41034	1	CCACCCGUCAAAUCGGUUGUGAUUUAUGUUUUAUGAAUUUUCAUAUCACAGCCAUAUUUGACAAGUCG	44	66
phum-mir-2a-1	mir-2	AAZO01007079.1	40128	40214	1	AGAGCAAAAUUCUUUCAUUACAGUUGACUGUAAUAGAUAAAUAUUGUAAUUCUAUCACAGCCAGCUUUGAUGAGCAGAACGUUGUUC	53	75
phum-mir-2a-2	mir-2	AAZO01007079.1	41168	41268	1	GAAAACUGUCUAAAGUUUUUCCUCUGCCACUCACAAAGUGGCUGGGGUAUGUUGAAUUCAUAUCACAGCCAGCUUUGAUGAGCGGGACGGGGAAAUAAAUC	61	83

phum-mir-2c	mir-2	AAZO01007079.1	41394	41470	1	GCUCCGCUCACAAAGUGGUUG GUAUAUGUUGUAUUCAUUUGA UUUAUCAUAUCACAGCCAGCUU UGAUGAGUGGGUC	50	72
phum-mir-210	mir-210	AAZO01006118.1	165	221	-1	AGCUGCUGGACACAGCCCAAG AUUAGUUUAAGACUCUUGUGC GUGAACAUCAGCUAG	1	22
phum-mir-219	mir-219	AAZO01005883.1	1045	1123	-1	CAUCAGGCUUUGAUUGUCCAA ACGCAAUUCUUGUUAUUUUAUU CAAUCAAGGACUGUGUGUGGA CAUCAAUGCUUGUCC	11	33
phum-mir-92a	mir-25	AAZO01005488.1	5121	5199	-1	ACCUUUGGAUGGCUCGUGACC GGUGGCAAUAUUUAAAUAUAUU AUAUAAUAUUGCACUUGUCCCG GCCUAUCUGGAGUG	50	70
phum-mir-92b	mir-25	AAZO01005488.1	4970	5052	-1	GCUAAGAGUGGAGGCCGAGUC AUUUGCAAACUAUGCAUUUAAU UAUUGAUCAUAUUGCACUUGU CCCGGCCUGCUUUCUUGCU	54	73
phum-mir-252	mir-252	AAZO01005296.1	52128	52212	-1	AAAUCUAUUUCACUUCUAAGUA CUAGUGCCGCAGGAGUUUAUA UUGAAACCUCCUGCUGCUCGG GUGCUUAUCACUGAAUAGAUU	16	37
phum-mir-263a	mir-263	AAZO01000403.1	2648	2722	1	GUUCCUGACAAUGGCACUGAA AGAAUUCACGGCUGAAUUUGA GGACCGUGGGUCUUUGGUGCC AUCUUCAGUAAC	11	30
phum-mir-263b	mir-263	AAZO01001458.1	4329	4406	1	UUUGUGAAACUUGGCACUGGA AGAAUUCACAGAUAGAAAAUGA AAAAUCGUGGGUCUUUUGGUG CCAAAGUUCACGGA	11	30
phum-mir-275	mir-275	AAZO01007192.1	28763	28833	1	CGCGCUGCUCAGGUCCUUUAG ACUUUUCUUAUGUUACAUGAAA GUCAGGUACCUAAAGUAGCGC GCGGAGC	45	67
phum-mir-276	mir-276	AAZO01003220.1	6659	6741	1	GGGUAUAAUUCAGCAGCGAGG UAUAGAGUUCCUACGUGCUUU GGUUACUGUAGGAACUUCAUA CCGUGCUCUUGGAUGUGCCU	14	36
phum-mir-277	mir-277	AAZO01001184.1	8494	8572	-1	UCGAAAUGCGUAUCAGAUGCG CGUUUACAAGUUUUUUUUUUAA UCAUCUGUAAAUGCACUAUCUG GUACGACAUUUCGG	51	73
phum-mir-278	mir-278	AAZO01005350.1	16264	16353	-1	AAAGGAAAAUGUUACGCUCGG AUGAAAGUUUUACCAAUCGUGU UAAAUAAAUUGAUCGGUGGGA CUUUCGUCCGUUUAAAACAUUU UAAU	56	77
phum-mir-279b	mir-279	AAZO01000045.1	15412	15504	1	AAGGAUUUUGUUAUUGGAUGA GUGAUGAUUUGGUUUCCUUAA AUUUUUUUUAUACUGGUGACU AGAUCUACACUCAUCACAAUAA	59	81

						UAAAAUCA		
phum-mir-305	mir-305	AAZO01007192.1	28951	29042	1	UUUGGUCUUUACGUUGAUUGUACUUCAUCAGGUGCUCUGGUAGUUCUUAAAUCCAGGCAUCUGGUGUAGUACUAUUAAUCUAAGAUCUAAAA	17	39
phum-mir-31a	mir-31	AAZO01004538.1	475	574	-1	AAAAGUUUGUUUUUAGACGGUGAGGUAGGCAAGAUGUCGGCAUAGCUGAAAUGAUUAUAAUAUGCGGCUGUGUCACAUCGAGCCAGCUUGAGCCGCCAUU	28	49
phum-mir-316	mir-316	AAZO01003258.1	2486	2576	-1	UACUCUUUAAAAUUCUGUCUUUUUCCGCUUUGCUGCCGGUGAUUUUAAAUAAACGACAGCAAAUAGGAAGAGGCCGAAUACAAAGAGUAUA	16	38
phum-mir-317	mir-317	AAZO01004205.1	881	971	1	GAGACCGGGUAUCAUCCUGUGUUGACUUCAUCGUUUUAUAAAUUAUUAUAAGAAAGUGAACACAGCUGGUGGUAUCUCAGCUUCUGGUCAG	57	81
phum-mir-34	mir-34	AAZO01004205.1	1295	1386	1	CGUCGUGCAGCAUUGGCAGUGUGGUUAGCUGGUUGUGUGGUGAUAACAAUAAAAUAGUCACAACCACUAUCUUCACUGGCAACGUAACACGG	14	35
phum-mir-375	mir-375	AAZO01003392.1	5884	5984	1	UUUGUCGGUUGCGUAUCGACCCGCGCCUUCUGAACAAAUUAGUAUUAUUAUCUAUGGUAAUUUGUUCGUUCGGCUCGAGUUAUUACGCAUCUCGCGAAACA	61	82
phum-mir-281	mir-46	1101323265948	198	264	-1	CGGACUAUCAAGAGAGCUAUCCGUCGACAGUACUGUUUAAAAACUGUCAUGGAAUUGCUCUCUUUAU	10	31
phum-mir-307	mir-67	AAZO01000007.1	12321	12418	1	UAUAUGCGUCAUUUCGGUCACUCACUCAACUUGGGUGUGUUACGUAAGUGAUUCGCCACAACCUCCUUGAGUGAGCGUCCGAGACAGUCGAUUCGUCU	56	75
phum-mir-7	mir-7	AAZO01007227.1	5984	6060	1	GGGUAUCUUGUGGAAGACUAGUGAUUUUGUUGUUCUUUAUUCGAGUAACAAGGAAUCACUAAUCAUCCCAAGAUGGU	11	33
phum-mir-71	mir-71	AAZO01007079.1	39876	39959	1	UCAACUUUUUGAAAGACAUGGGUAGUGAGAUGUUGGUUCACGUGUCAUUUUUACCUCUCACUACCUUGUCUUUCAUGAAGAAGU	10	31
phum-mir-8	mir-8	AAZO01008551.1	16823	16909	1	AGAGUAACUGUUUACAUCUUACCGGGCAGCAUUAGAUUGAUUUUAAGUAAUUCUAAUACUGUCAGGUAAUGAUGUCCUCAGGCUCU	54	76

						U		
phum-mir-87	mir-87	AAZO01003919.1	17390	17471	-1	AACCGAUUUACCGCCUGAAUCAUUGCUCGACCAAUUUCUUUAAUUAAAAAGCUGAGCAAAGUUUCAGGUGUGUCAAACGGUC	51	72
phum-mir-79a	mir-9	AAZO01004036.1	9071	9159	1	AGUCCUUUAUUAUUUUGCUUUGGUAAUUUAGUUUAUGAUGACUUUUAUACUCAUAAAGCUAGAUUACCAAAGCAUGAUUGAAAGGAAAG	54	75
phum-mir-79b	mir-9	AAZO01005481.1	27671	27770	-1	AAAUGAUGUCUGGUCUCUUCUUCUUUGGUUACCUAGUUUUGCGGACAUAUUUGAGAACCCAUAAAGCUAGGUCAGCAAAGCAGAAUGGGUCAGACGAAUC	61	81
phum-mir-9a	mir-9	AAZO01004036.1	9198	9299	1	GUCAGAAAUCAUAAAUCCUUUCUUUGGUGAUCUAGCUGUAUGGAUAAUAAUUUUAUUCCAUAAAGCUAUUUUACCGAUGUCAGGAAUAUUUGGUAAAUUCAA	21	43
phum-mir-9b	mir-9	AAZO01004036.1	9055	9176	-1	UAUGCAAUACACGUCGCCUUUCCUUUCAAUCAUGCUUUGGUAAUCUAGCUUUAUGAGUAUAAAAGUCAUCAUAAACUAAAUUACCAAAGCAAAAUAAUAAAGGACUUGAAAAAAAAAGUAAA	34	56
phum-mir-9c	mir-9	AAZO01004437.1	24805	24891	-1	UGGCGCUGAUUUUUUUCUUUGGUUAUCUAGCUGUAUGGAUAUUUUGAACUUCAUAAAGCUACAUUACCGAAGGUAAUAACAGCGCAA	16	38
phum-mir-927	mir-927	AAZO01006968.1	5667	5758	-1	AAGUUCGGUUUCGAUUUAGAAUUCCUACGCUUUACCUUUUUUAUAUAUAAUAUUGGCAAGGCGUUUGAAUUCUGAAUCCGAAUCAACUGUUU	15	36
phum-mir-929	mir-929	AAZO01002930.1	51332	51438	1	ACCUGACUGACUCAGUCGGUGUCAAAUUGACUCUAGUAGGGAGUCCAGCUUUUAUUAAAAUGGCGACUCCCUAACGGAGUCAGAUUGACUCCUUCUGGGAAAACCAG	26	46
phum-mir-965	mir-965	AAZO01002888.1	59224	59329	-1	ACAUUGUCGUAUUAGUACUCUCAUAGGGGAAAACUUUAUGCCUUAUGAUGCUAAUUAAAAUCCAUAAGCGUAUAGCUUUUCCCCUUUGGUGGCACUAAAAAUCUAC	65	86
phum-mir-981	mir-981	AAZO01001042.1	25366	25463	-1	AAAAAGAACCUACACGUUCCGGGUUUCUUGACAAUUUGAACCUUUUUUUUAUAUUUUAUGGUUCGUUGUCGACGAAACCUGCAUUGUGUAGGAUAAAA	62	83

phum-mir-100	mir-99	AAZO01005184.1	23040	23114	1	UUGCUCUUACCCGUAGAUCCGAACUUGUGUUUUUUUUACGUUAUGAUAUUCACAAGGUCGCCUCUGCAGGUAUCA	8	29
phum-mir-993	mir-993	AAZO01003084.1	30601	30691	-1	GGCACACUCGUAAUCUACCCUGUAGAUCCGGGCUUUUGUUGUUUUACCUCAUAAUCAGAAGCUCGUCUCUACAGGUAUCUUACGGGUGUAC	58	80
phum-mir-995	mir-995	AAZO01003562.1	18124	18213	-1	AUAGGCUUAGACGCAAACUGGAUAAUGUGAUUGCAACGUUUUCUAUAAAAAUCGUAGCACCACAUGAUUCAGCUUGCUUCUGGUCUUCUA	55	75
phum-mir-iab-4	mir-iab-4	1101323647795	626	692	1	UCUUGUACGUAUACUGAUGUAUCUGAGUGUAUUGCUUUCUGGUAUACUUUCAGUAUACGUAACAGGA	7	26
phum-mir-iab-4as	mir-iab-4as	1101323647795	626	692	-1	UCCUGUUACGUAUACUGAAAGUAUACCAGAAAGCAAUACACUCAGAUACAUCAGUAUACGUACAAGA	6	28

A.2 VIRUS GENOMICS

A.2.1 *New complete genome sequences of human rhinoviruses shed light on their phylogeny and genomic features*

Tapparel C, Junier T, Gerlach D, Cordey S, Van Belle S, Perrin L, Zdobnov EM, and Kaiser L. New complete genome sequences of human rhinoviruses shed light on their phylogeny and genomic features. *BMC Genomics* (2007) 8:224.

A.2.1.1 *Contributions*

Tapparel et al., 2007 presents a study of newly sequenced rhinovirus genomes focusing on evolution and conserved structural elements.

I predicting all conserved RNA structural elements and a new *cre* (*cis*-acting replication element) for the human rhinovirus B (HRV-B) species. I also participated in the analysis of the GC content of the different species.

A.2.1.2 *Main paper*

See pages 216–226 or at:
`http://www.biomedcentral.com/1471-2164/8/224`

BMC Genomics

BioMed Central

Research article

Open Access

New complete genome sequences of human rhinoviruses shed light on their phylogeny and genomic features

Caroline Tapparel*†1, Thomas Junier†2, Daniel Gerlach2, Samuel Cordey1, Sandra Van Belle1, Luc Perrin1, Evgeny M Zdobnov†2,3,4 and Laurent Kaiser†1

Address: [1]Central Laboratory of Virology, Division of Infectious Diseases, University of Geneva Hospitals, 24 Rue Micheli-du-Crest, 1211 Geneva 14, Switzerland, [2]Department of Genetic Medicine and Development, University of Geneva Medical School, 1 Rue Michel-Servet, 1211 Geneva 14, Switzerland, [3]Swiss Institute of Bioinformatics, 1 Rue Michel-Servet, 1211 Geneva 14, Switzerland and [4]Imperial College London, South Kensington Campus, SW7 2AZ London, UK

Email: Caroline Tapparel* - caroline.tapparel@hcuge.ch; Thomas Junier - Thomas.Junier@medecine.unige.ch; Daniel Gerlach - Daniel.Gerlach@medecine.unige.ch; Samuel Cordey - samuel.cordey@hcuge.ch; Sandra Van Belle - van-belle-sandra@diogenes.hcuge.ch; Luc Perrin - luc.perrin@hcuge.ch; Evgeny M Zdobnov - zdobnov@medecine.unige.ch; Laurent Kaiser - laurent.kaiser@hcuge.ch

* Corresponding author †Equal contributors

Published: 10 July 2007

BMC Genomics 2007, **8**:224 doi:10.1186/1471-2164-8-224

Received: 22 March 2007
Accepted: 10 July 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/224

Abstract

Background: Human rhinoviruses (HRV), the most frequent cause of respiratory infections, include 99 different serotypes segregating into two species, A and B. Rhinoviruses share extensive genomic sequence similarity with enteroviruses and both are part of the picornavirus family. Nevertheless they differ significantly at the phenotypic level. The lack of HRV full-length genome sequences and the absence of analysis comparing picornaviruses at the whole genome level limit our knowledge of the genomic features supporting these differences.

Results: Here we report complete genome sequences of 12 HRV-A and HRV-B serotypes, more than doubling the current number of available HRV sequences. The whole-genome maximum-likelihood phylogenetic analysis suggests that HRV-B and human enteroviruses (HEV) diverged from the last common ancestor after their separation from HRV-A. On the other hand, compared to HEV, HRV-B are more related to HRV-A in the capsid and 3B-C regions. We also identified the presence of a 2C *cis*-acting replication element (*cre*) in HRV-B that is not present in HRV-A, and that had been previously characterized only in HEV. In contrast to HEV viruses, HRV-A and HRV-B share also markedly lower GC content along the whole genome length.

Conclusion: Our findings provide basis to speculate about both the biological similarities and the differences (e.g. tissue tropism, temperature adaptation or acid lability) of these three groups of viruses.

Background

Human rhinovirus (HRV) is the most frequent cause of infection across all age groups of the population [1]. Replication is often restricted to the upper respiratory tract leading to self-limited illnesses such as the common cold. However, HRV infections can also exacerbate pre-existing airway disorders, invade the lower respiratory tract and lead to serious complications [2,3].

BMC Genomics 2007, **8**:224 http://www.biomedcentral.com/1471-2164/8/224

HRVs are single positive-stranded RNA viruses of approximately 7200 base pairs. They belong to the *Picornaviridae* family and are closely related to HEVs, another genus of the same family. The genome organization of *Picornaviridae* is conserved among the family with a long 5'-untranslated region (UTR), a single open reading frame (ORF) encoding a polyprotein, a short 3'UTR, and a poly(A) tail [4]. The 5'-terminal UMP of the viral RNA is covalently linked to the small viral protein VPg [5]. The 5'UTR contains two structural elements [6]. One is the 5'-cloverleaf structure involved in the plus-strand RNA synthesis and in the process of switching from translation to replication [7,8]. The other is the internal ribosomal entry site (IRES) which promotes translation of the polyprotein. The 3'-UTR is necessary for efficient RNA replication, but the exact mechanism is still not well understood [9,10]. In addition, species-specific internal *cis*-acting replication elements (*cre*) have been identified in HEV [11,12], HRV-A [13] and HRV-B [14,15].

HRV strains have been classified into 99 serotypes [16] based on the ability of a given serum to neutralize virus growth of a given strain in cell culture, although several serotypes share significant antigenic cross-reactivity [17]. According to nucleotide sequence relatedness of some serotypes [18-21] and to sequence comparison of all serotypes in the VP1 [16,22] and VP4-VP2 capsid protein-coding regions [23], the 99 serotypes segregate in two different groups: 74 belong to the HRV-A species and 25 to the HRV-B species. In addition to the division of HRVs into two species, they have also been classified into major and minor groups according to receptor usage. The major group ofHRVs (composed of 65 serotypes of species A and 25 serotypes of species B) binds ICAM1, whereas the minor group viruses (9 serotypes of species A) bind preferentially to LDL receptors [24-27]. The existence of multiple serotypes within each of these two lineages and different receptor usage support the hypothesis of significant differences at the protein level. Surprisingly, despite the fact that HRVs are the major cause of human respiratory infections, little is known about their genome variability at the full-length scale. To the best of our knowledge, part of the VP4-VP2 [23] and VP1-2A [16,22,28] regions have been sequenced for all serotypes and half of them for the 3D regions [19], but full-length sequences of only 8 serotypes are publicly available in the *Picornaviridae* database [29] (7 HRV-A and 1 HRV-B) [30-37]. While the present manuscript was in the process to be accepted, Kistler and coworkers published additional HRV-A and HRV-B full-length sequences increasing significantly the number of sequences available [38].

Among the *Picornaviridae*, HEVsare the closest relatives of HRVs and, as for HRVs, humans are the only known reservoir. Phylogenetic analyses of VP1-2A HRV and HEV sequences suggest thatHRVs and HEVs could be considered members of the same genus [28]. In addition, HRV-87 presents a high sequence similarity to HEVs and was recently reclassified as EV-68 [18,23,39,40]. Yet, the exact relation between HEV and HRV remains ambiguous without full-length genome comparison. At the phenotype level, however, HRV and HEV are clearly distinct: *in vitro*, cell tropism, pH tolerance and optimal growth temperature are significantly different; *in vivo*, the site of infection, organ tropism and the ability to disseminate are well-established characteristics that differentiate HEVs from HRVs. HRVs infections are restricted to the respiratory tract (temperature of 33°C), whereas most HEVs have the ability to replicate predominantly in the gastrointestinal tract (37°C). A large proportion can also disseminate, causing viremia and potentially invading the central nervous system [41]. Full-length genome comparison of these two genera helped us to identify genomic features and divergences at the amino acid level that might explain some biological differences.

We have sequenced 12 full-length genomes of different HRV serotypes and we present here the comparative analysis of 20 prototype HRV strains (13 HRV-A and 7 HRV-B) and 14 publicly available HEV strains that identifies the key elements differentiating these medically important viruses.

Results

The 12 newly sequencedHRV genomes have been deposited in GenBank [GenBank accession EF173414–EF173425]. They vary in sequence length from 7124 nucleotides (HRV-12) to 7219 nucleotides (HRV-17) which is similar to the length range of previously sequenced genomes (from 7102 to 7208 nucleotides). The average size ofHRVs type A (7131 nt) is smaller than the average size of HRVs type B (7215 nt), whereas the average size of 14 HEVs analysed in this study is 7417.

HRV-B is more closely related to HEV than to HRV-A

Phylogenetic analysis

The phylogeny of HRV-A, HRV-B and HEV was reconstructed by using maximum-likelihood phylogenetic method for the full polyproteins (Figure 1), as well as for each individual protein, using Simian picornavirus (SV-2) as the outgroup. The whole-polyprotein phlyogenetic analysis suggests the hypothesis that HRV-B and HEV lineages radiated from a common ancestor after its separation from HRV-A (Figure 1), where the percentage of bootstrap support of each grouping reflects the statistical confidence. Yet, when the analysis is conducted at the level of each individual protein, the subsequent reconstructed tree topology does not always support the same conclusion (Figures 2B-D and additional file 1). In the region of VP2, VP3, 3B and 3C proteins, the analysis sup-

ports the alternative hypothesis that HRV-B and HRV-A radiated from a common ancestor after separation from HEV. In the region of VP4, VP1 and 3A proteins, the analysis cannot discriminate the phylogenetic relationships between these three virus groups. The bootscanning experiment (see Methods) presented in Figure 2D is consistent with all the above-mentioned findings and also supports the hypothesis that in some parts of the capsid HRV-A and HRV-B share the last ancestor after their separation from HEV.

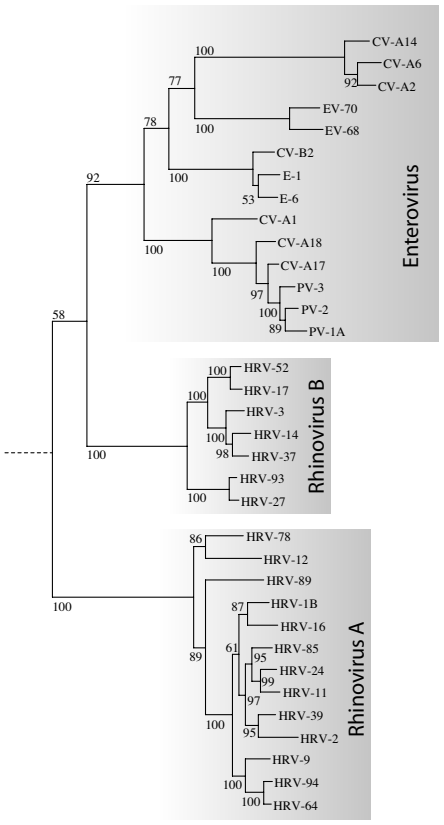


Figure 1
Whole-polyprotein phylogenetic tree. Whole-polyprotein, maximum likelihood phylogenetic tree shows a closer relation between HRV-B and HEV than between HRV-A and HRV-B. The figure indicates the percentage of bootstraps (out of 1000) that supports the corresponding clade. The sequence of simian picornavirus 1 (SV-2) was used as an outgroup. The branch lengths are measured in substitutions per site.

Protein product similarity

For each individual protein cleavage product, we also quantified pair-wise sequence identities among the HRV-A, HRV-B, and HEV genomes under consideration. Figure 2(A-B-C) shows the arrangement of these proteins along the picornavirus genomes and the corresponding protein identity matrices. Over all these three picornavirus (HRV-A, B and HEV), VP1 is the least conserved protein (48.5 % amino-acid identity) and VP4 the most conserved (66,6%). Globally, these comparisons are consistent with the phylogenetic analysis and confirm that the 2A, 2B, 2C, and 3D percentage of sequence similarities are higher between HRV-B and HEV than between HRV-A and B (see additional files 1 and 2). In contrast, at the level of VP1, VP2, VP3, 3B and 3C sequences, HRV-A and HRV-B have a higher percentage of homologies compared to HEV (see additional files 1 and 2).

Between the HRV-A and HRV-B groups, VP2 shows the highest overall conservation (> 60%) and 2A protein exhibits the lowest (< 40%). Within each of the species (HRV-A or HRV-B), the VP1 protein appears as the least conserved (< 80% of averaged amino-acid identity) and VP4 as the most conserved (> 96%) (see additional file 2). Protein 3B also shows poor conservation among HRV-B serotypes, but this may be an artefact due to its small size.

RNA structural elements

5'- and 3'- UTRs

Analysis of the 5'UTR of HRV-A, HRV-B and HEV shows that the well known 5' cloverleaf *cre* element as well as the IRES structure are highly conserved throughout the three groups (see additional file 3). The cloverleaf structure was originally discovered in polioviruses [42]. This secondary structure is deposited in the Rfam database (a collection of multiple sequence alignments covering many common non-coding RNA families and conserved RNA secondary structures [43]) under the accession number [RF00386]. The corresponding consensus structure of HRV-A, HRV-B and HEV, which was recovered without prior knowledge of it by comparative sequence analysis (see additional file 3A) matches well this Rfam consensus structure. In the same line, the IRES structure for HRV-A, HRV-B and HEVs is also very similar to the Rfam *Picornaviridae* consensus structure [RF00229] [44], except for the presence of two additional small helices (see additional file 3B).

The picornavirus 3'-UTR encodes astem-loop structure that may play a role in replication efficiency (through interaction with the 5'UTR) as well as in polyadenylation of genomic RNA [9,45,46]. In contrast to the 5'UTR structures, the 3'UTRs structures (see additional file 4)are not universally conserved in sequence and position among the three groups studied. The length of the 3' UTR between the groups varies between 47 nt (HRV-A), 50 nt

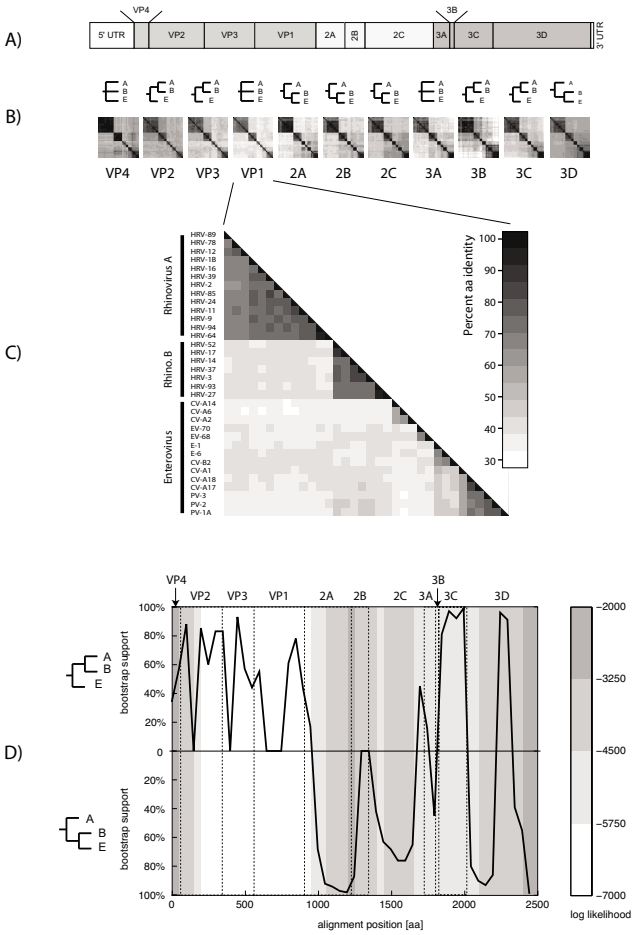


Figure 2

Protein and amino-acid similarity comparison between HRV-A, HRV-B and HEV. A) Schematic representation of HEV and HRV genome organization showing boundaries of encoded proteins. B) Protein similarity comparison. For each protein, the following is shown: – top row: a simplified tree representation of the relationships between HRV-A (A), HRV-B (B) and HEV (E), according to the corresponding ML tree (see additional file 1). Three cases are possible: HRV-B closer to HRV-A; HRV-B closer to HEV; and undecided (none of the above clearly more likely than the other). – bottom row: an all-versus-all sequence identity matrix (darker colorus indicate higher identity percentage). The similarity values are given in additional file 2. C) Close-up of the identity matrix for VP1. D) Bootscanning. The whole polyprotein alignment was divided into windows of 200 aa starting every 50 aa, and a 100-bootstrap ML tree was computed on each window. The black curve indicates the degree of support (as a percentage of bootstrap replicates) for either the "HRV-B closer to HRV-A" (upper half) or "HRV-B closer to HEV" (lower half) topology at each position along the whole genome (see Methods for details). The background colour reflects the log likelihood of the tree at each position which is a measure of overall confidence in the tree. Darker colour indicates higher confidence.

BMC Genomics 2007, **8**:224 http://www.biomedcentral.com/1471-2164/8/224

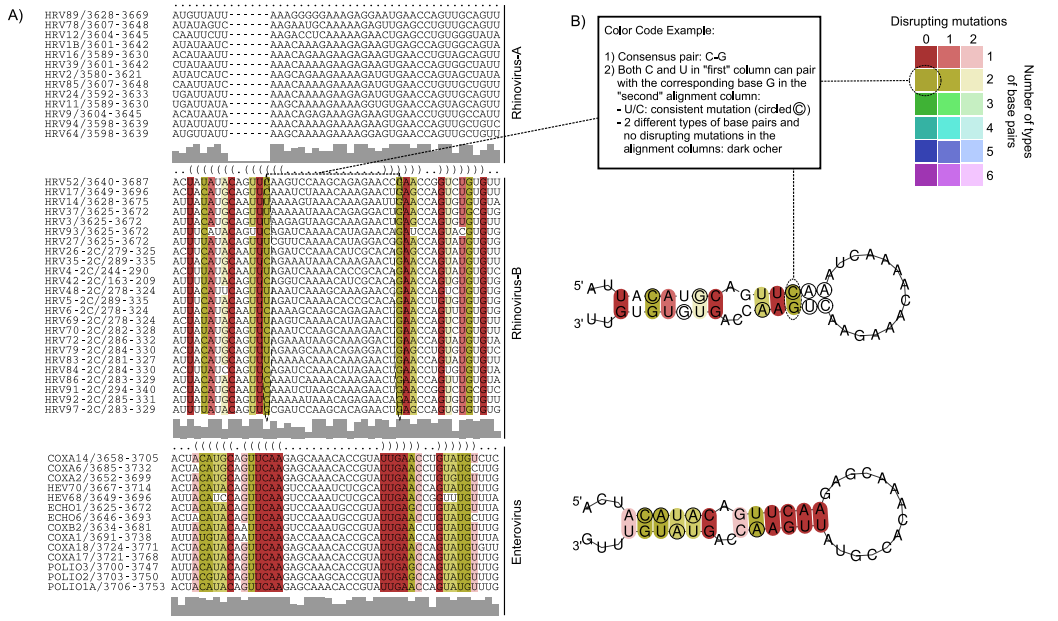


Figure 3
Alignments and conserved secondary structures for cis-acting 2C replication elements conserved within HRV-B and HEV. A) Multiple sequence alignment across all considered genomes that shows consensus secondary RNA structure (in dot bracket format, see first row); sequences are colour-coded according to RNA structure conservation; the sequence conservation profile for each group is shown in grey bars beneath the alignments. B) Secondary structure of the conserved *cre* 2C, colour-coded according to the different types of base pairs in the corresponding alignment columns. The more different the types of base pairs existing for two pairing alignment columns, the more evolutionary conservation of the structure (cp. compensatory and consistent mutations).

(HRV-B) and 83 nt (HEV). Furthermore, both HRV-A and HRV-B contain a stable stem-loop structure of 35 nt at the 3' end of the 3'UTR. HEV also contains a 42 nt stem-loop which is located closer to the middle of the 3'UTR. Nevertheless, there is a large amount of sequence variability within this whole group which leads to a less stable consensus stem-loop for all the analyzed HEV sequences.

In addition, there is a conserved stem-loop structure in HRV-A located close to the 3'UTR, yet the corresponding region in HRV-B and HEV suggest different structures, and the overall high sequence conservation in the region could give a misleading signal of structural conservation (see additional file 5).

Internal cis-acting elements

Besides the 5' and 3'UTR, disparate internal *cre* elements have been previously described among various rhinoviral serotypes of both HRV-A and HRV-B [15], and have been identified by our comparative analysis.

HRV-A cre

The HRV-2 2A internal *cre* motif [Rfam RF00220] [13] is conserved among all HRV-A genomes analysed in this study, but has not been identified in any HRV-B or HEV viruses (see additional file 6A). The same region of HRV-B also folds into a conserved secondary structure that seems specific to this group (data not shown).

HRV-B cre

Similarly, the internal *cre* motif reported for the HRV-14 VP1, a member of HRV-B, is present in all 7 HRV-B serotypes and is notably absent in all HRV-A and HEV analyzed (see additional file 6B).

Furthermore, the availability of new HRV-B sequences allowed us to identify another conserved *cre* motif within

BMC Genomics 2007, **8**:224 http://www.biomedcentral.com/1471-2164/8/224

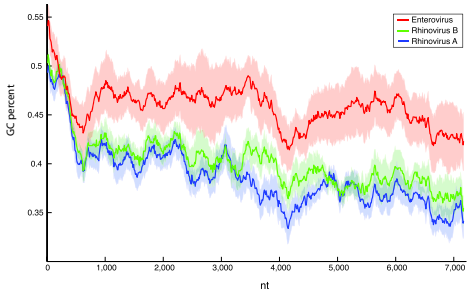


Figure 4
Local GC composition of HRV-A, HRV-B, and HEV. Average GC percentage computed over a sliding window of 600 nt and a step of 10 nt along whole-genome multiple alignments of HRV-A, HRV-B, and HEV, respectively (thick lines). The shaded areas represent one standard deviation above and below the average.

the HRV-B 2C coding sequence (Figure 3) that has the typical R_1NNNAAR_2NNNNNNR$_3$ *cre* motif [47-51] in all HRV-B serotypes analysed (the 7 full genomes plus 17 partial sequences), with the exception of HRV-27 that has a U instead of an R at position R_1. More importantly, the newly identified HRV-B 2C *cre* corresponds to the HEV 2C *cre*, previously identified in several HEVs [11,12].

GC content

The GC composition is an important genomic factor that can be evolutionary optimized for adaptation to multiple environmental constraints (such as ideal growth temperature). The GC content varies substantially between the groups of HEV, HRV-A and HRV-B (Figure 4), where HRV-B exhibits lowest values, HEV exhibits the highest values, and HRV-B is intermediate. This holds not only globally, but also locally, for each of the sliding windows along the whole genomes. These trends are statistically significant as the two-sided Kolmogorov-Smirnov test rejects the hypothesis that GC contents of HRV-A, HRV-B and HEV can be drawn from the same underlying distribution: HRV-A vs. HRV-B p-value < 10^{-15}; HRV-A vs. HEV p-value < 10^{-15}; HRV-B vs. HEV p-value < 10^{-15}.

Discussion

HRVs were first classified into two groups based on a differential sensitivity to a variety of antiviral compounds targeting VP1 [52]. The members of the HRV-A group were susceptible to most of these antiviral compounds, whereas the HRV-B were not. This classification was then confirmed by nucleotide sequence relatedness in the VP1 [16,22] and VP4-VP2 capsid protein-coding regions of all serotypes [23]. Analysis of other regions like the 3C protease has been restricted to a limited number of serotypes [18,20,21]. Whole genome comparisons have not been conducted since only one full-length HRV-B genome (HRV-14) as well as a limited number of HRV-A genomes were available. Complete sequencing and analysis of additional HRV-B and HRV-A genomes allowed us to describe their phylogeny and the similarity of individual proteins between the two HRV groups and HEV. For example, proteins such as 2A show a particularly pronounced difference in inter- versus intra-group conservation. Conversely, surface proteins such as VP2 (capsid) are better conserved across all groups.

It appears that HRV-B share a common ancestor with HEV as shown by the whole-genome phylogenetic analysis, which suggests that Rhinovirus is not monophyletic. This observation is reinforced by the identification of a new HRV-B 2C *cre* that is lacking in all HRV-A genomes studied. This *cre* consists of a hairpin structure with a conserved R_1NNNAAR_2NNNNNNR$_3$ motif in the loop [47-51] and was previously only known in HEV 2C gene. The first two As in this motif serve as the template for the VPg uridylylation reaction by the viral polymerase. Uridylylated VPg then serves as a primer for RNA synthesis [5]. Although the question is still open to debate, it has been suggested that the polyA tail may serve as template for VPg uridylylation and synthesis of the minus strand RNA, whereas internal *cre* are necessary for plus strand synthesis [53,54]. HRV-14 VP1 *cre*, as well as HRV-2 2A *cre*, were shown to be functionally equivalent to poliovirus RNA 2C *cre* in *in vitro* uridylylation assays with poliovirus VPg and polymerase [11-13,49]. Further studies are underway to define whether the HRV-B 2C *cre* identified in this study plays a role in VPg uridylylation or can be considered as an evolutionary leftover of HEV 2C *cre*. Concerning the replacement of an R position by a U in HRV-27 *cre*, it should be noted that the effect of a similar substitution on replication efficiency could not be studied for HRV-14 2C *cre* since it results in the introduction of a stop codon [47-51]. However, for Coxsackiesvirus B3 2C *cre*, a substitution at the R3 position by a U was shown permissive for replication [53,54].

Besides this putative new *cre* element in the 2C region of HRV-B, we could also identify the already known elements (cloverleaf structure and IRES in 5' as well as the stem-loop element in 3') in the 5' and 3' UTRs from all studied genomes. The cloverleaf structure and the IRES are highly conserved. Interestingly, we identified many compensatory mutations in the sequences of these structures, which points out that the selective pressure is "working"

on the structural level. The functionality of these elements is therefore more determined by their structure than by their primary sequence.

The observation that HRV-B and HEVs are more closely related to each other than either is to HRV-A seems paradoxical, given that HRVs differentiate themselves from most HEVs at the phenotypic level. This could be explained by recombination events that would have taken place soon after the divergence of HEV and HRV-A and during which regions were exchanged between HRV-A and the HEV ancestor of HRV-B. The protein identity plots and the bootscanning suggest that the recombining region may have included the capsid region. This is consistent with the fact that recombination breakpoints have been found to be largely restricted to nonstructural regions of the genome, mostly in the 2A-2C parts, and between the 5' UTR and the capsid-encoding region [55-57]. Recombination has been extensively documented as a driving force for the evolution of some *Picornaviridae* [58], although only hypothesized forHRVs [19]. Interspecies *in vivo* recombination was also suspected for HEVs [59]. Our hypothesis is that early recombination events may have occurred between HRV and HEV. Although these two virus species have often different tropism *in vivo*, both can easily infect the respiratory tract, an event that could provide opportunities for recombination.

Since VP1 is responsible for recognizing the receptor on the host cell surface, the hypothesis of capsid sequence transfer from HRV-A to HEV to yield HRV-B could explain a tissue tropism and a disease pattern similar to that of HRV-A rather than HEV. In addition, the similar GC content observed between HRV-B and HRV-A may account for some of these phenotypic differences. The relatively lower GC content of HRV-B may reflect an adaptation to the environment of the upper respiratory tract, whereas the higher GC content of HEVs might reflect convergent adaptation to the gastrointestinal tract of the central nervous system environment (such as higher temperature, acidity, etc.).

Conclusion

The analysis of new HRV full-length genomes statistically supports (> 90% bootsrap confidence) the current classification of HRV into two distinct species. HRV-B seems to be phylogenetycally more closely related to HEV, another important member of the *Picornaviridae* family, than it is to HRV-A. However, our observations suggest that this species classification accurately reflects the capsid type, but not all parts of the genome. We have shown that HRV-A and HRV-B differ significantly at the protein level and in the composition and structure of their *cis*-acting sequences. One of the evolutionary scenarios that would explain the differential grouping of HRV-B with HEVs or with HRV-A along the genome is that of an ancient recombination between HRV-A and HEV lineages, given that the HRV-B closer relation with HEVs is overall more statistically sound. However, without additional data, this remains only a hypothesis. The genomic features highlighted in our study help to contribute to our understanding of why these viruses maintain different phenotypic variations in humans, thereby enabling a more accurate analysis of their relationship.

Methods

Viruses

The prototype strains of 12 HRV serotypes (HRV-3, 17, 27, 37, 52, 93, 11, 12, 24, 78, 64 and 94) were obtained from the American Type Culture Collection (LGC Promochem, Molsheim, France) and the RNA was eitherextracted directly from ATCC stocks (HRV-17, HRV-52, HRV-64 and HRV-94) orthe stocks were first amplified by one (HRV-11, HRV-12, HRV-24 and HRV-78), two (HRV-27) or three (HRV-3, HRV-37 and HRV-93) passages in HeLa Ohio cell lines (kindly provided by Prof FG Hayden, University of Virginia, Charlottesville, VA, USA). These serotypes were chosen tobe well scattered on the trees performed previously with HRV VP1 and VP4-VP2 subregions [16,22,23] and to complete sequence analysis of clinical isolates studied in the laboratory [2].

The full-length genome sequences of the 8 additional HRV serotypes (HRV-1B [GenBank:D00239], 2 [GenBank: X02316], 14 [GenBank:X01087], 16 [GenBank: L24917], 39 [GenBank:AY751783], 89 [GenBank:M16248], 85 and 9 whose sequences were directly downloaded from the *Picornaviridae* sequence database [29]), as well as the sequences of the 14 HEV serotypes and the simian picornavirus (SV-2) outgroup [GenBank:AY064708] analyzed in this study, were obtained from GenBand at the NCBI. The 14 HEV sequences include the two members of the HEV-D subspecies: EV-68 [GenBank: EF107098] and EV-70 [GenBank:DQ201177], the three members of the poliovirus subspecies: PV-1 [GenBank:V01148], PV-2 [GenBank:X00595] and PV-3 [GenBank: X00925] as well as three representatives of the HEV-A, B and C subspecies randomly chosen: Coxsackie (CV)-A2 [GenBank:AY421760], CV-A6 [GenBank:AY421764]and CV-A14 for HEV-A [GenBank:AY421769]; Echovirus (E)-1 [GenBank:AF029859], E-6 [GenBank:AY302558] and CV-B2 [GenBank:AF081485]for HEV-B; and CV-A1 [GenBank:AF499635], CV-A17 [GenBank:AF499639] and CV-A18 [GenBank:AF499640] for HEV-C. A list of all viruses with their corresponding GenBank accession numbers can be found in the additional file 8 in the supplementary material.

Sequencing

Complete genome sequences were determined for each of the 12 above-mentioned strains. Reverse transcription (Superscipt II, Invitrogen, Basel, Switzerland) was performed with random hexamers on TRIzol- extracted (Invitrogen) RNA [60]. Overlapping fragments representing each complete viral genome were then amplified by PCR using degenerate primers designed to anneal highly conserved sequences among HRVs. Specific, non-degenerate primers were then designed to fill the gaps between the original PCR products. All primers used in this study are listed in the additional files (see additional file 7). The 5' and 3'ends were obtained with the 5'/3' RACE Kit (Roche Applied Science, Rotkreuz, Switzerland). PCR products were purified with the microcon columns (Millipore, Zug, Switzerland) before sequencing. Each PCR product was sequenced at least twice. Chromatograms produced with the ABI Prism 3130XL DNA Sequencer (Applied Biosystems, PE Europe BV, Basel, Switzerland) were directly imported for proofreading with the vector NTI Advance 10 program (Invitrogen).

Multiple sequence alignment

Open reading frames (ORFs) were extracted from the whole-genome nucleotide sequences of each virus species using the getorf programme from the EMBOSS package [61], using a minimal ORF length of 6000 nt to ensure that small, spurious ORFs were not reported. The multiple alignment of encoded polyprotein was produced with the extracted ORFs using MUSCLE [62] with default parameters. The alignments for each of the protein products were extracted from the full multiple alignment. The whole-genome nucleotide level alignment was assembled using T-Coffee [63] from 3 separate alignments: 5'-UTR and 3'-UTR aligned with MUSCLE (default parameters), and the amino acid level multiple alignment of the ORFs projected to the nucleotide level using the TRANALIGN programme from the EMBOSS package with default parameters. The alignments are available from [64].

Phylogenetic analysis

The maximum-likelihood phylogenetic analyses were performed using PhyML [65] with estimated proportion of variable sites, estimated Gamma distribution parameters and 16 substitution rate categories. Protein-level trees were made using the JTT [66] molecular evolution model, and nucleotide-level trees were made using the GTR model with empirical base frequency estimates. The consensus trees were reconstructed from bootstrap trees using PHYML or Tree-Puzzle [67] with the same parameters.

Protein identity plots

All-against-all protein product identity scores were produced using the Belvu programme [68], and reformatted into symmetrical square arrays of sequence identity percentage values (one for each cleavage product) represented as greyscale bitmaps in Figures 2B and 2C.

Bootscanning

A polyprotein multiple alignment was constructed (as described above) with 14 HEV sequences, 13 HRV-A sequences, 7 HRV-B sequences, and 1 SV-2 sequence. This alignment was subjected to bootscanning (as described in [69]) with a window size of 200 aa and a step of 50 aa. For each window, a maximum-likelihood tree with 100 bootstraps was computed as described above. HRV-A formed a single clade in all trees, HRV-B in all but one. HEV formed a single clade in many, but not all trees. The tree's topology was categorized as follows: i) the smallest clade that contained all rhinoviruses and at least one enterovirus was determined; ii) this clade was categorized as "HRV-B closest to HRV-A", "HRV-B closest to HEV", or "undecided" according to which clade was the sister clade of all HRV-B, iii) the bootstrap value of the clade determined in (i) was used as a measure of support of the topology. Finally, the log-likelihood of each tree was also recorded. For each window, this yielded: i) an indication of the most likely topology (with possibility of undecidedness); ii) a measure of support of this topology; and iii) a measure of confidence in the whole tree.

GC content

We extracted sub-alignments for HEV, HRV-A, and HRV-B from the above-described nucleotide-level, whole-genome alignment of 14 HEV, 13 HRV-A, 7 HRV-B sequences. This allows direct comparison of the GC content at the orthologous positions using a sliding window of 600 nt along the alignment, computing GC percentage over all sequences within the window, and with a step of 10 nt. The resulting set of three measures of local GC percentage content, one each for HEV, HRV-A, and HRV-B were plotted.

Identification of conserved RNA structural elements

The complete genome alignment of all 34 genomes spanning 7852 positions (5'UTR+ORF+3'UTR) was scanned for thermodynamically stable and structurally conserved RNA structures using RNAz [70] The structures were evaluated using a sliding window of 120 bp with 40 bp steps over the whole alignment, as well as separately for each of the three groups (HRV-A, HRV-B, HEV). To identify shorter secondary structure elements, the same procedure was performed using a window of 60 bp in steps of 20 bp. The consensus RNA structures of the selected alignment regions were folded using RNAalifold from the Vienna Package [71] with the least stringent option for consensus folding. These alignment regions were manually elongated and corrected in order to capture the whole RNA secondary structure. Furthermore, all alignment columns with more than 75% gaps were removed from the RNAal-

ifold consensus folding procedure, since gaps are not excluded for the folding energies evaluation. The resulting structures as well as the alignments were color-coded according to the amount of consistent, compensatory and inconsistent base changes at a certain alignment and structure position using Vienna RNA Utilities [72].

Abbreviations

HRV: Human Rhinovirus

HEV: Human Enterovirus

CV: Coxsackie Virus

E: Echovirus

PV: Poliovirus

EV-68: HEV-68

SV: Simian Picornavirus

cre: cis-acting replication element

UTR: untranslated region

UMP: Uridine monophosphate

IRES: internal ribosomal entry site

ORF: open reading frame

ML: maximum-likelihood

Authors' contributions

CT designed the original project, conducted and supervised the experiments (primer design, PCR conditions, sequence assembly and proofreading) and drafted the manuscript. SVB conducted most of the experiments, SC conducted the *cre* sequencing and revised the manuscript. TJ conducted all the sequences, phylogenetic analyses and calculation of GC content and participated in the writing of the manuscript. DG analyzed the RNA secondary structure, identified the *cre* elements and participated in the writing of the manuscript. LP participated to the analysis and the writing of the manuscript. EZ supervised and designed all the bioinformatics work and corrected the manuscript. LK designed the original project, supervised the complete work and corrected the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Whole-protein maximum likelihood phylogenetic trees for the 11 individual picornavirus proteins. *Each individual protein tree was performed as the whole polyprotein phylogenetic tree (Figure 1).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-224-S1.pdf]

Additional file 2

Similarity values among HRV-A, HRV-B and HEV for the 11 individual protein products. *The similarity values for protein comparison between HRV-A, HRV-B and HEV represented in Figure 2 are listed.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-224-S2.xls]

Additional file 3

5' UTR structure conservation. *A) 5'cloverleaf consensus structure for HRV-A, HRV-B and HEV identified by comparative sequence analysis. B) IRES consensus structure for HRV-A, HRV-B and HEV identified by comparative sequence analysis. See legend to Figure 3 for details.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-224-S3.pdf]

Additional file 4

3'UTR structure conservation. *3'UTR consensus structure for HRV-A, HRV-B and HEV identified by comparative sequence analysis. See legend to Figure 3 for details*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-224-S4.pdf]

Additional file 5

Conserved stem-loop structure in the ORF of HRV-A. *Conserved secondary structure located close to the 3'UTR of HRV-A and corresponding structures in HRV-B and HEV located in the same alignment region. See legend to Figure 3 for details*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-224-S5.pdf]

Additional file 6

Internal* cre *conservation among HRV-A and HRV-B serotypes. *A) Internal 2A* cre *conservation among HRV-A serotypes. B) Internal VP1* cre *conservation among HRV-B serotypes*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-224-S6.pdf]

Additional file 7

Primer list. *Degenerate and specific primers used to amplify and sequence the new rhinovirus genomes.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-224-S7.doc]

Additional file 8

Virus accession number. *List of all the accession numbers for the viruses used in the analyses.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-224-S8.xls]

Acknowledgements

We would like to thank Chantal Gaille for technical assistance and Sabine Yerly for support and comments on the manuscript. We also thank Rosemary Sudan for editorial assistance. This study was supported by the Swiss National Science Foundation (No. 3200B0-101670 to L. K. and 3100A0-112588/1 to E.Z.), the Canton of Geneva, and the University of Geneva Dean's programme for the promotion of women in science (CT). The Swiss Institute of Bioinformatics' Vital-IT facility [73] was used for the bootscanning.

References

1. Denny FW Jr.: **The clinical impact of human respiratory virus infections.** *Am J Respir Crit Care Med* 1995, **152(4 Pt 2):**S4-12.
2. Kaiser L, Aubert JD, Pache JC, Deffernez C, Rochat T, Garbino J, Wunderli W, Meylan P, Yerly S, Perrin L, Letovanec I, Nicod L, Tapparel C, Soccal PM: **Chronic Rhinoviral Infection in Lung Transplant Recipients.** *Am J Respir Crit Care Med* 2006.
3. Papadopoulos NG, Bates PJ, Bardin PG, Papi A, Leir SH, Fraenkel DJ, Meyer J, Lackie PM, Sanderson G, Holgate ST, Johnston SL: **Rhinoviruses infect the lower airways.** *J Infect Dis* 2000, **181(6):**1875-1884.
4. Kitamura N, Semler BL, Rothberg PG, Larsen GR, Adler CJ, Dorner AJ, Emini EA, Hanecak R, Lee JJ, van der Werf S, Anderson CW, Wimmer E: **Primary structure, gene organization and polypeptide expression of poliovirus RNA.** *Nature* 1981, **291(5816):**547-553.
5. Paul AV, van Boom JH, Filippov D, Wimmer E: **Protein-primed RNA synthesis by purified poliovirus RNA polymerase.** *Nature* 1998, **393(6682):**280-284.
6. Rohll JB, Percy N, Ley R, Evans DJ, Almond JW, Barclay WS: **The 5'-untranslated regions of picornavirus RNAs contain independent functional domains essential for RNA replication and translation.** *J Virol* 1994, **68(7):**4384-4391.
7. Huang H, Alexandrov A, Chen X, Barnes TW 3rd, Zhang H, Dutta K, Pascal SM: **Structure of an RNA hairpin from HRV-14.** *Biochemistry* 2001, **40(27):**8055-8064.
8. Paul AV: **Possible unifying mechanism of picornavirus genome replication.** In *Molecular Biology of Picornaviruses Volume 1*. Edited by: Semler BL, Wimmer E. Washington DC , ASM Press; 2002:227-246.
9. Brown DM, Cornell CT, Tran GP, Nguyen JH, Semler BL: **An authentic 3' noncoding region is necessary for efficient poliovirus replication.** *J Virol* 2005, **79(18):**11962-11973.
10. Todd S, Towner JS, Brown DM, Semler BL: **Replication-competent picornaviruses with complete genomic RNA 3' noncoding region deletions.** *J Virol* 1997, **71(11):**8868-8874.
11. Goodfellow I, Chaudhry Y, Richardson A, Meredith J, Almond JW, Barclay W, Evans DJ: **Identification of a cis-acting replication element within the poliovirus coding region.** *J Virol* 2000, **74(10):**4590-4600.
12. van Ooij MJ, Vogt DA, Paul A, Castro C, Kuijpers J, van Kuppeveld FJ, Cameron CE, Wimmer E, Andino R, Melchers WJ: **Structural and functional characterization of the coxsackievirus B3 CRE(2C): role of CRE(2C) in negative- and positive-strand RNA synthesis.** *J Gen Virol* 2006, **87(Pt 1):**103-113.
13. Gerber K, Wimmer E, Paul AV: **Biochemical and genetic studies of the initiation of human rhinovirus 2 RNA replication: identification of a cis-replicating element in the coding sequence of 2A(pro).** *J Virol* 2001, **75(22):**10979-10990.
14. McKnight KL, Lemon SM: **Capsid coding sequence is required for efficient replication of human rhinovirus 14 RNA.** *J Virol* 1996, **70(3):**1941-1952.
15. McKnight KL, Lemon SM: **The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication.** *Rna* 1998, **4(12):**1569-1584.
16. Ledford RM, Patel NR, Demenczuk TM, Watanyar A, Herbertz T, Collett MS, Pevear DC: **VP1 sequencing of all human rhinovirus serotypes: insights into genus phylogeny and susceptibility to antiviral capsid-binding compounds.** *J Virol* 2004, **78(7):**3663-3674.
17. **Rhinoviruses: a numbering system.** *Nature* 1967, **213(78):**761-762.
18. Binford SL, Maldonado F, Brothers MA, Weady PT, Zalman LS, Meador JW 3rd, Matthews DA, Patick AK: **Conservation of amino acids in human rhinovirus 3C protease correlates with broad-spectrum antiviral activity of rupintrivir, a novel human rhinovirus 3C protease inhibitor.** *Antimicrob Agents Chemother* 2005, **49(2):**619-626.
19. Savolainen C, Laine P, Mulders MN, Hovi T: **Sequence analysis of human rhinoviruses in the RNA-dependent RNA polymerase coding region reveals large within-species variation.** *J Gen Virol* 2004, **85(Pt 8):**2271-2277.
20. Horsnell C, Gama RE, Hughes PJ, Stanway G: **Molecular relationships between 21 human rhinovirus serotypes.** *J Gen Virol* 1995, **76 (Pt 10):**2549-2555.
21. Mori J, Clewley JP: **Polymerase chain reaction and sequencing for typing rhinovirus RNA.** *J Med Virol* 1994, **44(4):**323-329.
22. Laine P, Blomqvist S, Savolainen C, Andries K, Hovi T: **Alignment of capsid protein VP1 sequences of all human rhinovirus prototype strains: conserved motifs and functional domains.** *J Gen Virol* 2006, **87(Pt 1):**129-138.
23. Savolainen C, Blomqvist S, Mulders MN, Hovi T: **Genetic clustering of all 102 human rhinovirus prototype strains: serotype 87 is close to human enterovirus 70.** *J Gen Virol* 2002, **83(Pt 2):**333-340.
24. Greve JM, Davis G, Meyer AM, Forte CP, Yost SC, Marlor CW, Kamarck ME, McClelland A: **The major human rhinovirus receptor is ICAM-1.** *Cell* 1989, **56(5):**839-847.
25. Hofer F, Gruenberger M, Kowalski H, Machat H, Huettinger M, Kuechler E, Blass D: **Members of the low density lipoprotein receptor family mediate cell entry of a minor-group common cold virus.** *Proc Natl Acad Sci U S A* 1994, **91(5):**1839-1842.
26. Staunton DE, Merluzzi VJ, Rothlein R, Barton R, Marlin SD, Springer TA: **A cell adhesion molecule, ICAM-1, is the major surface receptor for rhinoviruses.** *Cell* 1989, **56(5):**849-853.
27. Uncapher CR, DeWitt CM, Colonno RJ: **The major and minor group receptor families contain all but one human rhinovirus serotype.** *Virology* 1991, **180(2):**814-817.
28. Laine P, Savolainen C, Blomqvist S, Hovi T: **Phylogenetic analysis of human rhinovirus capsid protein VP1 and 2A protease coding sequences confirms shared genus-like relationships with human enteroviruses.** *J Gen Virol* 2005, **86(Pt 3):**697-706.
29. **Picornaviridae.com** [http://www.picornaviridae.com/sequences/sequences.htm]
30. Stanway G, Hughes PJ, Mountford RC, Minor PD, Almond JW: **The complete nucleotide sequence of a common cold virus: human rhinovirus 14.** *Nucleic Acids Res* 1984, **12(20):**7859-7875.
31. Skern T, Sommergruber W, Blaas D, Gruendler P, Fraundorfer F, Pieler C, Fogy I, Kuechler E: **Human rhinovirus 2: complete nucleotide sequence and proteolytic processing signals in the capsid protein region.** *Nucleic Acids Res* 1985, **13(6):**2111-2126.
32. Lee WM, Wang W, Rueckert RR: **Complete sequence of the RNA genome of human rhinovirus 16, a clinically useful common cold virus belonging to the ICAM-1 receptor group.** *Virus Genes* 1995, **9(2):**177-181.
33. Hughes PJ, North C, Jellis CH, Minor PD, Stanway G: **The nucleotide sequence of human rhinovirus 1B: molecular relationships within the rhinovirus genus.** *J Gen Virol* 1988, **69 (Pt 1):**49-58.
34. Harris JR, Racaniello VR: **Amino acid changes in proteins 2B and 3A mediate rhinovirus type 39 growth in mouse cells.** *J Virol* 2005, **79(9):**5363-5373.
35. Duechler M, Skern T, Sommergruber W, Neubauer C, Gruendler P, Fogy I, Blaas D, Kuechler E: **Evolutionary relationships within the human rhinovirus genus: comparison of serotypes 89, 2, and 14.** *Proc Natl Acad Sci U S A* 1987, **84(9):**2605-2609.

36. Callahan PL, Mizutani S, Colonno RJ: **Molecular cloning and complete sequence determination of RNA genome of human rhinovirus type 14.** *Proc Natl Acad Sci U S A* 1985, **82(3):**732-736.
37. Lee WM, Monroe SS, Rueckert RR: **Role of maturation cleavage in infectivity of picornaviruses: activation of an infectosome.** *J Virol* 1993, **67(4):**2110-2122.
38. Kistler AL, Webster DR, Rouskin S, Magrini V, Credle JJ, Schnurr DP, Boushey HA, Mardis ER, Li H, DeRisi JL: **Genome-wide diversity and selective pressure in the human rhinovirus.** *Virol J* 2007, **4:**40.
39. Andeweg AC, Bestebroer TM, Huybreghs M, Kimman TG, de Jong JC: **Improved detection of rhinoviruses in clinical samples by using a newly developed nested reverse transcription-PCR assay.** *J Clin Microbiol* 1999, **37(3):**524-530.
40. Blomqvist S, Savolainen C, Raman L, Roivainen M, Hovi T: **Human rhinovirus 87 and enterovirus 68 represent a unique serotype with rhinovirus and enterovirus features.** *J Clin Microbiol* 2002, **40(11):**4218-4223.
41. Racaniello VR: **Picornaviridae: the viruses and their replication, Chapter 23.** *Fields VIrology, fourth edition* 2001, **1:**685-722.
42. Barton DJ, O'Donnell BJ, Flanegan JB: **5' cloverleaf in poliovirus RNA is a cis-acting replication element required for negative-strand synthesis.** *Embo J* 2001, **20(6):**1439-1448.
43. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005, **33(Database issue):**D121-4.
44. Liu Z, Carthy CM, Cheung P, Bohunek L, Wilson JE, McManus BM, Yang D: **Structural and functional analysis of the 5' untranslated region of coxsackievirus B3 RNA: In vivo translational and infectivity studies of full-length mutants.** *Virology* 1999, **265(2):**206-217.
45. Serrano P, Pulido MR, Saiz M, Martinez-Salas E: **The 3' end of the foot-and-mouth disease virus genome establishes two distinct long-range RNA-RNA interactions with the 5' end region.** *J Gen Virol* 2006, **87(Pt 10):**3013-3022.
46. van Ooij MJ, Polacek C, Glaudemans DH, Kuijpers J, van Kuppeveld FJ, Andino R, Agol VI, Melchers WJ: **Polyadenylation of genomic RNA and initiation of antigenomic RNA in a positive-strand RNA virus are controlled by the same cis-element.** *Nucleic Acids Res* 2006, **34(10):**2953-2965.
47. Goodfellow IG, Polacek C, Andino R, Evans DJ: **The poliovirus 2C cis-acting replication element-mediated uridylylation of VPg is not required for synthesis of negative-sense genomes.** *J Gen Virol* 2003, **84(Pt 9):**2359-2363.
48. Paul AV, Rieder E, Kim DW, van Boom JH, Wimmer E: **Identification of an RNA hairpin in poliovirus RNA that serves as the primary template in the in vitro uridylylation of VPg.** *J Virol* 2000, **74(22):**10359-10370.
49. Rieder E, Paul AV, Kim DW, van Boom JH, Wimmer E: **Genetic and biochemical studies of poliovirus cis-acting replication element cre in relation to VPg uridylylation.** *J Virol* 2000, **74(22):**10371-10380.
50. Thiviyanathan V, Yang Y, Kaluarachchi K, Rijnbrand R, Gorenstein DG, Lemon SM: **High-resolution structure of a picornaviral internal cis-acting RNA replication element (cre).** *Proc Natl Acad Sci U S A* 2004, **101(34):**12688-12693.
51. Yang Y, Rijnbrand R, McKnight KL, Wimmer E, Paul A, Martin A, Lemon SM: **Sequence requirements for viral RNA replication and VPg uridylylation directed by the internal cis-acting replication element (cre) of human rhinovirus type 14.** *J Virol* 2002, **76(15):**7485-7494.
52. Andries K, Dewindt B, Snoeks J, Wouters L, Moereels H, Lewi PJ, Janssen PA: **Two groups of rhinoviruses revealed by a panel of antiviral compounds present sequence divergence and differential pathogenicity.** *J Virol* 1990, **64(3):**1117-1123.
53. Morasco BJ, Sharma N, Parilla J, Flanegan JB: **Poliovirus cre(2C)-dependent synthesis of VPgpUpU is required for positive- but not negative-strand RNA synthesis.** *J Virol* 2003, **77(9):**5136-5144.
54. Murray KE, Barton DJ: **Poliovirus CRE-dependent VPg uridylylation is required for positive-strand RNA synthesis but not for negative-strand RNA synthesis.** *J Virol* 2003, **77(8):**4739-4750.
55. Cuervo NS, Guillot S, Romanenkova N, Combiescu M, Aubert-Combiescu A, Seghier M, Caro V, Crainic R, Delpeyroux F: **Genomic features of intertypic recombinant sabin poliovirus strains excreted by primary vaccinees.** *J Virol* 2001, **75(13):**5740-5751.
56. Lukashev AN: **Role of recombination in evolution of enteroviruses.** *Rev Med Virol* 2005, **15(3):**157-167.
57. Simmonds P, Welch J: **Frequency and dynamics of recombination within different species of human enteroviruses.** *J Virol* 2006, **80(1):**483-493.
58. Simmonds P: **Recombination and selection in the evolution of picornaviruses and other Mammalian positive-stranded RNA viruses.** *J Virol* 2006, **80(22):**11124-11140.
59. Santti J, Hyypia T, Kinnunen L, Salminen M: **Evidence of recombination among enteroviruses.** *J Virol* 1999, **73(10):**8741-8749.
60. Deffernez C, Wunderli W, Thomas Y, Yerly S, Perrin L, Kaiser L: **Amplicon sequencing and improved detection of human rhinovirus in respiratory samples.** *J Clin Microbiol* 2004, **42(7):**3212-3218.
61. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16(6):**276-277.
62. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32(5):**1792-1797.
63. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302(1):**205-217.
64. **Zdobnov's Computational Evolutionary Genomics group** [http://cegg.unige.ch/rhinoviruses/]
65. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52(5):**696-704.
66. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8(3):**275-282.
67. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18(3):**502-504.
68. **Belvu Homepage** [http://www.cgb.ki.se/cgb/groups/sonnhammer/Belvu.html]
69. Salemi M, Vandamme AM: **.** In *The Phylogenetic Handbook Volume 1*. Edited by: Salemi M, Vandamme AM. Cambridge University Press; 2003:348.
70. Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs.** *Proc Natl Acad Sci U S A* 2005, **102(7):**2454-2459.
71. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures.** *MonatshChem* 1994:167-188.
72. **Vienna RNA Utilities** [http://www.tbi.univie.ac.at/~ivo/RNA/utils.html]
73. **Swiss Institute of Bioinformatics** [http://www.vital-it.ch/vitalit-intro.htm]

A.2.1.3 *Supplementary information*

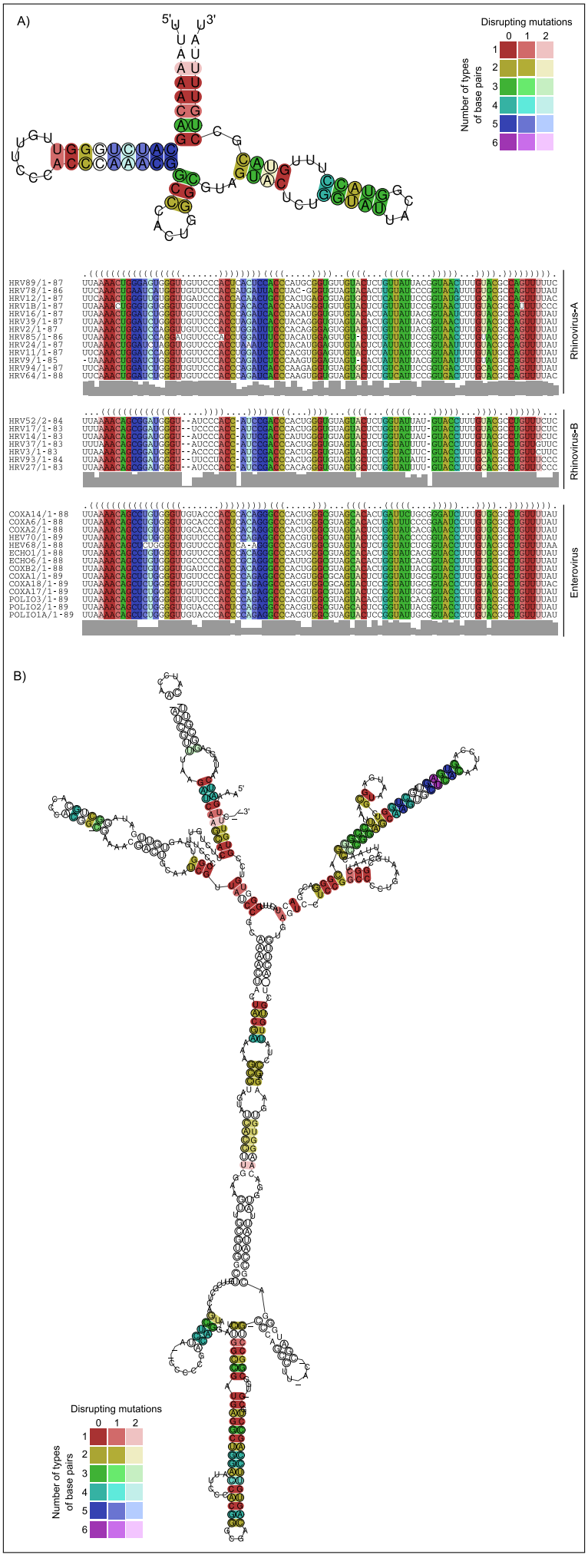
Figure legends for the corresponding original supplementary figures for the paper "New complete genome sequences of human rhinoviruses shed light on their phylogeny and genomic features". Note, only figures produced by me are shown, the original publication contains more supplementary material. Source:
`http://www.biomedcentral.com/1471-2164/8/224/additional/`

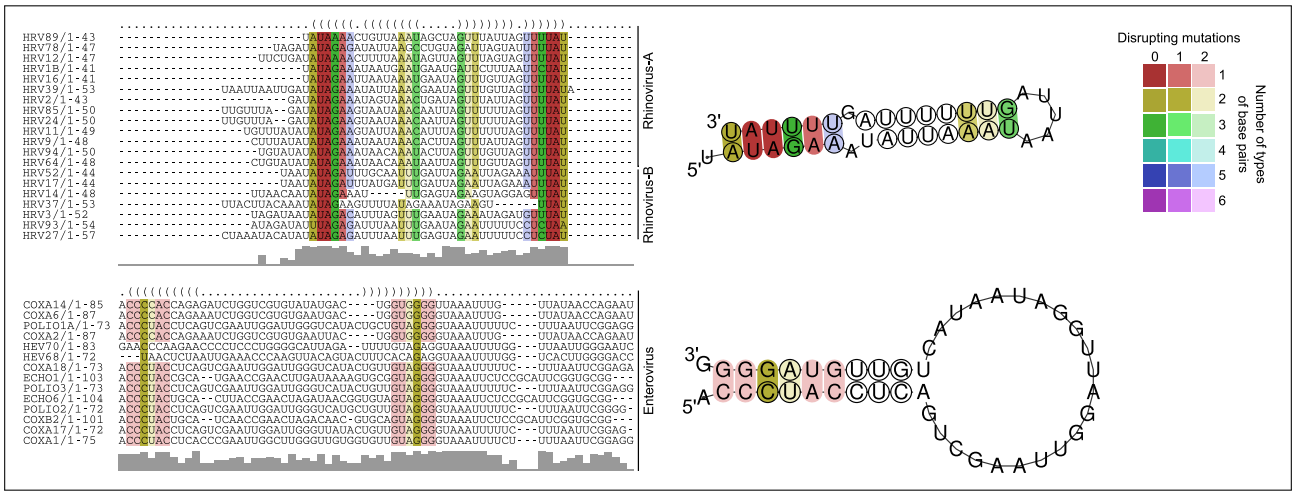
Additional file 3 on the next page 5′ UTR structure conservation. A) 5′ cloverleaf consensus structure for HRV-A, HRV-B and HEV identified by comparative sequence analysis. B) IRES consensus structure for HRV-A, HRV-B and HEV identified by comparative sequence analysis. See legend to Figure 3 for details.

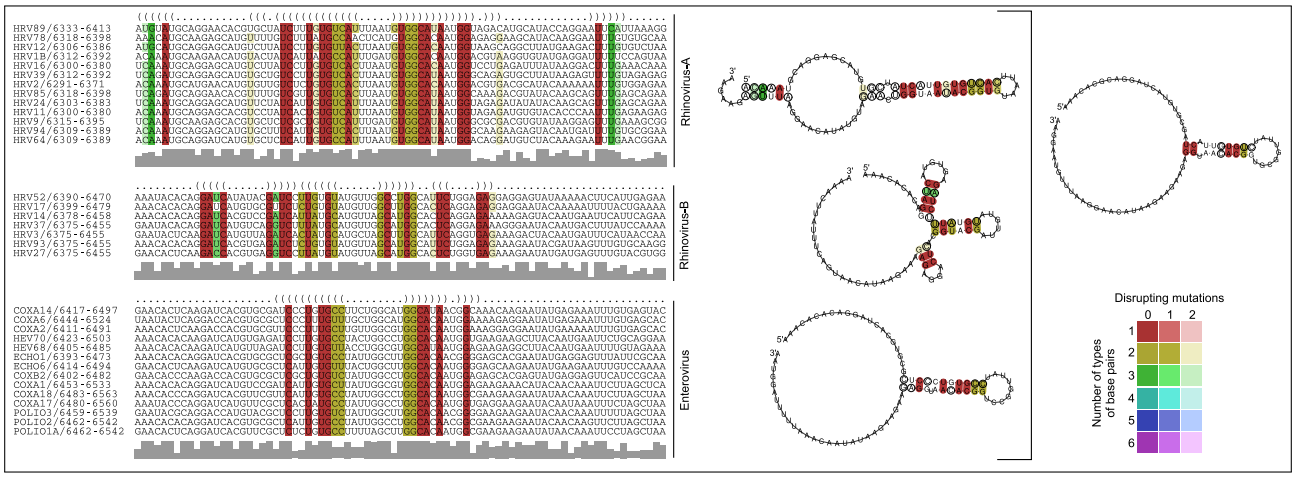
Additional file 4 on page 229 3′ UTR structure conservation. 3′ UTR consensus structure for HRV-A, HRV-B and HEV identified by comparative sequence analysis. See legend to Figure 3 for details.

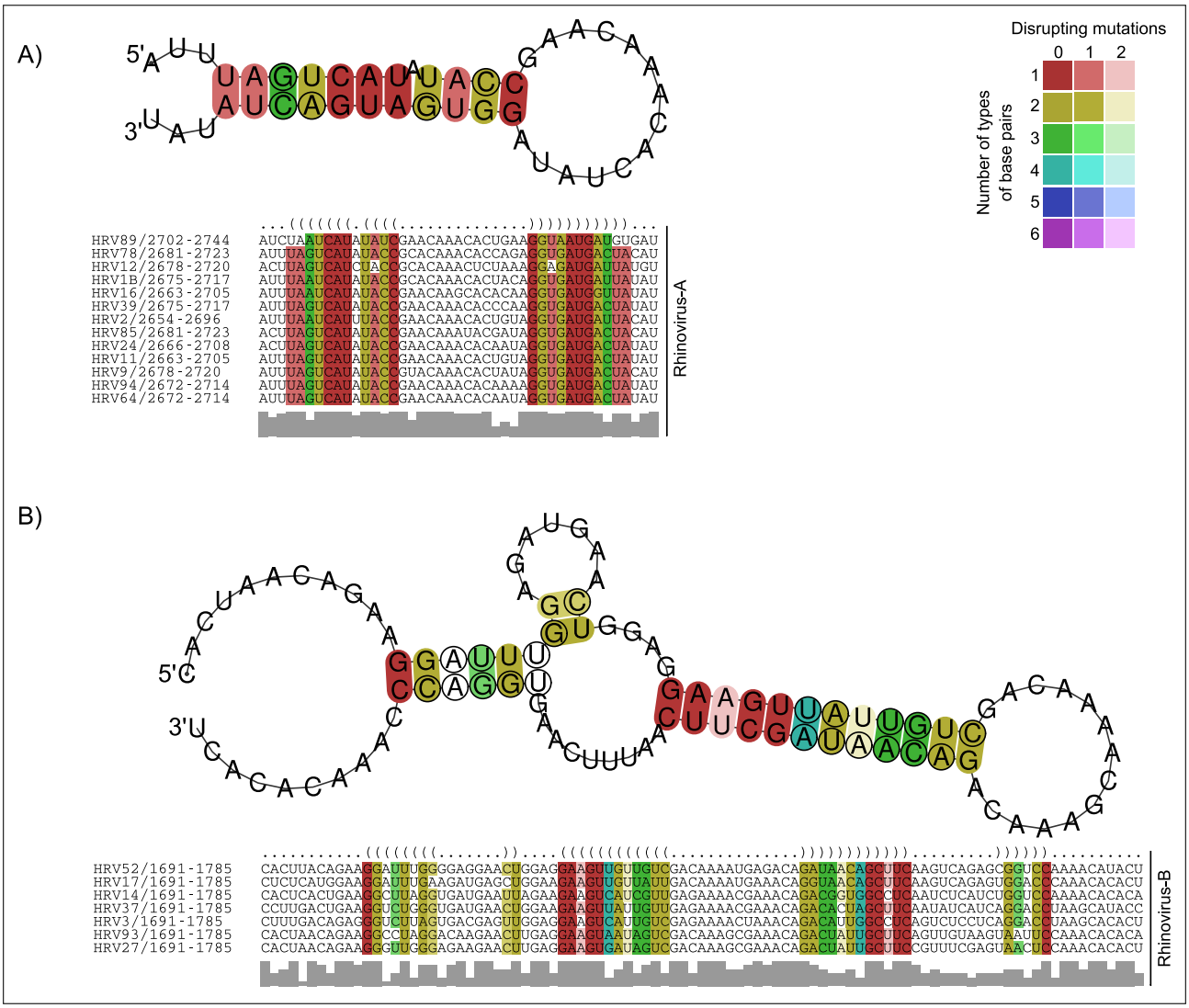
Additional file 5 on page 230 Conserved stem-loop structure in the ORF of HRV-A. Conserved secondary structure located close to the 3′ UTR of HRV-A and corresponding structures in HRV-B and HEV located in the same alignment region. See legend to Figure 3 for details.

Additional file 6 on page 231 Internal cre conservation among HRV-A and HRV-B serotypes. A) Internal 2A cre conservation among HRV-A serotypes. B) Internal VP1 cre conservation among HRV-B serotypes


A)
5'
3'
Disrupting mutations
0 1 2
Number of types of base pairs
1 2 3 4 5 6
HRV89/1-87
HRV78/1-86
HRV12/1-87
HRV1B/1-87
HRV16/1-87
HRV39/1-87
HRV2/1-87
HRV85/1-86
HRV24/1-87
HRV11/1-87
HRV9/1-85
HRV94/1-87
HRV64/1-88
Rhinovirus-A
HRV52/2-84
HRV17/1-83
HRV14/1-83
HRV37/1-83
HRV3/1-83
HRV93/1-84
HRV27/1-83
Rhinovirus-B
COXA14/1-88
COXA6/1-88
COXA2/1-88
HEV70/1-89
HEV68/1-88
ECHO1/1-88
ECHO6/1-88
COXB2/1-88
COXA1/1-89
COXA18/1-89
COXA17/1-89
POLIO3/1-89
POLIO2/1-89
POLIO1A/1-89
Enterovirus
B)
5'
3'
Disrupting mutations
0 1 2
Number of types of base pairs
1 2 3 4 5 6


Rhinovirus-A
Rhinovirus-B
Enterovirus
Disrupting mutations
0 1 2
1
2
3
4
5
6
Number of types of base pairs
3'
5'


Rhinovirus-A
Rhinovirus-B
Enterovirus
Disrupting mutations
0 1 2
Number of types of base pairs
1
2
3
4
5
6


A)
5'
3'
Disrupting mutations
0 1 2
Number of types of base pairs
1 2 3 4 5 6
HRV89/2702-2744
HRV78/2681-2723
HRV12/2678-2720
HRV1B/2675-2717
HRV16/2663-2705
HRV39/2675-2717
HRV2/2654-2696
HRV85/2681-2723
HRV24/2666-2708
HRV11/2663-2705
HRV9/2678-2720
HRV94/2672-2714
HRV64/2672-2714
Rhinovirus-A
B)
5'
3'
HRV52/1691-1785
HRV17/1691-1785
HRV14/1691-1785
HRV37/1691-1785
HRV3/1691-1785
HRV93/1691-1785
HRV27/1691-1785
Rhinovirus-B

A.2.2 *The cis-acting replication elements define human enterovirus and rhinovirus species*

Cordey S*, Gerlach D*, Junier T, Zdobnov EM, Kaiser L, and Tapparel C. The cis-acting replication elements define human enterovirus and rhinovirus species. *RNA* (2008) 14:1568–1578.

A.2.2.1 *Contributions*

Cordey et al., 2008 presents a novel *cre* element for a recently described human rhinovirus species (HRV-A2). They further show by genetic experiments that a previously predicted *cre* element for HRV-B (Tapparel et al., 2007) in non-functional.

My work is covered by the computational part of the project. I contributed in predicting the novel *cre* element HRV-A2 and its subsequent analysis. Figures done by me: Fig. 1, Figs. 3–5, Supplementary Figs. 2–4.

A.2.2.2 *Main paper*

See pages 233–243 or at:
`http://rnajournal.cshlp.org/content/14/8/1568.long`

* These authors contributed equally to the work

The *cis*-acting replication elements define human enterovirus and rhinovirus species

SAMUEL CORDEY,[1,2,6] DANIEL GERLACH,[3,4,6] THOMAS JUNIER,[3,4] EVGENY M. ZDOBNOV,[3,4,5] LAURENT KAISER,[1,2,6] and CAROLINE TAPPAREL[1,2,6]

[1]Central Laboratory of Virology, Division of Infectious Diseases, University of Geneva Hospitals, 1211 Geneva 14, Switzerland
[2]Medical School, University of Geneva, 1211 Geneva 14, Switzerland
[3]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva 14, Switzerland
[4]Swiss Institute of Bioinformatics, 1211 Geneva 14, Switzerland
[5]Imperial College London, South Kensington Campus, London, United Kingdom

ABSTRACT

Replication of picornaviruses is dependent on VPg uridylylation, which is linked to the presence of the internal *cis*-acting replication element (*cre*). *Cre* are located within the sequence encoding polyprotein, yet at distinct positions as demonstrated for poliovirus and coxsackievirus-B3, cardiovirus, and human rhinovirus (HRV-A and HRV-B), overlapping proteins 2C, VP2, 2A, and VP1, respectively. Here we report a novel distinct *cre* element located in the VP2 region of the recently reported HRV-A2 species and provide evolutionary evidence of its functionality. We also experimentally interrogated functionality of recently identified HRV-B *cre* in the 2C region that is orthologous to the human enterovirus (HEV) *cre* and show that it is dispensable for replication and appears to be a nonfunctional evolutionary relic. In addition, our mutational analysis highlights two amino acids in the 2C protein that are crucial for replication. Remarkably, we conclude that each genetic clade of HRV and HEV is characterized by a unique functional *cre* element, where evolutionary success of a new genetic lineage seems to be associated with an invention of a novel *cre* motif and decay of the ancestral one. Therefore, we propose that *cre* element could be considered as an additional criterion for human rhinovirus and enterovirus classification.

Keywords: picornavirus; rhinovirus; *cis*-acting replication element; replication; classification

INTRODUCTION

Human rhinoviruses (HRV), the most frequent cause of respiratory infection (Denny 1995), represent the largest genus of non-enveloped, single positive stranded RNA viruses in the *Picornaviridae* family. Using seroneutralization assays, HRV have been classified into 99 serotypes (Kapikian 1967) and further divided in two different species, HRV-A (74 serotypes) and HRV-B (25 serotypes), based on capsid protein sequences (Ledford et al. 2004; Savolainen et al. 2002). In addition, recent studies have reported new HRV strains clustering into a new HRV species designated as HRV-A2, HRV-C, or HRV-X (Arden et al. 2006; Lamson et al. 2006; Kistler et al. 2007; Lau et al. 2007; Lee et al. 2007; McErlean et al. 2007; Renwick et al. 2007). Similar to all *Picornaviridae* members, HRV genomes of ~7200 base pairs are organized in four different regions: a long 5′-untranslated region (UTR), a single open reading frame (ORF), a short 3′-UTR, and a poly(A) tail (Kitamura et al. 1981).

The 5′- and 3′-UTRs are known to contain important structural motifs. The 5′-UTR contains two highly conserved elements, the 5′-terminal cloverleaf and the internal ribosomal entry site (IRES). The 5′-terminal cloverleaf interacts with the viral protease $3CD^{pro}$ (Andino et al. 1990, 1993; Gamarnik and Andino 1997; Parsley et al. 1997; Rieder et al. 2003) and cellular proteins such as PCBP2 (Andino et al. 1990, 1993; Gamarnik and Andino 1997; Parsley et al. 1997; Perera et al. 2007) to form a ribonucleoprotein complex that is implicated in the switch from viral translation to replication (Huang et al. 2001; Sharma et al. 2005; Perera et al. 2007). This domain is also required both in *cis* and in *trans* for negative (Gamarnik and Andino 1998; Barton et al. 2001) and positive strand initiation (Andino et al. 1990), respectively. In contrast, the IRES

[6]These authors contributed equally to this work.

Reprint requests to: Samuel Cordey, Central Laboratory of Virology, Division of Infectious Diseases, University of Geneva Hospitals, 24 Rue Micheli-du-Crest, 1211 Geneva 14, Switzerland; e-mail: samuel.cordey@hcuge.ch; fax: ++41 22 3724097.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.1031408.

motif promotes initiation of translation from the uncapped viral genome. The 3′-UTR of piconaviruses is required for efficient viral RNA replication (Jacobson et al. 1993; Pilipenko et al. 1996; Melchers et al. 1997; Todd et al. 1997; Duque and Palmenberg 2001; Brown et al. 2004, 2005), although the exact mechanism is not yet understood.

In addition to the structural elements described above, *cis*-acting replication elements (*cre*) have been identified within the coding region of several picornavirus genomes (Fig. 1). The first *cre* motif was originally identified in the VP1 encoding region of HRV14 (a HRV-B member) (McKnight and Lemon 1996, 1998). Similar *cre* elements have been described for different picornaviruses in diverse genome regions such as in 2C for poliovirus (Goodfellow et al. 2000) and coxsackievirus-B3 (van Ooij et al. 2006), 2A for HRV2 (Gerber et al. 2001) and VP2 for Theiler's and Mengo viruses (two cardioviruses) (Lobert et al. 1999). Moreover, the presence of *cre* motif sequences for HRV3 and HRV72 has been identified in the VP1 coding sequence, similar to HRV14 VP1 *cre*, and in 2A for HRV1a and HRV16 serotypes, similar to HRV2 *cre* (McKnight 2003). In contrast to all these previous *cre* elements identified in the polyprotein-encoding region, the foot-and-mouth disease virus (FMDV) presents a *cre* element in the 5′-UTR adjacent to the IRES (Mason et al. 2002). These *cre* exhibit different nucleotide sequences, but are all comparable in size and share a similar stem–loop structure (Yang et al. 2002; Thiviyanathan et al. 2004). Based on sequence comparisons and mutational analysis of entero- and rhinovirus *cre* elements, a common motif, $R^1NNNA_1A_2R^2NNNNNNR^3$, has been proposed for the loop segment of these two genus (Yang et al. 2002; Yin et al. 2003). Extensive studies have been performed to understand the function of these *cre* motifs in virus replication. Both in vivo and in vitro experiments clearly demonstrated the involvement of these *cre* in VPg uridylylation (Paul et al. 1998; Rieder et al. 2000; Gerber et al. 2001; Yang et al. 2002; Yin et al. 2003; Richards et al. 2006). VPgpUpU is assumed to serve as primer for the viral polymerase 3Dpol in RNA synthesis. Not only does the loop appear to be important for the function of *cre* motif, the stem also seems to play a crucial role (Yin et al. 2003; Pathak et al. 2007). Since VPg is linked to the 5′-end of the positive and negative strand genomes, it is tempting to postulate that *cre* motifs are involved in the initiation of synthesis of both strands. However, this remains a subject of controversy, since some studies have argued in favor of a role of *cre* motif only in positive strand RNA synthesis (Goodfellow et al. 2003; Morasco et al. 2003), while others established an involvement of *cre* in negative sense synthesis (McKnight and Lemon 1998; Yang et al. 2002). More surprising were the observations that heterologous exchanges of *cre* elements were permissive in some cases (Gerber et al. 2001; Yin et al. 2003; Yang et al. 2004; Shen et al. 2007), thus suggesting that their roles in RNA replication are therefore independent of their location along the genome.

In a previous study, we observed that HRV-B and human enterovirus (HEV) were more closely related to each other than either is to HRV-A for 2A, 2B, 2C, and 3D proteins (Tapparel et al. 2007). This was explained as the result of hypothetical recombination events between HRV-A and the HEV ancestor of HRV-B soon after the divergence of HEV and HRV-A (Supplemental Fig. 1). Moreover, a putative 2C *cre* motif absent in all HRV-A genomes was identified in the 25 HRV-B serotypes that conserved significantly essential nucleotides within the consensus loop ($R^1NNNA_1A_2R^2NNNNNNR^3$). As a functional *cre* motif was already known in HRV14 VP1 protein (HRV-B), the first goal of this study was to determine whether this putative *cre* located within the HRV14 2C coding region was required for virus replication. By mutational analysis and creation of 2C *cre* disrupted-mutant (DM), we provide strong evidence that this putative 2C *cre* motif is nonfunctional for virus replication and appears to be an evolutionary relic. Interestingly, we pointed out two amino acids within the *cre* region of the 2C protein that are absolutely necessary for HRV14 replication. Moreover, by analyzing sequences of newly reported HRV strains, we were able to identify a unique and conserved VP2 *cre* motif present in all new HRV-A2 members, but absent in both HRV-A and HRV-B species.

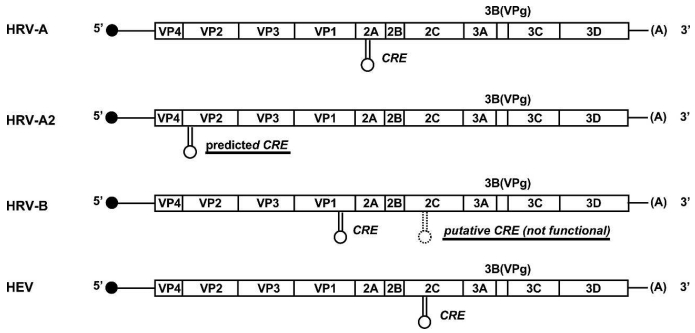


FIGURE 1. Schematic diagrams of human rhinovirus (HRV) and enterovirus (HEV) genomic organization showing the location of known *cre* elements in HEV, HRV-A, and HRV-B as well as newly predicted HRV-A2 *cre* and putative HRV-B *cre* (underlined) that are characterized in the text.

RESULTS

Here we discuss internal *cre* of human enterovirus and rhinovirus that are

essential for virus replication (Fig. 1) and provide experimental evidence of nonfunctionality of an additional putative HRV14 *cre* in the 2C region and evolutionary evidence of a novel predicted *cre* element located in the VP2 region of the recently reported HRV-A2 strains.

Mutations within the HRV14 2C *cre* loop affect virus growth

We previously identified a putative *cre* element in the HRV-B 2C coding sequence in addition to the already characterized HRV14 VP1 *cre* motif by full-length genome comparison of HRV-A, HRV-B, and HEV (Tapparel et al. 2007). In an attempt to determine the functionality of this second *cre* motif, we performed a mutational analysis in the corresponding region of the HRV14 infectious clone. Since nucleotides A_1, A_2, R^2, and R^3, known as essential for *cre* motif function, were conserved in HRV14 2C *cre*, as in all HRV-B (unlike R^1) (Tapparel et al. 2007), these positions were mutated as described (Fig. 2A,B). The effect of these mutations on HRV14 replication was then measured by immunofluorescence 12 h post-infection and confirmed by in situ hybridization with antisense RNA probe (data not shown).

As shown in Figure 2, A4277C (Lys → Gln), A4278C (Lys → Thr), A4279C (Lys → Asn), and A4286T (Thr → Ser) caused a decrease of six- to sevenfold in viral replication. The replication is further reduced (30- to 50-fold) with the following mutations: A4277G (Lys → Glu), A4278T (Lys → Ile), AA4277/8GG (Lys → Gly), A4286C (Thr → Pro), and A4286G (Thr → Ala). In contrast, A4279G does not change viral replication efficiency (Fig. 2A). Of note, none of these mutations disrupts the classical hairpin structure (data not shown). Taking into consideration the critical positions within the loop, the results were as expected apart from A4278G (Lys → Arg) and A4286G (Thr → Ala). Indeed, A4278G (Lys → Arg) changes the critical A_2 residue but replication is as efficient as wild type, whereas A4286G (Thr → Ala) respects the R^3 position but abrogates viral replication. The explanation for this can be found when observing the 2C amino acid sequence. All mutations changing the amino acid encoded by nucleotides $A_1A_2R^2$ (originally a Lys) in a nonconservative way decrease replication, whereas the only conservative change (A4278G [Lys → Arg]) has no effect. Similarly, R^3 is the first nucleotide of a codon encoding for a threonine in HRV14, and replacement of this threonine with any other amino acid disrupts replication such as mutations A4286G (Thr → Ala), A4286C (Thr → Pro), and A4286T (Thr → Ser). As all except one mutation (A4279G) changed the amino acid sequence within the 2C protein, the interpretation of results was complicated since any variations in virus growth could be caused by changes either in the 2C protein or in the putative 2C *cre* motif. Of note, sequencing of mutants after growth allowed us to exclude any reversion.

HRV14 replication is not affected by disrupting the putative 2C *cre* stem–loop

To firmly conclude that the putative HRV14 2C *cre* motif is not necessary for viral replication, we introduced 12 silent mutations disrupting HRV14 2C *cre* classical stem–loop (Fig. 2D, HRV14-DM). This disruption renders unlikely any use of this motif for VPg uridylylation by 3Dpol without altering the resulting coded viral protein product. Creation of disrupted *cre* mutants was successfully applied previously, notably in HRV14 VP1 *cre* (Yang et al. 2004) and coxsackievirus-B3 2C *cre* functional studies (van Ooij et al. 2006). After quantification of virus growth by immunofluorescence, we observed similar replication levels for HRV14-DM and HRV14-WT (Fig. 2C-E). Again, the absence of reversion was confirmed by sequencing.

In conclusion, the mutageneses approach together with the disruption mutant show that the putative HRV14 2C *cre* element is dispensable for viral replication, but the amino acid sequence is critical. In particular, we identified two amino acids (corresponding to 118th and 123rd amino acids in 2C coding sequence) essential for replication, since their individual changes resulted in a strong decrease in HRV14 replication efficiency.

Comparison of 2C *cre* structure between HEV and HRV-B

The absence of the putative HRV14 2C *cre* function suggests that the ancestral HRV-B 2C *cre* motif, probably originating from HEV 2C *cre* (Tapparel et al. 2007), has evolved into a nonfunctional motif and was replaced by a functional VP1 *cre* motif. To confirm this hypothesis and to determine if "intermediate" less conserved *cre* motifs could also exist within the enterovirus genus, we compared the evolution rate and conservation of the 2C *cre* motif among HEV and HRV-B. For this purpose, we calculated a structure-based phylogenetic tree of all HEV and HRV-B 2C *cre* elements. Our phylogenic analysis includes 86 publicly available full-length HEV (http://www.picornaviridae.com/sequences/sequences.htm) and the 25 HRV-B serotype sequences and could elucidate the evolutionary pathways differentiating these two groups. The tree depicted in Figure 3 represents a structural clustering tree based on the conserved secondary structures of 2C *cre* for each group and the measurement of structural distance between these groups. The tree topology clearly shows a highly conserved cluster for HEV 2C *cre* structures and five loosely conserved smaller clusters for the putative HRV-B 2C *cre* structures without any overlap between HEV and HRV-B members. Apparently, evolutionary pressure on the conservation of the HEV 2C *cre* stem–loop is much higher than for the HRV-B 2C *cre*. The length of terminal branches also presents important differences between the HEV and the HRV-B clusters. The mean structural distance between any of the HEV structures is smaller than for the HRV-B

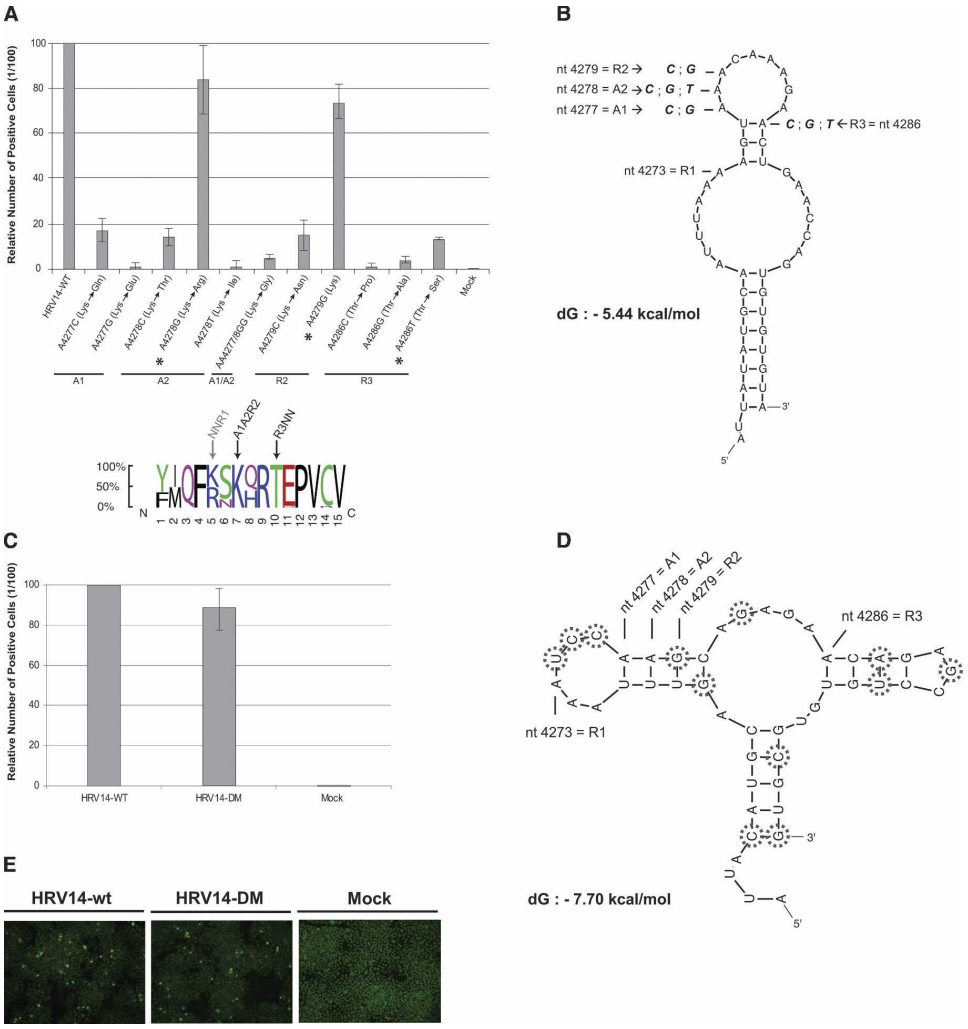


FIGURE 2. Effects of nucleotide changes within the putative HRV14 2C *cre* motif on viral growth. (*A*) Quantification of mutant virus growth in HeLa cells. Each HRV14 2C *cre* mutant is named according to the substituted nucleotide. The amino acid changes that follow the nucleotide mutations are indicated in parentheses after the mutation and those conserving the amino acid properties are marked by an asterisk. The frequency plot of amino acid residue conservation in 25 HRV-B serotypes is shown at the foot of Panel *A* (http://weblogo.berkeley.edu/). Quantification of virus growth was measured by immunofluorescence 12 h post-infection and expressed as the mean of positive cells per total cells and expressed relative to that of HRV14-WT. (*B*) Schematic representation of the putative HRV14 2C *cre* structure as predicted by MFOLD from nucleotides 4256 to 4303. Nucleotide substitutions at position 4277, 4278, 4279, and 4286 (corresponding, respectively, to A_1, A_2, R^2, and R^3 positions within the consensus $R^1NNNA_1A_2R^2NNNNNNR^3$ sequence) are represented in bold. (*C*) Quantification of virus growth for HRV14-DM versus HRV14-WT. Measurements were conducted by immunofluoresence as described in *A*. (*D*) Schematic representation of 2C *cre* motif for HRV14-DM (nucleotides 4256–4303) following the introduction of 12 silent mutations (surrounded nucleotides) as predicted by MFOLD. None of the crucial positions within the consensus *cre* loop sequence were mutated, except one at position 4279 (A → G) known to be permissive for efficient HRV14 replication (*A*). (*E*) HeLa cells infected by HRV14-WT, HRV14-DM, or Mock. Infected positive cells were detected by immunofluorescence (IF).

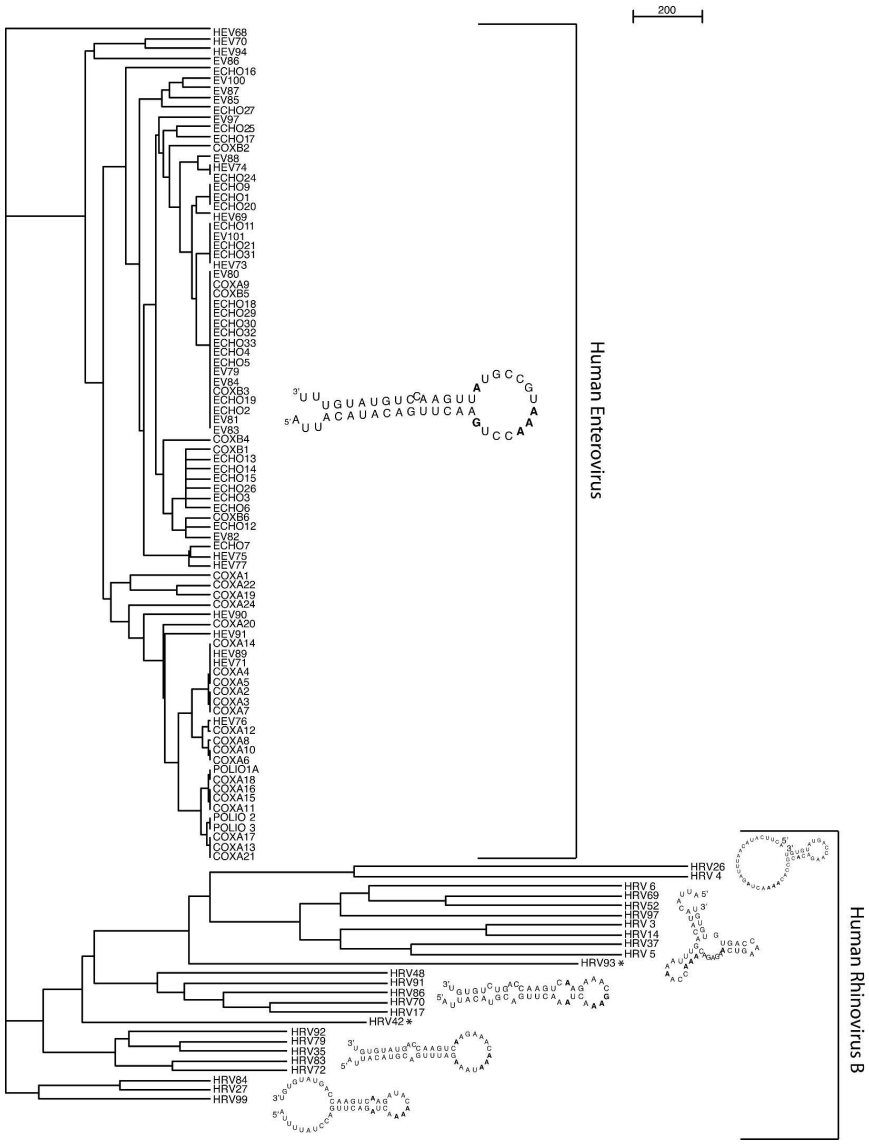


FIGURE 3. WPGMA structure-based cluster tree for 86 HEV and 25 HRV-B 2C *cre* elements. A consensus 2C *cre* structure is shown for each cluster except for HRV42 and HRV93, which do not belong to any specific cluster (marked *). The crucial positions within the consensus *cre* structures are marked by bold letters for each cluster.

sequences. In addition, no intermediate strain exists that would result in a clustering of a HRV-B strain within the HEV cluster. From these data, we conclude that there is a close link between a functional *cre* element and the short mean structural distance of that element within a group. This is further confirmed when looking at the structure-based phylogenetic trees of the two functional *cre* of HRV-A 2A and HRV-B VP1 (Supplemental Figs. 2, 4).

Taken together, these results support those obtained with HRV14-DM and the conclusion that the putative HRV14 2C *cre* is nonfunctional. In the absence of functional pressure, this motif structure evolves more rapidly within HRV-B. In addition, the tight structural cluster for 2C *cre* among HEV members, VP1 and 2A *cre* among HRV-B and HRV-A, respectively, suggests that a selective pressure, most likely related to the function of this motif, is present across all HEV, HRV-A, and HRV-B members.

Identification of a conserved predicted *cre* motif in the HRV-A2 VP2 coding region

New HRV strains diverging significantly from HRV-A and HRV-B species were identified recently. Different nomenclature names (HRV-A2, HRV-X, HRV-QPM, or HRV-C) have been assigned to these strains and we will use the HRV-A2 denomination for consistency. Indeed, whole genome comparison shows that HRV-A2 species is distinct but close to HRV-A (McErlean et al. 2008). Using the RNAz program, we scanned the six full-length genomes available for these new strains in order to localize their *cre* elements. In addition to the 5′-terminal cloverleaf, the IRES, and the 3′-UTR hairpin elements constantly found in picornaviruses (data not shown), we identified the presence of a unique and conserved predicted *cre* structure within the VP2 coding region for the six new HRV genomes analyzed (Fig. 4), and not within 2A or VP1 as is the case for HRV-A and HRV-B, respectively. The presence of this new predicted *cre* structure was then confirmed in the VP2 coding region for additional HRV-A2 partial sequences recently published (Kistler et al. 2007; McErlean et al. 2007). Although not yet investigated due to the lack of infectious clones and the inability of these new viruses to grow in standard cell culture, the probability that this predicted VP2 *cre* motif acts as a nondispensable functional element is extremely high. This assumption is based on four main observations: first, this element is only found in VP2 and matches perfectly the classical *cre* hairpin structure with a loop region of exactly 14 nucleotides (nt); second, the critical R^1, A_1, A_2, R^2, and R^3 nucleotides within the consensus *cre* sequence are conserved for all HRV-A2 analyzed; third, similar to the observations made for the HEV 2C *cre* (Fig. 3), HRV-A 2A *cre*, and HRV-B VP1 *cre* (Supplemental Figs. 2, 4), the structural tree of HRV-A2 VP2 *cre* presents very short structural distances between the individual strains elements (Supplemental Fig. 3). Fourth, the mean base pair distance, defined as the number of base pairs to change to transform one structure into another, for HRV-A2 VP2 *cre*, is as low as for the other functional *cre* elements (HEV 2C *cre*: 2.97, HRV-A 2A *cre*: 2.39, putative HRV-B 2C *cre*: 16.27, HRV-B VP1 *cre*: 5.15, predicted HRV-A2 VP2 *cre*: 1.81). According to the structure trees, this also shows that the *cre* structures of all HRV-A2 strains are very similar.

In summary, each HRV species identified to date, and from whom a full-length genome is available, possesses a

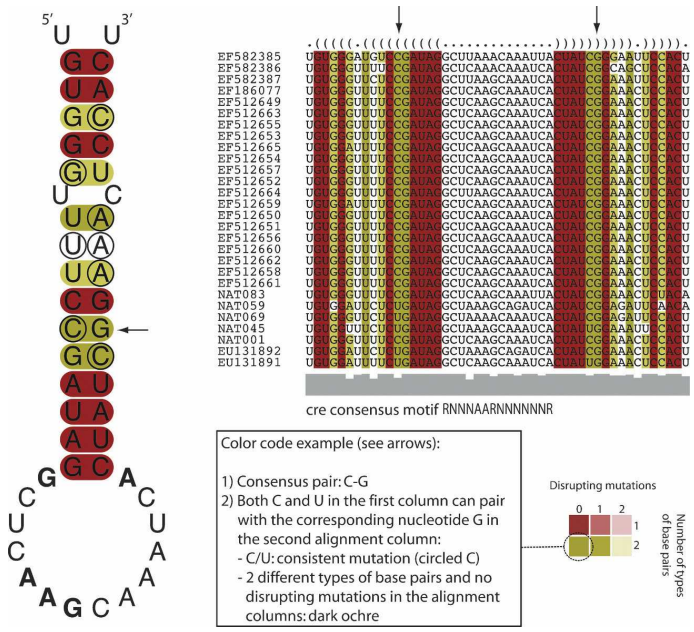


FIGURE 4. Conservation of a predicted VP2 *cre* secondary structure for HRV-A2. Multiple sequence alignment across all available full-length HRV-A2 (NAT001, NAT045, EF186077, and EF582385-7) and VP2 partial sequences (Nat059, NAT069, Nat083, EU131891-2, and EF512649-65) showing a consensus secondary *cre* structure in VP2. The structure is shown in the dot-bracket format *above* alignment. Each corresponding bracket represents a consensus base pair of the alignment columns *beneath*. Sequences are color-coded according to consistent and compensatory mutations in the aligned sequences regarding the conserved structure (see figure text box). The sequence conservation profile is shown in gray bars *below* the alignment. The secondary structure of the conserved predicted *cre* VP2, color-coded according to the different types of base pairs in the corresponding alignment columns, is shown on the *left* side. The conserved *cre* motif nucleotides are marked in bold. Note that the amino acid sequence corresponding to the loop region is almost 100% conserved in all species (C-G-F-S-D-R-L-K-Q-I-T-I-G/N-S-T). Mutations supporting the structure (consistent mutations) occur almost exclusively at the third codon position, which leads to synonymous codons and the amino acid conservation.

functional *cre* element located at different, but specific, positions (Fig. 1).

Boxplot analysis

The analysis of the distances between the terminal nodes in the WPGMA structural cluster trees of HRV-A 2A *cre*, putative HRV-B 2C *cre*, HRV-B VP1 *cre*, predicted HRV-A2 VP2 *cre*, and HEV 2C *cre* shows striking differences between these elements (Fig. 5). The size of the distance distribution as well as its mean describe how "tight" the individual *cre* are clustered. In other words the more similar the elements are to each other within one species the more conserved are the secondary structures between the strains. The experimental evidence for the nonfunctionality of the putative HRV-B 2C *cre* element is strengthened by the data of the distance distributions in the boxplot for this element. There is a large variety in distances as well as a high mean distance between single strain elements, which means that these structures do not form a species-specific evolutionary conserved secondary structure. On the other hand, the newly predicted *cre* element for HRV-A2 VP2 shows a very narrow distance distribution for its members as well as the lowest mean structural distance compared to all the other functional *cre* elements. It is therefore very likely that this predicted HRV-A2 VP2 *cre* is functional, although further genetic and biochemical studies will be necessary to finally confirm this assumption.

DISCUSSION

The presence of internal *cre* motif was first described for HRV14 (McKnight and Lemon 1996), and the requirement of this motif for viral RNA replication came as a surprising observation. Indeed, for many years, it was considered that the different structures previously identified within the 5′- and 3′-UTR of rhino- and enteroviruses were sufficient for this function. Since then, *cre* motifs have been identified within the protein-coding sequence of poliovirus, coxsackievirus-B3, HRV2, Theiler's, and Mengo viruses (Lobert et al. 1999; Goodfellow et al. 2000; Gerber et al. 2001; van Ooij et al. 2006). Although these *cre* elements were positioned in different regions according to the different viruses (Fig. 1), their classical hairpin structure of ~60 nt was demonstrated as essential for the VPg uridylylation process. Uridylylated VPg then serves as primer at the genome 3′ extremities for RNA replication. Whether it is necessary for replication of both strands or for only one strand remains an open question.

In a recent study, we identified the presence of a new putative *cre* motif within the HRV14 2C coding region in addition to the one already described in VP1 (Tapparel et al. 2007). Although it would be surprising that HRV14 required the presence of two functional *cre* motifs for its genome replication, we cannot rule out that these two elements may have complementary roles in RNA replication, one being committed to the synthesis of the positive strain and the other to the negative synthesis. Both the approaches of point mutation of the critical putative 2C *cre* motif residues and the total disruption of the *cre* structure give a clear indication that the putative HRV14 2C *cre* motif is nonfunctional. Indeed, derivatives with disrupted motif but conserved amino acid sequences replicate at levels equivalent to wild-type virus. We recently showed that HEV and HRV-B were more closely related to each other than either is to HRV-A for 2A, 2B, 2C, and 3Dpol coding regions (Tapparel et al. 2007). Thus, it can be postulated that the presence of the putative 2C *cre* motif in all HRV-B strains may only result from a leftover from HEV and HRV-A early recombination events. The topology of

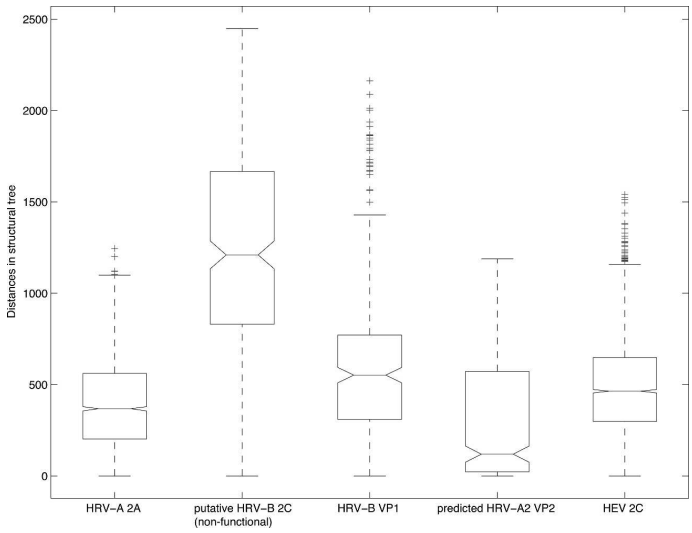


FIGURE 5. Boxplots of structural distances within the respective WPGMA trees of HRV-A 2A *cre*, putative HRV-B 2C *cre*, HRV-B VP1 *cre*, predicted HRV-A2 VP2 *cre*, and HEV 2C *cre*. The mean distances are 400, 1229, 654, 282, and 475, respectively. Note that the nonfunctional putative HRV-B 2C *cre* shows the largest distances, thus showing that the structures of the individual *cre* sequences of these species do not form a tight cluster of mostly similar structures as is the case for the other *cre* motifs. The distance distributions of all the other elements are below that of the nonfunctional *cre* element. The distances are directly related to the scores for the clusters from the LocARNA package.

the HRV-B and HEV 2C *cre* structural trees further corroborates this hypothesis. Whereas the evolution of the HEV 2C *cre* is very slow, the putative HRV-B 2C *cre* tree shows rapid evolution. Of note, some HRV serotypes (HRV92, HRV79, HRV35, HRV83, and HRV72) exhibit a better classical 2C *cre* stem–loop structure than the one displayed by HRV14, and we cannot totally exclude that the putative HRV-B 2C *cre* is absolutely nonfunctional for these specific serotypes. Nevertheless, this hypothesis seems highly unlikely based on the high conservation of the HRV-B VP1 structural tree throughout all HRV-B members. Finally, our mutational analysis revealed the presence of two essential amino acids located between positions 4277–9 and 4286–8, corresponding to the 118th and 123rd amino acids within the HRV14 2C coding sequence, respectively. The presence of these two amino acids lying exactly within the putative HRV14 2C *cre* loop sequence may explain in part the conservation of such motifs, as their mutation may strongly affect one of the different 2C protein functions and, consequently, HRV14 replication.

Protein 2C (also known as 2CATP^{ase}) is highly conserved among all picornaviruses and has been shown to be required for several steps during virus infection. Indeed, 2C is known to be involved in viral RNA replication, encapsidation, and intracellular membrane remodeling (Cho et al. 1994; Pfister et al. 2000; Suhy et al. 2000; Teterina et al. 2006). This protein consists of three different parts. The central domain, where the putative HRV14 2C *cre* is located, contains the nucleoside triphosphate (NTP) binding motif, including motifs A and B known to contribute to binding the phosphate moiety of NTP (Gorbalenya et al. 1990; Mirzayan and Wimmer 1992) and to NTP hydrolysis (Gorbalenya et al. 1990; Rodriguez and Carrasco 1993; Mirzayan and Wimmer 1994; Samuilova et al. 2006), respectively. HRV14 2C *cre* is not situated within a functionally known motif, but only 9 nt upstream of the A motif. Our results suggest that amino acids 118 and 123 of the *cre* motif are critical for one of the 2C protein functions, but further experiments are required to define their precise roles.

The whole-genome scan of new HRV strains resulted in the discovery of a new predicted *cre* motif located within the VP2 protein-coding region. Based on the fact that (1) this *cre* motif is not found elsewhere along the genome; (2) it perfectly respects the stem–loop structure including the R^1, A_1, A_2, R^2, and R^3 positions; and (3) the structural tree of the predicted HRV-A2 VP2 *cre* shows no important discrepancies in branch length, similarly to those of HRV-B VP1 *cre*, HRV-A 2A *cre*, and HEV 2C *cre* structures, we assume that this classical hairpin structure is essential for HRV-A2 replication. The validity of our newly predicted *cre* in HRV-A2 is strengthened by the following points: This study used more sequences than the previous one (Tapparel et al. 2007) to predict the putative new HRV-A2 VP2 *cre*. Furthermore, the mean base pair distance between all the individual strain structures is as low as for the other *cre* motifs known to be functional. This reflects the high evolutionary constraints on this locus in keeping with its stable stem–loop structure and is additionally supported by consistent nucleotide mutations in the stem region. This point is also strengthened by the fact that all HRV-A2 strains form one common structure cluster in the tree (Supplemental Fig. 3), which was not the case for the putative HRV-B 2C *cre* (Fig. 3). This type of analysis was not used in our previous paper, but points out that a limited number of species and an exclusive view on the consensus structure can lead to a prediction of a *cre* that is shown here to be nonfunctional in HRV14. However, the most important difference in the putative HRV-B 2C *cre* and the predicted HRV-A2 VP2 *cre* stems from the fact that in HRV-B an experimentally verified *cre* was already known, whereas no *cre* element has been described in HRV-A2 until now.

This new rhinovirus species was previously identified based on VP4-VP2 and VP1 phylogenetic analysis. In addition to their differences in capsid sequences, HRV-A, HRV-B, and HRV-A2 are thus distinguishable from each other by the localization of their respective *cre* motifs. This signifies that the classification of HRV genomes based on their capsid sequence matches the classification based on *cre* location. Following this observation, it will be interesting to scan the genome of the other newly identified rhinoviruses whose full-length genome sequences are not yet available (Lee et al. 2007) to find the location of their *cre* elements. Using *cre* as an additional classification criterion, rhinovirus A, A2, and B could be considered as three independent species, whereas the four enterovirus species could be reclassified into one unique species. Interestingly, this observation correlates with the homology comparison made between each HEV and HRV species. Indeed, the percentages of homologies between each HEV species are significantly higher than those between each HRV species at both nucleotide (full-length genome) and amino acid levels (Supplemental Fig. 5). Beyond the classification per se, our findings suggest that secondary functional structures are key elements in shaping the evolutionary pathways of human picornaviruses and that each human genogroup evolves independently within the borders determined by these nondispensable constraints. When we extended our computational analysis with structural constraints based on HEV and HRV *cre* elements to the whole full-length picornavirus sequences available, no classical *cre* motif could be observed for the other *Picornaviridae* (data not shown). This suggests that either the structure of this element strongly diverges among other picornavirus genera, as already observed for cardiovirus (Lobert et al. 1999) and FMDV *cre* (Mason et al. 2002) or that *cre* is only present in some *Picornaviridae*. However, the latter argument is unlikely due to the presence of the VPg protein in all *Picornaviridae* members. Finally, our data confirm that

cre are functionally independent of their position along the genome. However, the driving force that directs *cre* location for each species remains an open question, but, clearly, the localization of this structure on the genome results from specific constraints.

In conclusion, our analysis demonstrated that the conserved putative HRV14 2C *cre* found also in all HRV-B serotypes is nonfunctional and likely represents an evolutionary residue from the HEV and HRV-B common ancestor. This study clearly shows the importance of the structural constraints for entero- and rhinovirus *cre* functionality. Moreover, we were able to identify a highly conserved *cre* in the VP2 protein-coding region of newly identified HRV-A2 strains. Beyond the potential usefulness for picornavirus classification, our observation highlights *cre* motifs as key determinants in shaping human picornavirus evolutionary ability. Finally, we propose to use variations in *cre* locations as an additional criterion facilitating human rhinovirus and enterovirus classification.

MATERIALS AND METHODS

Cell and media

HeLa-OH cells were grown in Eagle's Minimum Essential Medium (EMEM; Lonza) supplemented with 2 mM L-glutamine, 1 μg/mL amphotericin, 100 μg/mL gentamicin, 20 μg/mL vancomycin, 10% fetal calf serum (FCS) at 37°C in a 5% CO_2-containing atmosphere.

Construction of 2C *cre* mutants

A PCR fragment of 1602 nt containing HRV14 2C coding sequence amplified with the forward primer 5′-GGCATTCA GAATAGTAAATGAACATG-3′ and the reverse primer 5′-GTTGGGGGGCTTAGTGTGTT-3′ from the plasmid pWR3.26-HRV14 (HRV14 full-length sequence under the T7 promoter control, kindly provided by Wai-Ming Lee [University of Wisconsin]) was subcloned into the plasmid pCR2.1-TOPO (Invitrogen). The resulting plasmid was named pCR2.1-TOPO-HRV14 and used to perform the mutagenesis with the Quickchange site-directed mutagenesis kit (Stratagene). The primers used for mutagenesis are listed in Supplemental Fig. 6. The mutated fragments were then excised at AvrII and XcmI unique restriction sites and ligated back into pWR3.26-HRV14. The mutation was confirmed by sequencing.

In vitro transcription and transfection

Twenty micrograms of plasmid pWR3.26-HRV14 harboring wt or putative HRV14 2C *cre* mutants were linearized at a unique MluI restriction site downstream of the 3′-viral poly(A) genome. Using T7 RNA polymerase, RNA transcripts were synthesized from the linear templates with the MEGAscript T7 kit (Ambion) 3 h at 37°C and purified with the RNeasy Mini Kit (Qiagen). Transcript RNA was quantified and checked by 0.1% sodium dodecyl sulfate–1% agarose gel analysis. HeLa-OH cells were seeded at 6×10^5 cells in 35-mm wells of a six-well plate. The following day, cells were transfected with 2 μg of RNA transcripts containing wt or mutant HRV14 2C *cre* sequence using the TransMessenger Transfection Reagent kit (Qiagen). After 3 h at 37°C, 2 mL of McCoy's 5A Medium (Invitrogen)–2% FCS was used to replace the transfection mix. Cells were then incubated at 33°C for 48 h.

Quantification of virus growth

Virus growth was measured by immunofluorescence. Two days post-transfection, cell supernatant was recovered and clarified 5 min at 1000 rpm in a Multifuge 4 KR. Two hundred microliters of clarified supernatant were mixed with 2 mL of McCoy's 5A Medium–2% FCS to infect HeLa cells grown overnight on 22 mm × 22 mm coverslips in 33-mm wells. Virus growth of wt and mutant HRV14 was finally quantified by immunofluorescence 12 h post-infection. The cells were washed twice with PBS lacking Ca^{2+} and Mg^{2+} (PBS^-) and fixed 1.5 h in acetone at −20°C. Cells were air-dried for a few minutes at room temperature before incubation with the primary antibody, a rabbit anti-HRV14 serum (ATCC number: VR-284AS/Rb, diluted 1/1000 in PBS^-–1% BSA), for 45 min at 37°C in a humidity chamber. After intensive washing with PBS^-, Alexa Fluor 488 goat anti-rabbit antibody (Invitrogen) was added and the cells were incubated for 45 min at 37°C in the dark. After final rinsing with PBS^-, coverslips were mounted in fluorotec embedding medium (Bio-Science products AG). Quantification of virus growth was calculated as the percentage of positive cells in three independent experiments.

Comparative sequence analysis

The virus sequences were aligned using MAFFT version 6.240 (Katoh et al. 2002). All alignments were then analyzed with RNAz version 1.0 (Washietl et al. 2005). The conserved secondary structure elements found by the program were manually investigated for the presence of the typical 14-nt *cre* hairpin–loop as well as the conserved sequence motif. Subregions of the alignments were extracted using the EMBOSS package version 5.0.0 (Rice et al. 2000) and a consensus sequence and structure was calculated with RNAalifold, which is part of the Vienna Package version 1.7. The trees for Figure 3 and Supplemental Figures 2, 3, and 4 were calculated using multiple sequence alignments of the respective *cre* motifs and the mlocARNA pipeline version 1.0 (Will et al. 2007) for structural based cluster trees. Secondary structures and energies for individual *cre* sequences were computed using the RNAfold program (part of the Vienna Package) and the program MFOLD version 3.2 (Zuker 2003). The boxplots in Figure 5 were created using the pairwise distances between every terminal node in every single weighted pair group method with averaging (WPGMA) tree of the species-specific *cre* elements from mlocARNA. The length distributions were plotted group wise using Matlab 7.2.0.283 (Mathworks, http://www.mathworks.com).

SUPPLEMENTAL DATA

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

We would like to thank Lara Turin, Sandra Frayard Van Bell, and Chantal Gaille for technical assistance and Luc Perrin for support

and comments on the manuscript. We also thank Rosemary Sudan for editorial assistance. This study was supported by the Swiss National Science Foundation (No 3200B0-101670 to L.K. and 3100A0112588/I to E.M.Z.), the Canton of Geneva, and the University of Geneva Dean's programme for the promotion of women in science (C.T.).

Received February 13, 2008; accepted April 24, 2008.

REFERENCES

Andino, R., Rieckhof, G.E., and Baltimore, D. 1990. A functional ribonucleoprotein complex forms around the 5′-end of poliovirus RNA. *Cell* **63:** 369–380.

Andino, R., Rieckhof, G.E., Achacoso, P.L., and Baltimore, D. 1993. Poliovirus RNA synthesis utilizes an RNP complex formed around the 5′-end of viral RNA. *EMBO J.* **12:** 3587–3598.

Arden, K.E., McErlean, P., Nissen, M.D., Sloots, T.P., and Mackay, I.M. 2006. Frequent detection of human rhinoviruses, paramyxoviruses, coronaviruses, and bocavirus during acute respiratory tract infections. *J. Med. Virol.* **78:** 1232–1240.

Barton, D.J., O'Donnell, B.J., and Flanegan, J.B. 2001. 5′ cloverleaf in poliovirus RNA is a *cis*-acting replication element required for negative-strand synthesis. *EMBO J.* **20:** 1439–1448.

Brown, D.M., Kauder, S.E., Cornell, C.T., Jang, G.M., Racaniello, V.R., and Semler, B.L. 2004. Cell-dependent role for the poliovirus 3′-noncoding region in positive-strand RNA synthesis. *J. Virol.* **78:** 1344–1351.

Brown, D.M., Cornell, C.T., Tran, G.P., Nguyen, J.H., and Semler, B.L. 2005. An authentic 3′-noncoding region is necessary for efficient poliovirus replication. *J. Virol.* **79:** 11962–11973.

Cho, M.W., Teterina, N., Egger, D., Bienz, K., and Ehrenfeld, E. 1994. Membrane rearrangement and vesicle induction by recombinant poliovirus 2C and 2BC in human cells. *Virology* **202:** 129–145.

Denny Jr., F.W. 1995. The clinical impact of human respiratory virus infections. *Am. J. Respir. Crit. Care Med.* **152:** S4–12.

Duque, H. and Palmenberg, A.C. 2001. Phenotypic characterization of three phylogenetically conserved stem-loop motifs in the mengovirus 3′-untranslated region. *J. Virol.* **75:** 3111–3120.

Gamarnik, A.V. and Andino, R. 1997. Two functional complexes formed by KH domain containing proteins with the 5′-noncoding region of poliovirus RNA. *RNA* **3:** 882–892.

Gamarnik, A.V. and Andino, R. 1998. Switch from translation to RNA replication in a positive-stranded RNA virus. *Genes & Dev.* **12:** 2293–2304.

Gerber, K., Wimmer, E., and Paul, A.V. 2001. Biochemical and genetic studies of the initiation of human rhinovirus 2 RNA replication: Identification of a *cis*-replicating element in the coding sequence of $2A^{pro}$. *J. Virol.* **75:** 10979–10990.

Goodfellow, I., Chaudhry, Y., Richardson, A., Meredith, J., Almond, J.W., Barclay, W., and Evans, D.J. 2000. Identification of a *cis*-acting replication element within the poliovirus coding region. *J. Virol.* **74:** 4590–4600.

Goodfellow, I.G., Polacek, C., Andino, R., and Evans, D.J. 2003. The poliovirus 2C *cis*-acting replication element-mediated uridylylation of VPg is not required for synthesis of negative-sense genomes. *J. Gen. Virol.* **84:** 2359–2363.

Gorbalenya, A.E., Koonin, E.V., and Wolf, Y.I. 1990. A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS Lett.* **262:** 145–148.

Huang, H., Alexandrov, A., Chen, X., Barnes III., T.W., Zhang, H., Dutta, K., and Pascal, S.M. 2001. Structure of an RNA hairpin from HRV-14. *Biochemistry* **40:** 8055–8064.

Jacobson, S.J., Konings, D.A., and Sarnow, P. 1993. Biochemical and genetic evidence for a pseudoknot structure at the 3′-terminus of the poliovirus RNA genome and its role in viral RNA amplification. *J. Virol.* **67:** 2961–2971.

Kapikian, A.Z. 1967. Rhinoviruses: A numbering system. *Nature* **213:** 761–762.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30:** 3059–3066.

Kistler, A., Avila, P.C., Rouskin, S., Magrini, V., Credle, J.J., Schnurr, D.P., Boushey, H.A., Mardis, E.R., Li, H., and DeRisi, J.L. 2007. Pan-viral screening of respiratory tract infections in adults with and without asthma reveals unexpected human coronavirus and human rhinovirus diversity. *J. Infect. Dis.* **196:** 817–825.

Kitamura, N., Semler, B.L., Rothberg, P.G., Larsen, G.R., Adler, C.J., Dorner, A.J., Emini, E.A., Hanecak, R., Lee, J.J., Van der Werf, S., et al. 1981. Primary structure, gene organization and polypeptide expression of poliovirus RNA. *Nature* **291:** 547–553.

Lamson, D., Renwick, N., Kapoor, V., Liu, Z., Palacios, G., Ju, J., Dean, A., St George, K., Briese, T., and Lipkin, W.I. 2006. MassTag polymerase-chain-reaction detection of respiratory pathogens, including a new rhinovirus genotype, that caused influenza-like illness in New York State during 2004–2005. *J. Infect. Dis.* **194:** 1398–1402.

Lau, S.K., Yip, C.C., Tsoi, H.W., Lee, R.A., So, L.Y., Lau, Y.L., Chan, K.H., Woo, P.C., and Yuen, K.Y. 2007. Clinical features and complete genome characterization of a distinct human rhinovirus genetic cluster, probably representing a previously undetected HRV species, HRV-C, associated with acute respiratory illness in children. *J. Clin. Microbiol.* **196:** 986–993.

Ledford, R.M., Patel, N.R., Demenczuk, T.M., Watanyar, A., Herbertz, T., Collett, M.S., and Pevear, D.C. 2004. VP1 sequencing of all human rhinovirus serotypes: Insights into genus phylogeny and susceptibility to antiviral capsid-binding compounds. *J. Virol.* **78:** 3663–3674.

Lee, W.M., Kiesner, C., Pappas, T., Lee, I., Grindle, K., Jartti, T., Jakiela, B., Lemanske, R.F., Shult, P.A., and Gern, J.E. 2007. A diverse group of previously unrecognized human rhinoviruses are common causes of respiratory illnesses in infants. *PLoS ONE* **2:** e966. doi: 10.1371/journal.pone.0000966.

Lobert, P.E., Escriou, N., Ruelle, J., and Michiels, T. 1999. A coding RNA sequence acts as a replication signal in cardioviruses. *Proc. Natl. Acad. Sci.* **96:** 11560–11565.

Mason, P.W., Bezborodova, S.V., and Henry, T.M. 2002. Identification and characterization of a *cis*-acting replication element (*cre*) adjacent to the internal ribosome entry site of foot-and-mouth disease virus. *J. Virol.* **76:** 9686–9694.

McErlean, P., Shackelton, L.A., Lambert, S.B., Nissen, M.D., Sloots, T.P., and Mackay, I.M. 2007. Characterisation of a newly identified human rhinovirus, HRV-QPM, discovered in infants with bronchiolitis. *J. Clin. Virol.* **39:** 67–75.

McErlean, P., Shackelton, L.A., Andrews, E., Webster, D.R., Lambert, S.B., Nissen, M.D., Sloots, T.P., and Mackay, I.M. 2008. Distinguishing molecular features and clinical characteristics of a putative new rhinovirus species, human rhinovirus C (HRV C). *PLoS ONE* **3:** e1847. doi: 10.1371/journal.pone.0001847.

McKnight, K.L. 2003. The human rhinovirus internal *cis*-acting replication element (*cre*) exhibits disparate properties among serotypes. *Arch. Virol.* **148:** 2397–2418.

McKnight, K.L. and Lemon, S.M. 1996. Capsid coding sequence is required for efficient replication of human rhinovirus 14 RNA. *J. Virol.* **70:** 1941–1952.

McKnight, K.L. and Lemon, S.M. 1998. The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *RNA* **4:** 1569–1584.

Melchers, W.J., Hoenderop, J.G., Bruins Slot, H.J., Pleij, C.W., Pilipenko, E.V., Agol, V.I., and Galama, J.M. 1997. Kissing of the two predominant hairpin loops in the coxsackie B virus 3′-untranslated region is the essential structural feature of the origin of replication required for negative-strand RNA synthesis. *J. Virol.* **71:** 686–696.

Mirzayan, C. and Wimmer, E. 1992. Genetic analysis of an NTP-binding motif in poliovirus polypeptide 2C. *Virology* **189:** 547–555.
Mirzayan, C. and Wimmer, E. 1994. Biochemical studies on poliovirus polypeptide 2C: Evidence for ATPase activity. *Virology* **199:** 176–187.
Morasco, B.J., Sharma, N., Parilla, J., and Flanegan, J.B. 2003. Poliovirus cre(2C)-dependent synthesis of VPgpUpU is required for positive- but not negative-strand RNA synthesis. *J. Virol.* **77:** 5136–5144.
Parsley, T.B., Towner, J.S., Blyn, L.B., Ehrenfeld, E., and Semler, B.L. 1997. Poly(rC) binding protein 2 forms a ternary complex with the 5′-terminal sequences of poliovirus RNA and the viral 3CD proteinase. *RNA* **3:** 1124–1134.
Pathak, H.B., Arnold, J.J., Wiegand, P.N., Hargittai, M.R., and Cameron, C.E. 2007. Picornavirus genome replication: Assembly and organization of the VPg uridylylation ribonucleoprotein (initiation) complex. *J. Biol. Chem.* **282:** 16202–16213.
Paul, A.V., van Boom, J.H., Filippov, D., and Wimmer, E. 1998. Protein-primed RNA synthesis by purified poliovirus RNA polymerase. *Nature* **393:** 280–284.
Perera, R., Daijogo, S., Walter, B.L., Nguyen, J.H., and Semler, B.L. 2007. Cellular protein modification by poliovirus: The two faces of poly(rC)-binding protein. *J. Virol.* **81:** 8919–8932.
Pfister, T., Jones, K.W., and Wimmer, E. 2000. A cysteine-rich motif in poliovirus protein 2C(ATPase) is involved in RNA replication and binds zinc in vitro. *J. Virol.* **74:** 334–343.
Pilipenko, E.V., Poperechny, K.V., Maslova, S.V., Melchers, W.J., Slot, H.J., and Agol, V.I. 1996. *Cis*-element, oriR, involved in the initiation of (−) strand poliovirus RNA: A quasi-globular multidomain RNA structure maintained by tertiary (''kissing'') interactions. *EMBO J.* **15:** 5428–5436.
Renwick, N., Schweiger, B., Kapoor, V., Liu, Z., Villari, J., Bullmann, R., Miething, R., Briese, T., and Lipkin, W.I. 2007. A recently identified rhinovirus genotype is associated with severe respiratory-tract infection in children in Germany. *J. Infect. Dis.* **196:** 1754–1760.
Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16:** 276–277.
Richards, O.C., Spagnolo, J.F., Lyle, J.M., Vleck, S.E., Kuchta, R.D., and Kirkegaard, K. 2006. Intramolecular and intermolecular uridylylation by poliovirus RNA-dependent RNA polymerase. *J. Virol.* **80:** 7405–7415.
Rieder, E., Paul, A.V., Kim, D.W., van Boom, J.H., and Wimmer, E. 2000. Genetic and biochemical studies of poliovirus *cis*-acting replication element cre in relation to VPg uridylylation. *J. Virol.* **74:** 10371–10380.
Rieder, E., Xiang, W., Paul, A., and Wimmer, E. 2003. Analysis of the cloverleaf element in a human rhinovirus type 14/poliovirus chimera: Correlation of subdomain D structure, ternary protein complex formation and virus replication. *J. Gen. Virol.* **84:** 2203–2216.
Rodriguez, P.L. and Carrasco, L. 1993. Poliovirus protein 2C has ATPase and GTPase activities. *J. Biol. Chem.* **268:** 8105–8110.
Samuilova, O., Krogerus, C., Fabrichniy, I., and Hyypia, T. 2006. ATP hydrolysis and AMP kinase activities of nonstructural protein 2C of human parechovirus 1. *J. Virol.* **80:** 1053–1058.
Savolainen, C., Blomqvist, S., Mulders, M.N., and Hovi, T. 2002. Genetic clustering of all 102 human rhinovirus prototype strains: serotype 87 is close to human enterovirus 70. *J. Gen. Virol.* **83:** 333–340.
Sharma, N., O'Donnell, B.J., and Flanegan, J.B. 2005. 3′-Terminal sequence in poliovirus negative-strand templates is the primary *cis*-acting element required for VPgpUpU-primed positive-strand initiation. *J. Virol.* **79:** 3565–3577.
Shen, M., Wang, Q., Yang, Y., Pathak, H.B., Arnold, J.J., Castro, C., Lemon, S.M., and Cameron, C.E. 2007. Human rhinovirus type 14 gain-of-function mutants for oriI utilization define residues of 3C(D) and 3Dpol that contribute to assembly and stability of the picornavirus VPg uridylylation complex. *J. Virol.* **80:** 12485–12495.
Suhy, D.A., Giddings Jr., T.H., and Kirkegaard, K. 2000. Remodeling the endoplasmic reticulum by poliovirus infection and by individual viral proteins: An autophagy-like origin for virus-induced vesicles. *J. Virol.* **74:** 8953–8965.
Tapparel, C., Junier, T., Gerlach, D., Cordey, S., Van Belle, S., Perrin, L., Zdobnov, E.M., and Kaiser, L. 2007. New complete genome sequences of human rhinoviruses shed light on their phylogeny and genomic features. *BMC Genomics* **8:** 224.
Teterina, N.L., Levenson, E., Rinaudo, M.S., Egger, D., Bienz, K., Gorbalenya, A.E., and Ehrenfeld, E. 2006. Evidence for functional protein interactions required for poliovirus RNA replication. *J. Virol.* **80:** 5327–5337.
Thiviyanathan, V., Yang, Y., Kaluarachchi, K., Rijnbrand, R., Gorenstein, D.G., and Lemon, S.M. 2004. High-resolution structure of a picornaviral internal *cis*-acting RNA replication element (*cre*). *Proc. Natl. Acad. Sci.* **101:** 12688–12693.
Todd, S., Towner, J.S., Brown, D.M., and Semler, B.L. 1997. Replication-competent picornaviruses with complete genomic RNA 3′-noncoding region deletions. *J. Virol.* **71:** 8868–8874.
van Ooij, M.J., Vogt, D.A., Paul, A., Castro, C., Kuijpers, J., van Kuppeveld, F.J., Cameron, C.E., Wimmer, E., Andino, R., and Melchers, W.J. 2006. Structural and functional characterization of the coxsackievirus B3 CRE(2C): Role of CRE(2C) in negative- and positive-strand RNA synthesis. *J. Gen. Virol.* **87:** 103–113.
Washietl, S., Hofacker, I.L., and Stadler, P.F. 2005. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci.* **102:** 2454–2459.
Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F., and Backofen, R. 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.* **3:** e65. doi: 10.1371/journal.pcbi.0030065.
Yang, Y., Rijnbrand, R., McKnight, K.L., Wimmer, E., Paul, A., Martin, A., and Lemon, S.M. 2002. Sequence requirements for viral RNA replication and VPg uridylylation directed by the internal *cis*-acting replication element (*cre*) of human rhinovirus type 14. *J. Virol.* **76:** 7485–7494.
Yang, Y., Rijnbrand, R., Watowich, S., and Lemon, S.M. 2004. Genetic evidence for an interaction between a picornaviral *cis*-acting RNA replication element and 3CD protein. *J. Biol. Chem.* **279:** 12659–12667.
Yin, J., Paul, A.V., Wimmer, E., and Rieder, E. 2003. Functional dissection of a poliovirus *cis*-acting replication element [PV-*cre*(2C)]: Analysis of single- and dual-*cre* viral genomes and proteins that bind specifically to PV-*cre* RNA. *J. Virol.* **77:** 5152–5166.
Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31:** 3406–3415.

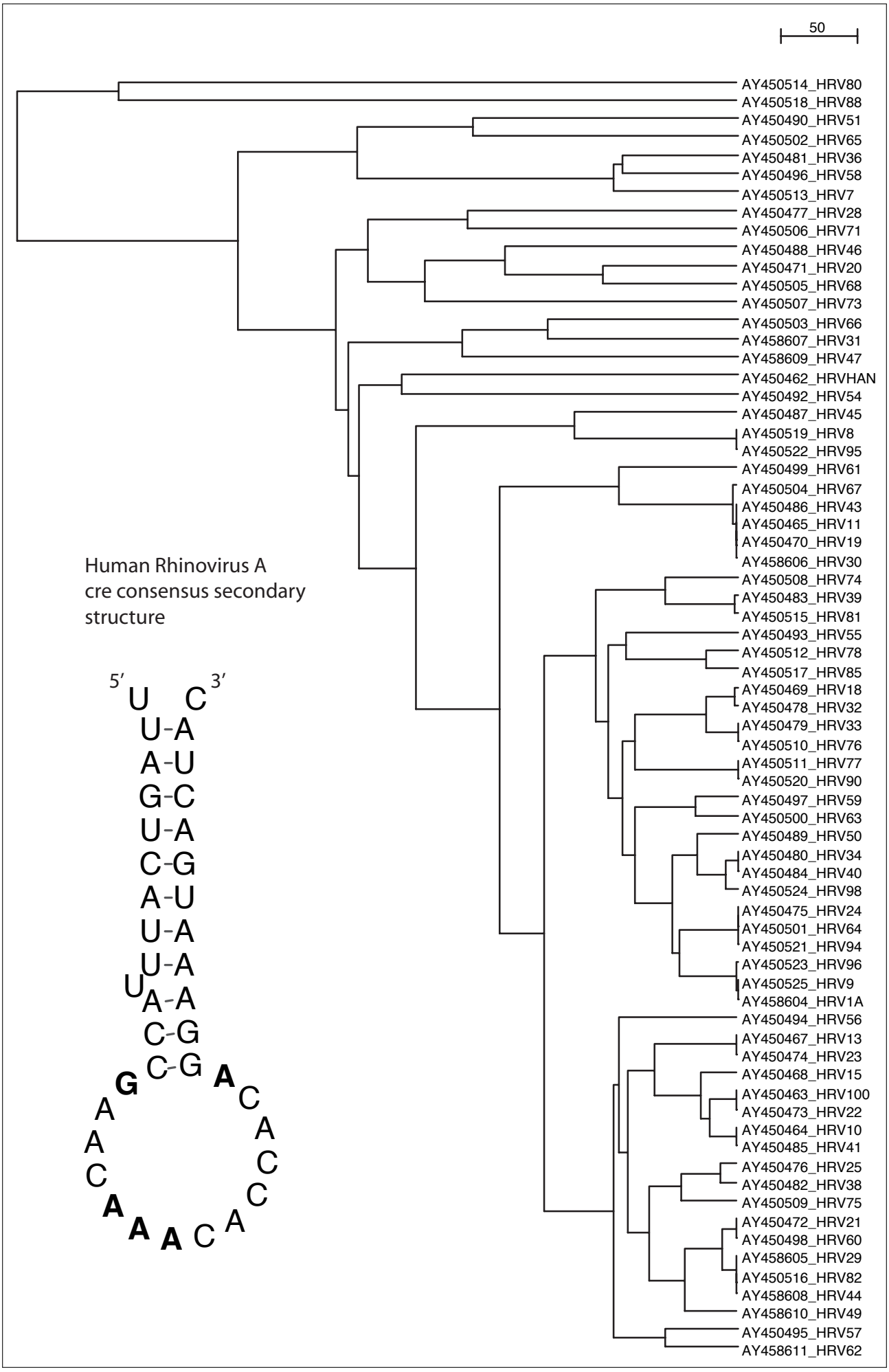
A.2.2.3 *Supplementary information*

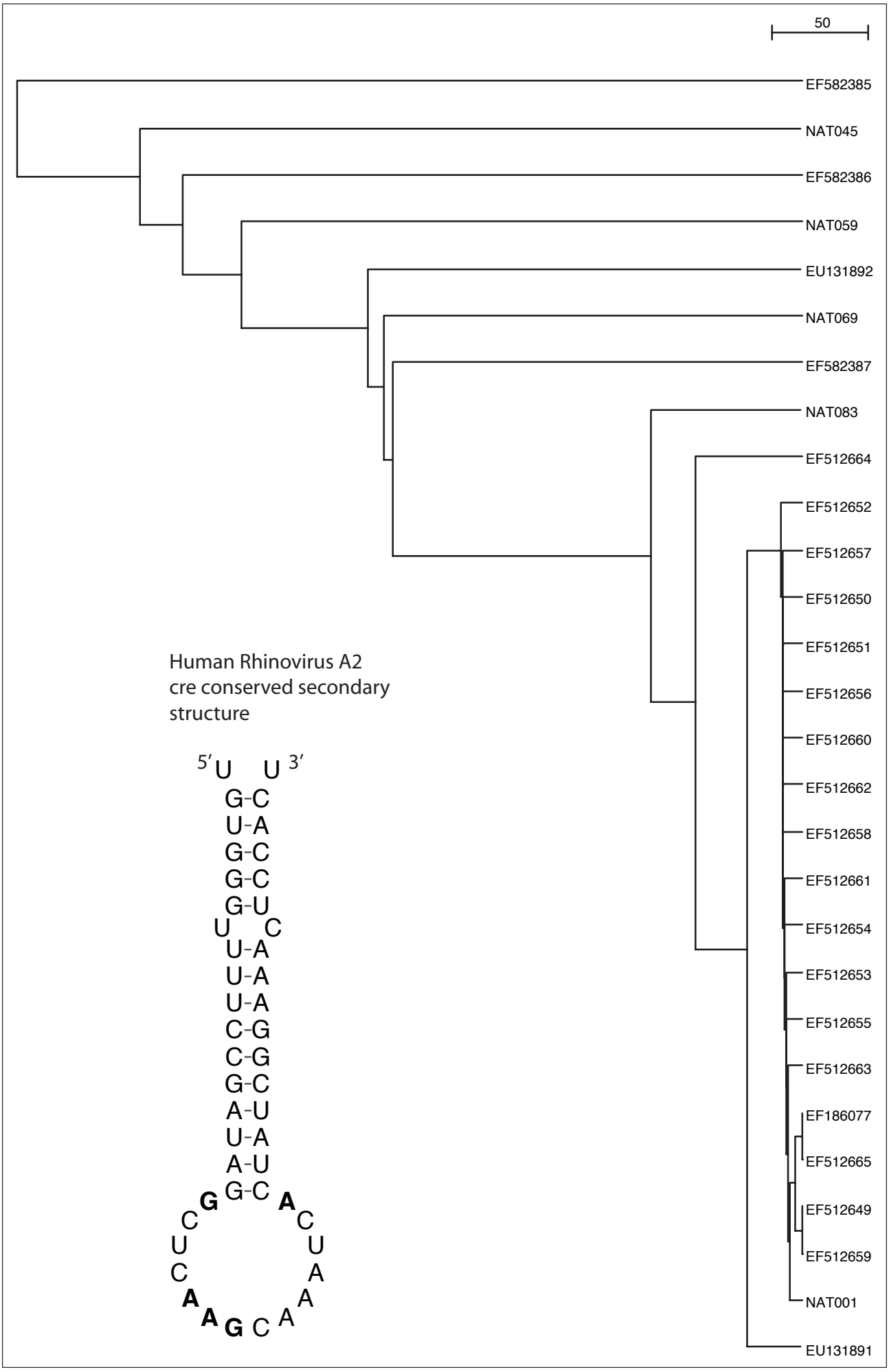
Figure legends for the corresponding original supplementary figures for the paper "The cis-acting replication elements define human enterovirus and rhinovirus species". Note, only figures produced by me are shown, the original publication contains more supplementary material. Source:
`http://rnajournal.cshlp.org/content/14/8/1568/suppl/DC1`

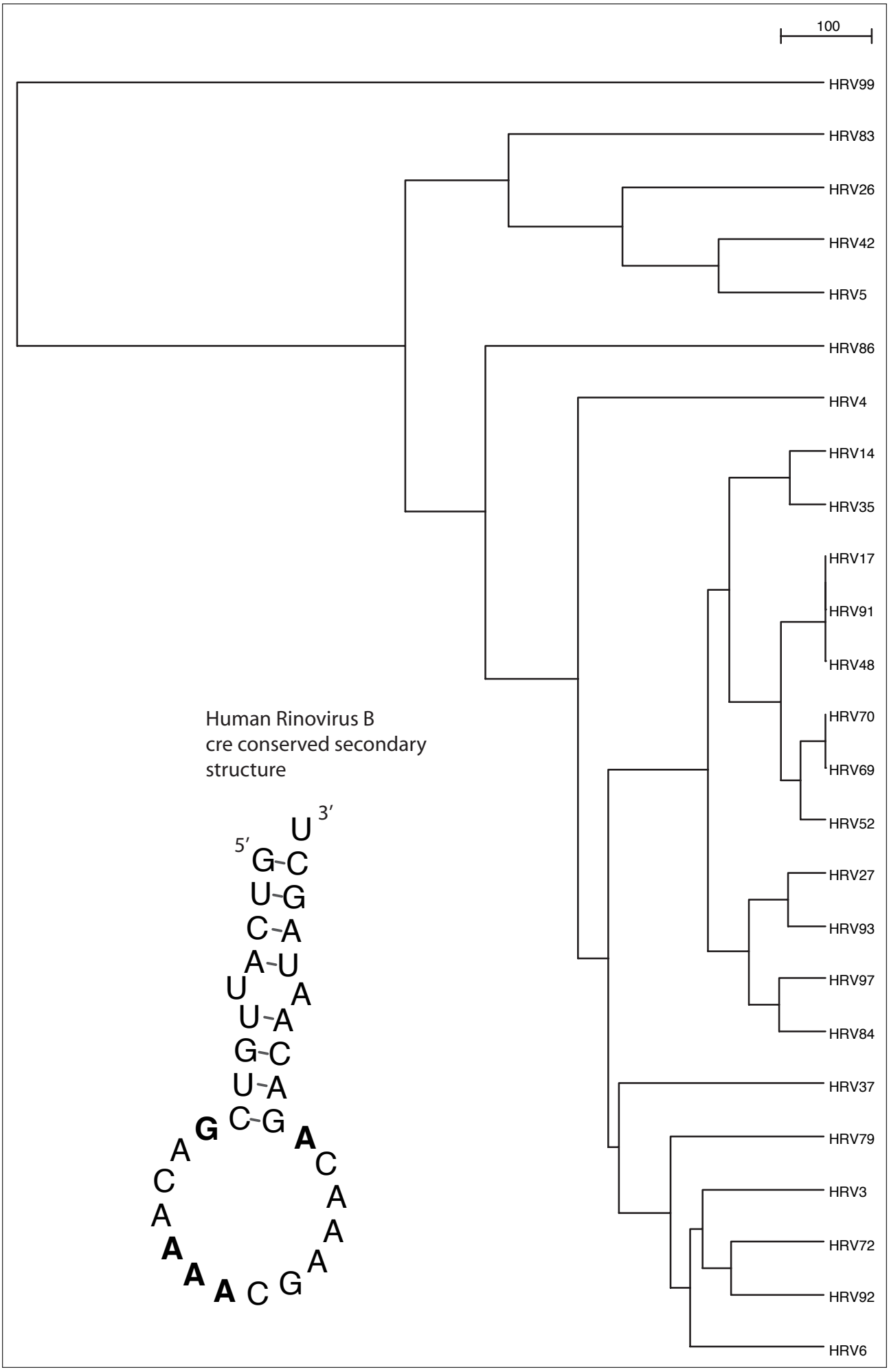
Supplementary Figure 2 on the next page WPGMA cluster tree of the human rhinovirus A 2A cre element plus its conserved consensus secondary structure. The conserved *cre* motif nucleotides are marked in bold in the consensus structure.

Supplementary Figure 3 on page 246 WPGMA cluster tree of the human rhinovirus A2 VP2 cre element plus its conserved consensus secondary structure. The conserved *cre* motif nucleotides are marked in bold in the consensus structure.

Supplementary Figure 4 on page 247 WPGMA cluster tree of the human rhinovirus B VP1 cre element plus its conserved consensus secondary structure. The conserved *cre* motif nucleotides are marked in bold in the consensus structure.


50
AY450514_HRV80
AY450518_HRV88
AY450490_HRV51
AY450502_HRV65
AY450481_HRV36
AY450496_HRV58
AY450513_HRV7
AY450477_HRV28
AY450506_HRV71
AY450488_HRV46
AY450471_HRV20
AY450505_HRV68
AY450507_HRV73
AY450503_HRV66
AY458607_HRV31
AY458609_HRV47
AY450462_HRVHAN
AY450492_HRV54
AY450487_HRV45
AY450519_HRV8
AY450522_HRV95
AY450499_HRV61
AY450504_HRV67
AY450486_HRV43
AY450465_HRV11
AY450470_HRV19
AY458606_HRV30
AY450508_HRV74
AY450483_HRV39
AY450515_HRV81
AY450493_HRV55
AY450512_HRV78
AY450517_HRV85
AY450469_HRV18
AY450478_HRV32
AY450479_HRV33
AY450510_HRV76
AY450511_HRV77
AY450520_HRV90
AY450497_HRV59
AY450500_HRV63
AY450489_HRV50
AY450480_HRV34
AY450484_HRV40
AY450524_HRV98
AY450475_HRV24
AY450501_HRV64
AY450521_HRV94
AY450523_HRV96
AY450525_HRV9
AY458604_HRV1A
AY450494_HRV56
AY450467_HRV13
AY450474_HRV23
AY450468_HRV15
AY450463_HRV100
AY450473_HRV22
AY450464_HRV10
AY450485_HRV41
AY450476_HRV25
AY450482_HRV38
AY450509_HRV75
AY450472_HRV21
AY450498_HRV60
AY458605_HRV29
AY450516_HRV82
AY458608_HRV44
AY458610_HRV49
AY450495_HRV57
AY458611_HRV62
Human Rhinovirus A
cre consensus secondary
structure
5′
3′


50
EF582385
NAT045
EF582386
NAT059
EU131892
NAT069
EF582387
NAT083
EF512664
EF512652
EF512657
EF512650
EF512651
EF512656
EF512660
EF512662
EF512658
EF512661
EF512654
EF512653
EF512655
EF512663
EF186077
EF512665
EF512649
EF512659
NAT001
EU131891
Human Rhinovirus A2
cre conserved secondary
structure
5′ U U 3′
G-C
U-A
G-C
G-C
G-U
U C
U-A
U-A
U-A
C-G
C-G
G-C
A-U
U-A
A-U
G-C
G A
C C
U U
C A
A A
G C A


100
HRV99
HRV83
HRV26
HRV42
HRV5
HRV86
HRV4
HRV14
HRV35
HRV17
HRV91
HRV48
HRV70
HRV69
HRV52
HRV27
HRV93
HRV97
HRV84
HRV37
HRV79
HRV3
HRV72
HRV92
HRV6
Human Rinovirus B
cre conserved secondary
structure
5′
3′

A.2.3 *New respiratory enterovirus and recombinant rhinoviruses among circulating picornaviruses*

Tapparel C, Junier T, Gerlach D, Van-Belle S, Turin L, Cordey S, Mühlemann K, Regamey N, Aubert JD, Soccal PM, Eigenmann P, Zdobnov E, and Kaiser L. New respiratory enterovirus and recombinant rhinoviruses among circulating picornaviruses. *Emerging Infect. Dis.* (2009) 15:719–726.

A.2.3.1 *Contributions*

The study by Tapparel et al., 2009 presents new respiratory enterovirus and evidence for recombination among rhinoviruses.

I participated in writing parts of the manuscript and the preparation of the Figs. 1–2.

A.2.3.2 *Main paper*

See pages 249–256 or at:
`http://www.cdc.gov/eid/content/15/5/719.htm`

New Respiratory Enterovirus and Recombinant Rhinoviruses among Circulating Picornaviruses

Caroline Tapparel,[1] Thomas Junier,[1] Daniel Gerlach, Sandra Van Belle, Lara Turin, Samuel Cordey, Kathrin Mühlemann, Nicolas Regamey, John-David Aubert, Paola M. Soccal, Philippe Eigenmann, Evgeny Zdobnov,[1] and Laurent Kaiser[1]

Rhinoviruses and enteroviruses are leading causes of respiratory infections. To evaluate genotypic diversity and identify forces shaping picornavirus evolution, we screened persons with respiratory illnesses by using rhinovirus-specific or generic real-time PCR assays. We then sequenced the 5′ untranslated region, capsid protein VP1, and protease precursor 3CD regions of virus-positive samples. Subsequent phylogenetic analysis identified the large genotypic diversity of rhinoviruses circulating in humans. We identified and completed the genome sequence of a new enterovirus genotype associated with respiratory symptoms and acute otitis media, confirming the close relationship between rhinoviruses and enteroviruses and the need to detect both viruses in respiratory specimens. Finally, we identified recombinants among circulating rhinoviruses and mapped their recombination sites, thereby demonstrating that rhinoviruses can recombine in their natural host. This study clarifies the diversity and explains the reasons for evolution of these viruses.

Human rhinoviruses (HRVs) and enteroviruses (HEVs) are leading causes of infection in humans. These 2 picornaviruses share an identical genomic organization, have similar functional RNA secondary structures, and are classified within the same genus (www.ictvonline.org/virusTaxonomy.asp) because of their high sequence homology (*1*). However, despite their common genomic features, these 2 groups of viruses have different phenotypic characteristics. In vivo, rhinoviruses are restricted to the respiratory tract, whereas enteroviruses infect primarily the gastrointestinal tract and can spread to other sites such as the central nervous system. However, some enteroviruses exhibit specific respiratory tropism and thus have properties similar to rhinoviruses (*2*–*5*). In vitro, most HRVs and HEVs differ by their optimal growth temperature, acid tolerance, receptor usage, and cell tropism. The genomic basis for these phenotypic differences between similar viruses is not yet fully understood.

HRVs and HEVs are characterized by ≈100 serotypes. Recently, molecular diagnostic tools have shown that this diversity expands beyond those predefined serotypes and encompasses also previously unrecognized rhinovirus and enterovirus genotypes. As an example, a new HRV lineage named HRV-C was recently identified and now complements the 2 previously known A and B lineages (*6*–*8*) (N.J. Knowles, pers. comm.). The C lineage has not only a distinct phylogeny (*9*–*16*) but is also characterized by specific cis-acting RNA structures (*17*).

In this study, we screened a large number of persons with acute respiratory diseases by using assays designed to overcome the diversity of both rhinoviruses and enteroviruses circulating in humans. Whenever possible, we systematically sequenced 5′ untranslated region (UTR), capsid protein VP1, and protease precursor 3CD regions of strains. Our goals were 1) to characterize the diversity of circulating rhinoviruses and, to a lesser extent, enteroviruses, to identify putative new picornavirus variants, and 2) to assess whether recombination may drive HRV evolution, which has not been shown in natural human infections (*18*).

Author affiliations: University of Geneva Hospitals, Geneva, Switzerland (C. Tapparel, S. Van Belle, L. Turin, S. Cordey, P. M. Soccal, P. Eigenmann, L. Kaiser); University of Geneva Medical School, Geneva (C. Tapparel, T. Junier, D. Gerlach, S. Van Belle, L. Turin, S. Cordey, E. Zdobnov, L. Kaiser); Swiss Institute of Bioinformatics, Geneva (T. Junier, D. Gerlach, E. Zdobnov); University Hospital of Bern, Bern, Switzerland (K. Mühlemann, N. Regamey); University Hospital of Lausanne, Lausanne, Switzerland (J.-D. Aubert); and Imperial College London, London, UK (E. Zdobnov)

DOI: 10.3201/eid1505.081286

[1]These authors contributed equally to this article.

Materials and Methods

RNA Extraction, Reverse Transcription-PCR, and Real-Time PCR

Reverse transcription–PCR (Superscript II; Invitrogen, Carlsbad, CA, USA) was performed on RNA extracted by using the HCV Amplicor Specimen Preparation kit (Roche, Indianapolis, IN, USA), TRIzol (Invitrogen), or the QIAamp Viral RNA Mini kit (QIAGEN, Valencia, CA, USA). Real-time PCR specific for HRV-A, HRV-B, and HEV (*19*), and a generic panenterhino real-time PCR (forward primer 5′-AGCCTGCGTGGCKGCC-3′, reverse primer 5′-GAAACACGGACACCCAAAGTAGT-3′, and probe 5-FAM-CTCCGGCCCCTGAATGYGGCTAA-TAMRA-3′), were performed in several cohort studies (Table).

Clinical Specimens

Picornavirus-positive samples were detected from patients enrolled in cohort studies in different regions of Switzerland during 1999–2008. The main characteristics of these populations, type of respiratory specimens, and screening methods are shown in the Table. The rhinovirus serotypes used for 3CD sequencing were obtained from the American Type Culture Collection (Manassas, VA, USA).

PCR and Sequencing

Sequencing was performed directly from the clinical specimen except for samples selected by routine isolation methods on human embryonic (HE) primary fibroblast cell lines (Table) or for HRV reference serotypes. Primers used to amplify the 5′-UTR and the VP1 and 3CD regions are listed in online Technical Appendix 1 Table 1A (available from www.cdc.gov/EID/content/15/5/719-Techapp1.pdf).

Full-length genome sequences of CL-1231094, a related clinical strain of enterovirus, and partial sequences of CL-Fnp5 and CL-QJ274218 were obtained as follows. RNA extracted by using the QIAamp Viral RNA Mini kit (QIAGEN) plus DNase treatment or with Trizol was reverse transcribed with random-tagged primer FR26RV-N and amplified with the SMART RACE cDNA Amplification kit (Clontech, Mountain View, CA, USA) with a specific forward primer and FR20RV reverse primer (online Technical Appendix 1 Table 1B) (*23*). Amplification products were separated by electrophoresis on agarose gels and fragments (0.6–2.5 kb) were extracted by using the QIAquick Gel Extraction kit (QIAGEN). Purified products were cloned by using the TOPO TA cloning kit (Invitrogen).

Minipreps were prepared from individual colonies and clones with the largest inserts were chosen for sequencing. Sequences obtained were used to design a new forward primer (online Technical Appendix 1 Table 1) to advance toward the 3′ end of the genome. PCR products of 3′ genomic ends were obtained by using the BD Smart

Table. Characteristics of screened study populations and respiratory samples, Switzerland*

Type of study (no. enrolled)	Age group	Patient characteristics	Years of study	Type of specimens	PCR	No. (%) positive	Reference
Respiratory infection in newborns (243)	<1 y	Nonhospitalized children with initial respiratory episode with cough	1999–2005	NPS	HRV-A and HRV-B specific real time for the first 203 and panenterhino for 40	36 (15)	(*20*)
Lower respiratory tract infection in hospitalized patients (147)	Adults	Mainly immunocompromised patients with lower respiratory tract complications and comorbidities	2001–2003	BAL, NPS	HRV-A and HRV-B specific real time	16 (11)	(*21*)
Acute respiratory tract infection in children (653)	<17 y	Nonhospitalized children with AOM or pneumonia	2004–2007	NPS	Panenterhino	121 (18)	(*22*) and ongoing study
Lower respiratory tract infection in hospitalized patients (485)	Adults	Mainly immunocompromised patients with lower respiratory tract complications and concurrent illnesses	2003–2006	BAL, NPS	Panenterhino	52 (11)	(*21*) and ongoing study
Acute respiratory tract infection in children (64)	<12 y	Children at an emergency department with fever and acute respiratory symptoms treated with antimicrobial drugs	2006–2007	NPS	Panenterhino	23 (36)	NP
Isolation in routine procedures (NA)	Children and adults	Hospitalized patients	1999–2008	BAL, NPS	HE culture isolation	NA	NP

*NPS, nasopharyngeal samples; HRV, human rhinovirus; BAL, bronchoalveolar lavage; AOM, acute otitis media; NP, not published; NA, not available; HE, human embryonic primary fibroblast cell line.

Race cDNA amplification kit (Becton Dickinson, Franklin Lakes, NJ, USA) according to manufacturer's instructions. All PCR products were purified by using microcon columns (Millipore, Billerica, MA, USA) and sequenced by using the ABI Prism 3130XL DNA Sequencer (Applied Biosystems, Foster City, CA, USA). Chromatograms were imported for proofreading with the vector NTI Advance 10 program (Invitrogen). Overlapping fragments were assembled with the contigExpress module of the vector NTI Advance 10.

Sequence Analysis, Phylogeny, and Bootscanning of Recombinants

Alignments were constructed by using MUSCLE (*24*) with a maximum of 64 iterations. (For detailed analyses, see http://cegg.unige.ch/picornavirus.) Multiple FastA was converted into PHYLIP format (for tree building) with the EMBOSS program Seqret (*25*). Trees were built with PhyML (*26*) by using the general time reversible model, BIONJ for the initial tree, and optimized tree topology and branch lengths. Trees with <50 species and larger trees used 16 and 8 rate categories, respectively. Transition/transversion ratios, proportions of invariant sites, and shape parameters of the γ distribution were estimated.

To investigate the hypothesis of recombination and map the breakpoints, we adapted the bootscanning method (*27*) as follows. The alignment was sliced into windows of constant size and fixed overlap and a 100-replicate maximum-likelihood (using HRV-93 as an outgroup) was computed for each window. From each tree, the distance between the candidate recombinant and all other sequences was extracted. This extraction yielded a matrix of distances for each window and for each alignment position. A threshold was defined as the lowest distance plus a fraction (15%) of the difference between the highest and lowest distances. The nearest neighbors of the candidate recombinant were defined as sequences at a distance smaller than this threshold. This distance ensured that the nearest neighbor, as well as any close relative, was always included. Possible recombination breakpoints thus corresponded to changes of nearest neighbors. Serotypes included in this analysis represented serotypes close to CL-013775 and CL-073908 on the basis of 5′-UTR and VP1 phlyogenetic trees (online Technical Appendix 2 Figure 1, panels A, B, available from www.cdc.gov/EID/content/15/5/719-Techapp2.pdf), as well as serotypes close to CL-135587 on the basis of VP1 and 3CD phlyogenetic trees (online Technical Appendix 2 Figure 1, panels B, C) and whose full-length sequence was available.

Distance matrices were computed from alignments with the distmat program in EMBOSS (http://bioweb2.pasteur.fr/docs/EMBOSS/embossdata.html) by using the Tamura distance correction. This method uses transition and transversion rates and takes into account the deviation of GC content from the expected value of 50%. Gap and ambiguous positions were ignored. Final values were then converted to similarity matrices by subtracting each value from 100.

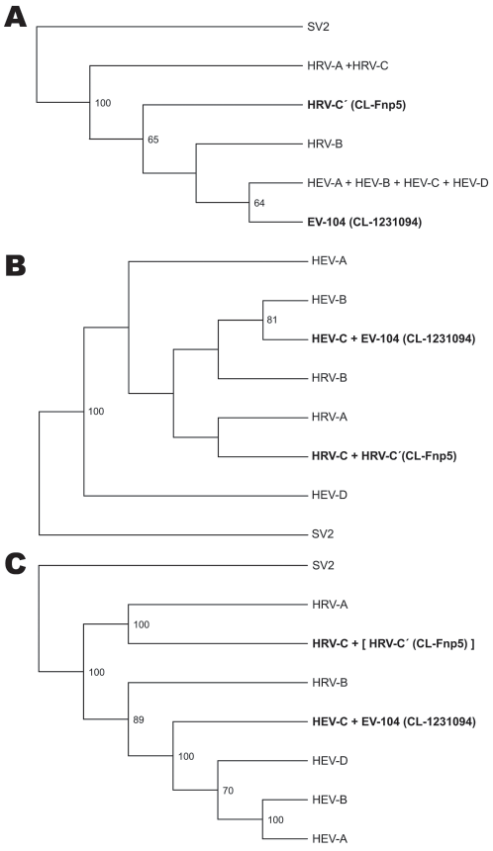


Figure 1. 5′ untranslated region (A), capsid protein VP1 (B), and complete genome (C) phylogeny of the virus clades studied. Trees were produced by condensing the full phylogeny shown in online Technical Appendix 2 Figure 1, panels A, B, and D (available from www.cdc.gov/EID/content/15/5/719-Techapp2.pdf). Human rhinovirus C′ (HRV-C') includes the divergent rhinoviruses described in 2007 (*13*) and a related clinical strain (CL-Fnp5). HRV-C includes the new clade described since 2006 (*9*–*14*,*16*). Enterovirus 104 (EV-104) and the related strain CL-1231094 refer to a previously unknown enterovirus clade described in this study. In panel C, HRV-C′ is shown in brackets to indicate its expected location (based on VP1 and 3D sequences). Simian picornavirus 1 (SV2) was used as an outgroup. HEV, human enterovirus. Bootstrap support values <50 are not shown in the trees. New viruses are shown in **boldface**.

RESEARCH

Results

Screening of Persons with Respiratory Tract Infections

Persons enrolled in several cohorts of children and adults with respiratory infections (Table) were screened for picornavirus by culture isolation on HE cell lines, real-time PCR specific for HRV-A and HRV-B (*19*), or by a panenterhino real-time PCR designed to theoretically detect all rhinoviruses and enteroviruses with publicly available sequences. Of 1,592 respiratory samples tested by real-time PCR, 248 were virus positive (Table). The 5′-UTR sequences were obtained for 77 real-time PCR or culture-positive samples and VP1 and 3CD sequences for 48 of these (Table; online Technical Appendix 1 Table 2). In parallel, the 3CD sequences were identified for all reference serotypes. The results of this screening are summarized in online Technical Appendix 1 Table 2, and all sequences are available from GenBank (accession nos. EU840726–EU840988).

On the basis of these results, respiratory infections caused by HRV-B might be less frequent than those caused by HRV-A, and HRV-A infections are distributed among the whole library of reference serotypes. A specific real-time PCR used to detect enteroviruses in respiratory specimens from some of the cohorts studied indicated that these viruses are rare in children (2.5% vs. 6.3% for HRV) and even rarer or absent in adults (0% vs. 24% for HRV) (*28*).

Phylogeny and Molecular Epidemiology of 5′-UTR

To include all 99 HRV reference strains and new divergent rhinoviruses described recently by Lee et al. (*13*), we reconstructed a phylogenetic tree (online Technical Appendix 2 Figure 1, panel A) on the basis of a sequence of 280 nt in the 5′-UTR. This sequence provided a correct clustering of HRV-A, HRV-B, and HEV strains according to the accepted whole-genome phylogeny (online Technical Appendix 2 Figure 1, panel D) (*15*) but did not resolve appropriately the phylogeny of the 4 HEV species and the HRV-A and HRV-C viruses. The condensed tree version (Figure 1, panel A) enabled us to identify 2 groups phylogenetically distant from all previously known HRVs and HEVs. The first group, referred to as HRV-C′, contained some of our clinical samples and rhinoviruses sequenced by Lee et al. (*13*). The second group was a new clade and was named EV-104. This clade included 8 clinical samples collected in different regions of Switzerland without direct epidemiologic links (online Technical Appendix 1 Table 2).

Identification of HRV-C Viruses by Sequencing of HRV Viruses with Divergent 5′-UTRs

Characterization of HRVs newly identified during 2006–2008 showed that they all belong to the same HRV-C species (*9*–*16*). Recently, Lee et al. (*13*) identified another cluster of viruses (HRV-C′; Figure 1, panel A) and suggested that this group was phylogenetically distinct from all other HRVs on the basis of analysis of their 5′-UTR sequences. To define the phylogeny, we adapted a previously described method (*23*) to complete the genome sequence directly from our clinical strains (CL-Fnp5 and CL-QJ274218) that showed a similar divergent 5′-UTR (online Technical Appendix 2 Figure 1, panel A). A condensed version (Figure 1, panel B) of the phylogenetic tree based on VP1 sequences (online Technical Appendix 2 Figure 1, panel B) indicated that CL-Fnp5 clustered with the new HRV-C clade, a finding further confirmed by CL-QJ 274218 partial sequences. This finding supports the view that new HRVs variants described since 2006 (*9*–*16*) all belong to the same lineage.

New Divergent Lineage of HEV Species C

As shown in Figure 1, panel A, the panenterhino real-time PCR enabled detection of a new HEV strain phylogenetically distinct from all previously known HEV species and associated with respiratory diseases. Enterovirus-specific real-time PCRs or reference VP1 primer sets routinely used to type enteroviruses (primers 222 and 224 and nested primers AN88 and 89) (*29*,*30*) did not amplify this new genotype. We could not grow this virus on HeLa and HE cell lines. Consequently, we applied the method described above to complete the genome sequence directly from the CL-1231094 (EU840733) clinical specimen. VP1 and full-length genome sequences showed that, albeit divergent at the 5′-UTR level, this new variant belonged to the HEV-C species (Figure 1, panels B, C). Full-length genome phylogenetic tree (Figure 2) and VP1 protein identity plots (online Technical Appendix 2 Figure 2) with all members of the HEV-C species indicated that this virus represents a new HEV-C genotype that shares 68%, 66%, and 63% nucleotide and 77%, 75%, and 68% amino acid sequence identity, respectively, with coxsackieviruses A19 (CV-A19), A22, and A1, the closest serotypes. This new virus was named EV-104 (www.picornastudygroup.com/types/enterovirus_genus.htm).

Specific primers (Ent_P1.29/P2.13 and Ent_P3.30/P3.32; online Technical Appendix 1 Table 1C) were then designed to amplify the VP1 and 3D regions of the 7 other samples of this cluster collected from children with acute respiratory tract infections and otitis media. VP1 nucleotide homology among these strains was 94%–98%, except for 1 distantly related sample (74%–76%), which may represent an additional genotype. Additional sequencing is ongoing to verify this assumption.

At the 5′-UTR level, the strain described by Lee et al. (*13*) and EV-104 diverged from other members of HRV-C and HEV-C species, respectively. Thus, the 5′-UTR–based

phylogeny was inconsistent with that based on VP1 sequences and suggested possible recombination events (Figure 1, panels A, B). Because the 5′-UTR is the target of most molecular diagnostic assays, this sequence divergence needs to be taken into account in future studies.

Recombination Events between 5′-UTR, VP1, and 3CD Genome Regions

Other studies have provided sequences of clinical strains, but genetic characterization was often limited to 1 genomic region. Our goal was to sequence 3 genomic regions for each analyzed strain to determine definitively whether recombination events could represent a driving force for the evolution of rhinoviruses in their natural environment. Although recombination events have been suggested for reference serotypes, they have never been shown for circulating clinical strains (*18*,*31*,*32*). In contrast, recombination is well established as a driving force of enterovirus evolution. Thus, we completed the 5′-UTR, VP1, and 3CD sequences of 43 clinical strains by using a pool of adapted and degenerated primers (online Technical Appendix 1 Table 1A).

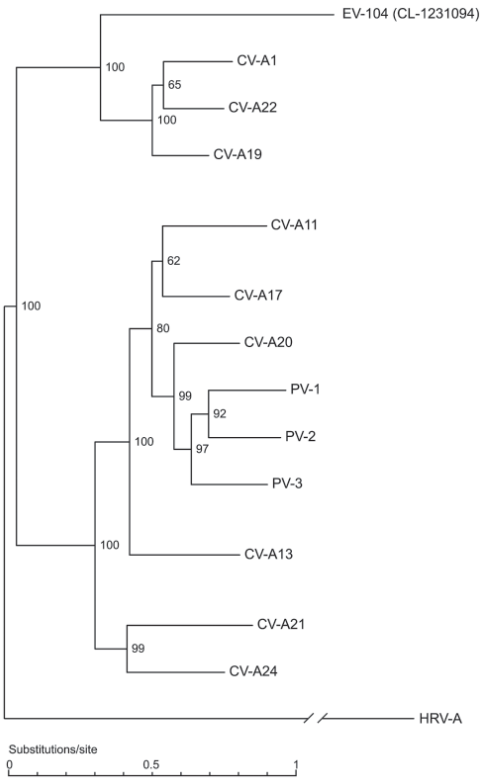


Figure 2. Full genome phylogenetic tree of enterovirus 104 (EV-104), representative strain CL-1231094, and members of the human enterovirus C (HEV-C) species. Human rhinovirus A (HRV-A) (GenBank accession no. DQ473509) was used as outgroup. Coxsackievirus A1 (CV-A1) (AF499635), CV-A21 (AF546702), CV-A20 (AF499642), CV-A17 (AF499639), CV-A13 (AF499637), CV-A11 (AF499636), CV-A19 (AF499641), CV-A22 (AF499643), CV-A24 (D90457), poliovirus 1 (PV-1) (V01148), PV-2 (X00595), and PV-3 (X00925) sequences were obtained from GenBank.

Independent phylogenetic trees (online Technical Appendix 2) and similarity matrices were constructed for the 3 genomic regions. Since the last common ancestor and as depicted on the distance matrices and highlighted by boxplots of maximum-likelihood branch length distributions (online Technical Appendix 2 Figure 3), there are more mutations fixed in the VP1 region than in the 3CD region, and more in the 3CD region than in 5′-UTR, which is indicative of a variable rate of evolution in these regions. Accordingly, VP1 sequences enabled genotyping of all but 3 clinical strains analyzed (online Technical Appendix 2, Figure 1, panel B). These strains may represent rhinovirus genotypes only distantly related to predefined reference serotypes. In contrast, genotyping based on 3CD and 5′-UTR was less accurate, as expected. These results confirmed that molecular typing of rhinoviruses, similarly to other picornaviruses, must use capsid sequences.

Phylogeny of the 5′-UTR, VP1, and 3CD of reference serotypes showed many incongruities caused by insufficient tree resolution or recombinant viruses as previously proposed (*18*,*31*). As an example, 2 VP1 clusters including HRV-85/HRV-40 and HRV-18/HRV-50/HRV-34 (online Technical Appendix 2 Figure 1, panel B) were reorganized as HRV-85/HRV-18/HRV-40 and HRV-50/HRV-34, respectively, on 3CD (online Technical Appendix 2 Figure 1, panel C). The differential cosegregations between these virus strains suggested recombination events. When available, full-length genome sequence bootscanning applied to all serotypes will give an estimate of the number of reference strains with mosaic genomes.

Similarly, the noncoding region, VP1, and 3CD trees showed major phylogenetic incongruities for 3 clinical isolates (online Technical Appendix 2 Figure 1). Two of these isolates (CL-013775 and CL-073908) were typed as HRV-67 on the basis of VP1 sequence and were closest to this serotype in 3CD, whereas the 5′-UTR cosegregated with HRV-36 (see 5′-UTR recombinant; online Technical Appendix 2 Figure 1, panels A–C). These viruses were isolated by cell culture from 2 epidemiologically linked cases and thus represented transmission of the same virus. To confirm the recombination, we completed the sequencing by obtaining the 5′-UTR, VP4, and VP2 sequences (EU840918 and EU840930) and compared them with HRV-36, HRV-67, and other closely related serotypes. Bootscanning analysis

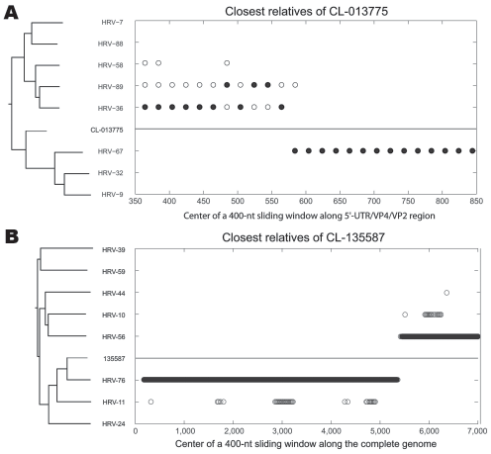


Figure 3. Nearest-neighbor relatedness of rhinovirus CL-013775 (and CL-073908) along the 5′ untranslated region/VP4/VP2 region (A), and nearest-neighbor relatedness of rhinovirus CL-135587 along the complete genome (B), identified by bootscanning. At each position of a sliding window, the solid circles indicate the closest relative within a defined threshold of the phylogenetic distance to CL-013775 (A) and CL-135587 (B). Both panels show phylogenetic trees of analyzed serotypes over the entire scanned region. Human rhinovirus 7 (HRV-7), -9, -10, -11, -24, -32 (accession nos. EU096019, AF343584), -36, -39, -44, -56, -58 (EU096045, AY040236), -59, -67 (EU096054, AF343603, and DQ473505), -76, -88, and -89 sequences were obtained from GenBank (see online Technical Appendix 2 Figure 1, available from www.cdc.gov/EID/content/15/5/719-Techapp2.pdf, for full-length genome accession numbers).

(Figure 3, panel A) enabled mapping of the recombination site within the 5′-UTR, just before the polyprotein start codon. Sequence alignment mapped recombination breakpoints more precisely between positions 524 and 553 with reference to HRV-2 (X02316).

The other incongruent isolate (CL-135587) was typed as HRV-76 on the basis of VP1 sequence and was closest to this serotype in the 5′-UTR, but 3CD cosegregates with HRV-56 (3C recombinant; online Technical Appendix 2 Figure 1, panels B, C). Similarly, we completed the full-length sequence of this isolate (EU840726) and HRV-56 (EU840727). The same approach enabled mapping of the recombination site at the N terminus of protein 3C between positions 1511 and 1523 with reference to HRV-2 (Figure 3, panel B). These results demonstrate that recombination occurs among clinical rhinoviruses. In our analysis of 40 rhinovirus-positive samples collected over 9 years (3 additional samples were duplicates of 2 different viruses; online Technical Appendix 1 Table 2) for 3 genomic regions, 2 of the analyzed viruses appeared to be recombinants. The 2 documented recombinations occurred in members of the HRV-A species. The design of this study and technical issues (e.g., inability to sequence low viral loads) limited the ability to calculate a recombination rate, particularly for HRV-B and HRV-C.

Discussion

Our genomic analysis of picornaviruses associated with upper or lower respiratory diseases in adults and children indicates that rhinoviruses circulating in the community are widely diverse. The large number of circulating genotypes supports the view that rhinoviruses do not circulate by waves or outbreaks of a given dominant genotype, which might explain the high frequency of reinfection during short periods. As expected, the observed variability is higher for surface capsid proteins, the targets of most immune pressure, and this region remains the only accurate one for genotyping and defining phylogeny. Technical constraints such as the limited amount of clinical specimens, the use of different screening methods, and the need to sequence an unknown target of extreme variability might have limited the representativeness of our sequence collection. Therefore, our study should not be considered as an exhaustive epidemiologic analysis of rhinoviruses and enteroviruses associated with respiratory diseases.

By using a systematic approach, we have identified a new enterovirus genotype (EV-104) that has a divergent 5′-UTR. Undetectable by conventional methods, EV-104 could be detected by using a more generic real-time PCR assay designed to match all known available rhinovirus and enterovirus sequences. Such diagnostic tools have and will lead to constant discovery of new picornavirus genotypes (*9–14*,*16*,*33–36*). These genotypes may represent viruses, in most instances, that have remained undetected because of insensitive cell cultures or overly restrictive molecular tools. In addition, enterovirus genotypes causing respiratory infections, such as EV-68 and CV-A21, might be underrepresented because enteroviruses are usually searched for in fecal specimens (*37*).

EV-104 belongs to the HEV-C species: CV-A19, CV-A22, and CV-A1 are its closest serotypes. These HEV-C subgroup viruses are genetically distinct from all other serotypes of the species. These viruses show no evidence of recombination with other HEV-C strains and, similar to EV-104, do not grow in cell culture (*29*). On the basis of our epidemiologic data, we conclude that EV-104 was found in 8 children from different regions of Switzerland who had respiratory illnesses such as acute otitis media or pneumonia. Future studies using adapted detection tools will provide more information on the range of this virus. On the basis of its genomic features and similarities with coxsackieviruses and poliovirus, EV-104 could theoreti-

cally infect the central nervous system (*2*,*38*). Detection of new subtypes of picornaviruses indicates that viruses with new phenotypic traits could emerge, and conclusions on tropism of new strains should be substantiated by extensive experimental or clinical investigations (*39*).

By completing the sequence of a seemingly divergent rhinovirus (*13*), we assigned this virus to the new HRV-C species, thus limiting currently to 3 the number of HRV species. For the sake of simplicity, we propose to consider this virus as a member of the HRV-C clade.

Finally, we demonstrated that rhinovirus evolves by recombination in its natural host. Known to be a driving force of enterovirus evolution, rhinovirus recombination among clinical strains has never been observed. Two clinical isolates of 40 viruses analyzed resulted from recombination events and their breakpoints were identified within the 5′-UTR sequence and the N terminus of protein 3C, respectively. These findings are consistent with the fact that recombination breakpoints in picornaviruses are restricted to nonstructural regions of the genome or between the 5′-UTR and the capsid-encoding region (*40*). Our observations provide new insight on the diversity and ability of rhinovirus to evolve in its natural host. The fact that only 2 of 40 analyzed viruses over a 9-year period were recombinants is suggestive of a lower recombination frequency in rhinoviruses than in other picornaviruses (*32*,*40*) and might be related, but not exclusively, to the short duration of rhinovirus infection (*18*,*31*,*32*). Recombination events occurred between HRV-A genotypes, but whether they can occur in species B and C remains unknown. Interspecies recombination is rare in picornaviruses and is mainly the result of in vitro experiments. For rhinoviruses, the different location of *cre* elements in each species might be an additional limiting constraint (*17*).

In summary, we have highlighted the large genomic diversity of the most frequent human respiratory viral infection. Our phylogenetic analysis has characterized circulating strains relative to reference strains and has identified a previously unknown enterovirus genotype. We have shown that recombination also contributes to rhinovirus evolution in its natural environment.

Acknowledgments

We thank Rosemary Sudan for editorial assistance and the Swiss Institute of Bioinformatics' Vital-IT facility for bootscanning and computing infrastructure.

This study was supported by the Swiss National Science Foundation (grants 3200B0-101670 to L.K. and 3100A0112588/I to E.Z.), the Department of Medicine of the University Hospitals of Geneva, the University of Geneva Dean's Program for the Promotion of Women in Science (C.T.), and the Infectigen Foundation.

Dr Tapparel is a molecular virologist at the University Hospitals of Geneva. Her research interests are the molecular epidemiology of picornaviruses (rhinoviruses and enteroviruses), development of new diagnostic methods, and determination of fundamental aspects of these viruses.

References

1. Tapparel C, Junier T, Gerlach D, Cordey S, Van Belle S, Perrin L, et al. New complete genome sequences of human rhinoviruses shed light on their phylogeny and genomic features. BMC Genomics. 2007;8:224. DOI: 10.1186/1471-2164-8-224
2. Newcombe NG, Andersson P, Johansson ES, Au GG, Lindberg AM, Barry RD, et al. Cellular receptor interactions of C-cluster human group A coxsackieviruses. J Gen Virol. 2003;84:3041–50. DOI: 10.1099/vir.0.19329-0
3. Pulli T, Koskimies P, Hyypia T. Molecular comparison of coxsackie A virus serotypes. Virology. 1995;212:30–8. DOI: 10.1006/viro.1995.1450
4. Dufresne AT, Gromeier M. A nonpolio enterovirus with respiratory tropism causes poliomyelitis in intercellular adhesion molecule 1 transgenic mice. Proc Natl Acad Sci U S A. 2004;101:13636–41. DOI: 10.1073/pnas.0403998101
5. Oberste MS, Maher K, Schnurr D, Flemister MR, Lovchik JC, Peters H, et al. Enterovirus 68 is associated with respiratory illness and shares biological features with both the enteroviruses and the rhinoviruses. J Gen Virol. 2004;85:2577–84. DOI: 10.1099/vir.0.79925-0
6. Ledford RM, Patel NR, Demenczuk TM, Watanyar A, Herbertz T, Collett MS, et al. VP1 sequencing of all human rhinovirus serotypes: insights into genus phylogeny and susceptibility to antiviral capsid-binding compounds. J Virol. 2004;78:3663–74. DOI: 10.1128/JVI.78.7.3663-3674.2004
7. Laine P, Blomqvist S, Savolainen C, Andries K, Hovi T. Alignment of capsid protein VP1 sequences of all human rhinovirus prototype strains: conserved motifs and functional domains. J Gen Virol. 2006;87:129–38. DOI: 10.1099/vir.0.81137-0
8. Savolainen C, Blomqvist S, Mulders MN, Hovi T. Genetic clustering of all 102 human rhinovirus prototype strains: serotype 87 is close to human enterovirus 70. J Gen Virol. 2002;83:333–40.
9. Arden KE, McErlean P, Nissen MD, Sloots TP, Mackay IM. Frequent detection of human rhinoviruses, paramyxoviruses, coronaviruses, and bocavirus during acute respiratory tract infections. J Med Virol. 2006;78:1232–40. DOI: 10.1002/jmv.20689
10. Kistler A, Avila PC, Rouskin S, Wang D, Ward T, Yagi S, et al. Pan-viral screening of respiratory tract infections in adults with and without asthma reveals unexpected human coronavirus and human rhinovirus diversity. J Infect Dis. 2007;196:817–25. DOI: 10.1086/520816
11. Lamson D, Renwick N, Kapoor V, Liu Z, Palacios G, Ju J, et al. MassTag polymerase-chain-reaction detection of respiratory pathogens, including a new rhinovirus genotype, that caused influenza-like illness in New York State during 2004–2005. J Infect Dis. 2006;194:1398–402. DOI: 10.1086/508551
12. Lau SK, Yip CC, Tsoi HW, Lee RA, So LY, Lau YL, et al. Clinical features and complete genome characterization of a distinct human rhinovirus (HRV) genetic cluster, probably representing a previously undetected HRV species, HRV-C, associated with acute respiratory illness in children. J Clin Microbiol. 2007;45:3655–64. DOI: 10.1128/JCM.01254-07
13. Lee WM, Kiesner C, Pappas T, Lee I, Grindle K, Jartti T, et al. A diverse group of previously unrecognized human rhinoviruses are common causes of respiratory illnesses in infants. PLoS One. 2007;2:e966. DOI: 10.1371/journal.pone.0000966

14. McErlean P, Shackelton LA, Lambert SB, Nissen MD, Sloots TP, Mackay IM. Characterisation of a newly identified human rhinovirus, HRV-QPM, discovered in infants with bronchiolitis. J Clin Virol. 2007;39:67–75. DOI: 10.1016/j.jcv.2007.03.012
15. McErlean P, Shackelton LA, Andrews E, Webster DR, Lambert SB, Nissen MD, et al. Distinguishing molecular features and clinical characteristics of a putative new rhinovirus species, human rhinovirus C (HRV C). PLoS One. 2008;3:e1847.
16. Renwick N, Schweiger B, Kapoor V, Liu Z, Villari J, Bullmann R, et al. A recently identified rhinovirus genotype is associated with severe respiratory-tract infection in children in Germany. J Infect Dis. 2007;196:1754–60. DOI: 10.1086/524312
17. Cordey S, Gerlach D, Junier T, Zdobnov EM, Kaiser L, Tapparel C. The cis-acting replication elements define human enterovirus and rhinovirus species. RNA. 2008;14:1568–78. DOI: 10.1261/rna.1031408
18. Savolainen C, Laine P, Mulders MN, Hovi T. Sequence analysis of human rhinoviruses in the RNA-dependent RNA polymerase coding region reveals large within-species variation. J Gen Virol. 2004;85:2271–7. DOI: 10.1099/vir.0.79897-0
19. Deffernez C, Wunderli W, Thomas Y, Yerly S, Perrin L, Kaiser L. Amplicon sequencing and improved detection of human rhinovirus in respiratory samples. J Clin Microbiol. 2004;42:3212–8. DOI: 10.1128/JCM.42.7.3212-3218.2004
20. Regamey N, Kaiser L, Roiha HL, Deffernez C, Kuehni CE, Latzin P, et al. Viral etiology of acute respiratory infections with cough in infancy: a community-based birth cohort study. Pediatr Infect Dis J. 2008;27:100–5.
21. Garbino J, Gerbase MW, Wunderli W, Deffernez C, Thomas Y, Rochat T, et al. Lower respiratory viral illnesses: improved diagnosis by molecular methods and clinical impact. Am J Respir Crit Care Med. 2004;170:1197–203. DOI: 10.1164/rccm.200406-781OC
22. Kronenberg A, Zucs P, Droz S, Muhlemann K. Distribution and invasiveness of *Streptococcus pneumoniae* serotypes in Switzerland, a country with low antibiotic selection pressure, from 2001 to 2004. J Clin Microbiol. 2006;44:2032–8. DOI: 10.1128/JCM.00275-06
23. Allander T, Tammi MT, Eriksson M, Bjerkner A, Tiveljung-Lindell A, Andersson B. Cloning of a human parvovirus by molecular screening of respiratory tract samples. Proc Natl Acad Sci U S A. 2005;102:12891–6. DOI: 10.1073/pnas.0504666102
24. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7. DOI: 10.1093/nar/gkh340
25. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7. DOI: 10.1016/S0168-9525(00)02024-2
26. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52:696–704. DOI: 10.1080/10635150390235520
27. Salminen MO, Carr JK, Burke DS, McCutchan FE. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses. 1995;11:1423–5.
28. Garbino J, Soccal PM, Aubert JD, Rochat T, Meylan P, Thomas Y, et al. Respiratory viruses in bronchoalveolar lavage: a hospital-based cohort study in adults. Thorax. 2009; [Epub ahead of print].
29. Brown B, Oberste MS, Maher K, Pallansch MA. Complete genomic sequencing shows that polioviruses and members of human enterovirus species C are closely related in the noncapsid coding region. J Virol. 2003;77:8973–84. DOI: 10.1128/JVI.77.16.8973-8984.2003
30. Nix WA, Oberste MS, Pallansch MA. Sensitive, seminested PCR amplification of VP1 sequences for direct identification of all enterovirus serotypes from original clinical specimens. J Clin Microbiol. 2006;44:2698–704. DOI: 10.1128/JCM.00542-06
31. Kistler AL, Webster DR, Rouskin S, Magrini V, Credle JJ, Schnurr DP, et al. Genome-wide diversity and selective pressure in the human rhinovirus. Virol J. 2007;4:40. DOI: 10.1186/1743-422X-4-40
32. Simmonds P. Recombination and selection in the evolution of picornaviruses and other mammalian positive-stranded RNA viruses. J Virol. 2006;80:11124–40. DOI: 10.1128/JVI.01076-06
33. Smura T, Blomqvist S, Paananen A, Vuorinen T, Sobotova Z, Bubovica V, et al. Enterovirus surveillance reveals proposed new serotypes and provides new insight into enterovirus 5′-untranslated region evolution. J Gen Virol. 2007;88:2520–6. DOI: 10.1099/vir.0.82866-0
34. Oberste MS, Maher K, Michele SM, Belliot G, Uddin M, Pallansch MA. Enteroviruses 76, 89, 90 and 91 represent a novel group within the species *Human enterovirus A*. J Gen Virol. 2005;86:445–51. DOI: 10.1099/vir.0.80475-0
35. Junttila N, Leveque N, Kabue JP, Cartet G, Mushiya F, Muyembe-Tamfum JJ, et al. New enteroviruses, EV-93 and EV-94, associated with acute flaccid paralysis in the Democratic Republic of the Congo. J Med Virol. 2007;79:393–400. DOI: 10.1002/jmv.20825
36. Norder H, Bjerregaard L, Magnius L, Lina B, Aymard M, Chomel JJ. Sequencing of 'untypable' enteroviruses reveals two new types, EV-77 and EV-78, within human enterovirus type B and substitutions in the BC loop of the VP1 protein for known types. J Gen Virol. 2003;84:827–36. DOI: 10.1099/vir.0.18647-0
37. Witso E, Palacios G, Cinek O, Stene LC, Grinde B, Janowitz D, et al. High prevalence of human enterovirus a infections in natural circulation of human enteroviruses. J Clin Microbiol. 2006;44:4095–100. DOI: 10.1128/JCM.00653-06
38. Jiang P, Faase JA, Toyoda H, Paul A, Wimmer E, Gorbalenya AE. Evidence for emergence of diverse polioviruses from C-cluster coxsackie A viruses and implications for global poliovirus eradication. Proc Natl Acad Sci U S A. 2007;104:9457–62. DOI: 10.1073/pnas.0700451104
39. Domingo E, Martin V, Perales C, Escarmis C. Coxsackieviruses and quasispecies theory: evolution of enteroviruses. Curr Top Microbiol Immunol. 2008;323:3–32. DOI: 10.1007/978-3-540-75546-3_1
40. Lukashev AN. Role of recombination in evolution of enteroviruses. Rev Med Virol. 2005;15:157–67. DOI: 10.1002/rmv.457

Address for correspondence: Caroline Tapparel, Laboratory of Virology, Division of Infectious Diseases, University of Geneva Hospitals, 24 Rue Micheli-du-Crest, 1211 Geneva 14, Switzerland; email: caroline.tapparel@hcuge.ch

A.2.4 *Insights into rhinovirus genome evolution during experimental human and cells infections*

Cordey S, Junier T, Gerlach D, Gobbini F, Farinelli L, Zdobnov EM, Winther B, Tapparel C, and Kaiser L. Insights into rhinovirus genome evolution during experimental human and cells infections. *PLoS ONE* (2010) 5:e10588.

A.2.4.1 *Contributions*

Cordey et al., 2010 presents a study of rhinovirus genome evolution within humans and cell cultures using high-throughput genomic sequencing techniques.

I participated in writing parts of the manuscript and the preparation of the Figs. 2–3.

A.2.4.2 *Main paper*

See pages 258–267 or at:
`http://www.plosone.org/article/info:doi/10.1371/journal.pone.0010588`

OPEN ACCESS Freely available online

PLoS one

Rhinovirus Genome Evolution during Experimental Human Infection

Samuel Cordey[1,2*☯], Thomas Junier[3,4☯], Daniel Gerlach[3,4], Francesca Gobbini[1,2], Laurent Farinelli[5], Evgeny M. Zdobnov[3,4], Birgit Winther[6], Caroline Tapparel[1,2☯], Laurent Kaiser[1,2☯]

1 Laboratory of Virology, Division of Infectious Diseases and Division of Laboratory Medicine, University of Geneva Hospitals, Geneva, Switzerland, **2** Medical School, University of Geneva, Geneva, Switzerland, **3** Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland, **4** Swiss Institute of Bioinformatics, Geneva, Switzerland, **5** Fasteris SA, Plan-les-Ouates, Switzerland, **6** Division of General Pediatrics, Department of Pediatrics, University of Virginia, Charlottesville, Virginia, United States of America

Abstract

Human rhinoviruses (HRVs) evolve rapidly due in part to their error-prone RNA polymerase. Knowledge of the diversity of HRV populations emerging during the course of a natural infection is essential and represents a basis for the design of future potential vaccines and antiviral drugs. To evaluate HRV evolution in humans, nasal wash samples were collected daily for five days from 15 immunocompetent volunteers experimentally infected with a reference stock of HRV-39. In parallel, HeLa-OH cells were inoculated to compare HRV evolution in vitro. Nasal wash in vivo assessed by real-time PCR showed a viral load that peaked at 48–72 h. Ultra-deep sequencing was used to compare the low-frequency mutation populations present in the HRV-39 inoculum in two human subjects and one HeLa-OH supernatant collected 5 days post-infection. The analysis revealed hypervariable mutation locations in VP2, VP3, VP1, 2C and 3C genes and conserved regions in VP4, 2A, 2B, 3A, 3B and 3D genes. These results were confirmed by classical sequencing of additional samples, both from inoculated volunteers and independent cell infections, and suggest that HRV inter-host transmission is not associated with a strong bottleneck effect. A specific analysis of the VP1 capsid gene of 15 human cases confirmed the high mutation incidence in this capsid region, but not in the antiviral drug-binding pocket. We could also estimate a mutation frequency in vivo of 3.4×10^{-4} mutations/nucleotides and 3.1×10^{-4} over the entire ORF and VP1 gene, respectively. In vivo, HRV generate new variants rapidly during the course of an acute infection due to mutations that accumulate in hot spot regions located at the capsid level, as well as in 2C and 3C genes.

Citation: Cordey S, Junier T, Gerlach D, Gobbini F, Farinelli L, et al. (2010) Rhinovirus Genome Evolution during Experimental Human Infection. PLoS ONE 5(5): e10588. doi:10.1371/journal.pone.0010588

Editor: Darren P. Martin, Institute of Infectious Disease and Molecular Medicine, South Africa

Received March 22, 2010; **Accepted** April 21, 2010; **Published** May 11, 2010

Funding: This study was supported by the Swiss National Science Foundation (grants 3200B0-101670 to L.K. and 3100A0112588/1 to E.Z.), the Department of Medicine of the University Hospitals of Geneva (fonds de péréquation, PRD 06-II-06), the University of Geneva Dean's Program for the Promotion of Women in Science (C.T.), and the Infectigen Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: samuel.cordey@hcuge.ch

☯ These authors contributed equally to this work.

Introduction

Human rhinoviruses (HRV) are the most frequent cause of respiratory infection in humans [1]. These viruses belong to the *Picornaviridae*, one of the oldest and most diversified human virus family, characterized by a non-enveloped, single positive-stranded RNA genome. Although rhinovirus replication is often restricted to the upper respiratory tract leading to self-limited illnesses of short duration, such as the common cold, HRV can also invade the lower respiratory tract and lead to more serious infections [2,3].

Similar to many other RNA viruses, the error-prone rhinoviral polymerase can accumulate a large number of nucleotide mutations over a very short period of time, a feature that favors viral adaptation. The error rate of picornavirus RNA polymerases has been estimated to range between 10^{-3} and 10^{-4} errors/nucleotide/cycle of replication [4,5]. This variability is a driving force for virus evolution and results in a large genetic and phenotypic diversity illustrated by the very high number of different HRV serotypes identified to date (http://www.picornaviridae.com/enterovirus/enterovirus.htm). As for other RNA viruses, the adaptive immune-mediated positive selection, which targets the capsid region for rhinoviruses, is probably one of the main HRV evolutionary forces at both the intra- and inter-host levels. The in vivo selection of resistant variants during exposure to anti-VP1 agents confirms the virus ability to rapidly mutate the capsid protein while still conserving replicative fitness [6–8]. The different environmental conditions of the upper and the lower respiratory tracts could also impact on internal genes. These observations are consistent with the full-length genome data of all HRV serotypes that show the capsid genes VP2, VP3 and VP1 as the least conserved [9].

Rhinovirus species and their respective serotypes have thus emerged due to the ability to accumulate a large number of mutations along the whole genome while preserving both replicative capacity and transmissibility. Opportunities for these mutational events likely occur within the context of each human infection that is usually limited to only a few days. Therefore, it is

important to study the patterns and kinetics of viral genome evolution during the course of HRV infection in individuals, although this does not take into account recombination events that are also used by rhinoviruses to generate new species [10]. In addition, it is likely that the respiratory mucosal surface is exposed to a cloud of different variants of a quasispecies after deposition of infectious droplets. This raises the question of whether rhinovirus infection in humans results from the selection of a given clone among the quasispecies (bottleneck effect) or from concomitant infection by several different variants that are part of the transmitted quasispecies population. Rhinovirus-positive clinical samples collected in humans with naturally-acquired respiratory disease are not suitable for such a study as the time elapsed since the beginning of the infection remains unknown. For this reason, we took advantage of samples from experimentally inoculated adult volunteers to assess the genome evolution over a 5-day course of infection, which is likely to represent the peak window period of transmissibility. In parallel, we infected HeLa-OH cells with the same HRV-39 inoculum to compare in vitro and in vivo adaptation.

The classical Sanger sequencing method usually reliably detects viral variants present at a frequency of at least 20% within a heterogeneous virus population [11,12]. However, the development of ultra-deep sequencing technologies, such as pyrophosphate-based sequencing (pyrosequencing) or reversible chain-terminator extension, now allows efficient detection of viral variants present in only 1–2% of the population [13]. In the present investigation, we used both methods and compared results.

The aim of our study was to describe the HRV-39 genome evolution over a 5-day period in its natural host and to analyze the kinetics of minority mutations (i.e. low-frequency mutations present between 2–50% within the viral population) in vivo and in vitro following inoculation with a quasispecies cloud of viral variants. This allowed us to point out regions along the HRV-39 open reading frame (ORF) that were enriched for mutations or conserved, and to determine whether the HRV genome evolution after 5 days of infection shares similarities with the general long-term HRV evolution history. Collectively, these results should contribute to evaluate the ability of HRV to generate new variants, as well as their ability to escape antivirals and vaccines.

Results

Viral load and HRV-39 kinetics infection in inoculated subjects

Nineteen human volunteers were infected with a standardized quantified inoculum (~1000 50% tissue culture infective dose/mL) of an HRV-39 viral stock. We then collected nasal wash (NW) samples from days 0 to 5 to assess the kinetics of the HRV-39 viral load by real-time PCR. Four of 19 subjects were found to be rhinovirus RNA- positive the day prior to inoculation, either from a recent or ongoing infection, and were excluded from this analysis. As expected, real-time PCR performed on NW samples were positive for all but one subject at day 5. Although differences in HRV-39 viral loads were observed between individuals, viral titers peaked consistently between days 2 and 3 before starting to decline (Fig. 1). These results confirm previous observations with HRV-16 inoculation that relate a similar peak 48 h post-infection correlating with the peak of symptom scores [14].

Mutation analysis along the whole ORF by ultra-deep and classical sequencing methods

Generation of new mutations in the course of infection and adaptation to a specific host. The dynamics of HRV-39 minority mutation evolution were analyzed by ultra-deep sequencing of three samples collected at day 5 post-infection (two clinical, one in vitro) by comparing the entire ORF of the initial HRV-39 inoculum with the sequence obtained in two human subjects (P1073 and P1077) and in HeLa-OH cells (HeLa-A). To exclude mutation introductions linked to RT or DNA polymerase errors, each RNA extracted from human NW or HeLa-OH cell samples was reverse transcribed in duplicate and amplified by PCR in quadruplicate. The presence of single nucleotide mutations identified by the ultra-deep sequencing approach was considered only if statistically reproducible in each of our replicates (see Materials and Methods).

A Venn diagram provides an overview of the repartition of minority mutations in the four different viral populations analysed and shows individual overlaps (Fig. 2A). The total number of minority mutations detected in the four viral populations was 45. The number of minority mutations detected in the original inoculum was 32, thus revealing that 13 additional minority mutations appeared in the three populations studied after 5 days of viral replication. Most minority mutations present in the inoculum were detected in at least one of the two subjects studied (25/32) with 14 kept within the two viral populations (Fig. 2A, B), thus suggesting that most initial variants had passed in these infected volunteers. Similar results were observed in HeLa-OH cells (21/32 minority mutations conserved, Fig. 2A, B). Viral populations present in NW samples of subjects P1073 and P1077 and the cell supernatant were composed of 24, 26 and 26 minority mutations, respectively, with 21%, 23% and 19% representing new minority mutations having emerged during the 5-day infection period.

During the course of infection, four minority mutations present in the inoculum were counter-selected (one in VP1, one in 2C NTPase A site [15], two in 3D) and two, five and three appeared specifically in patients P1073 (one in VP1 NIm-IB domain [16], one in 2C NTPase A site), P1077 (one in VP2, one in 2C and three in 3D) and in HeLa-A (two in 2C NTPase A site, one in 2C zinc finger motif [17]), respectively, and were present at similar frequencies compared to all minority mutations. Finally, two minority mutations were present both in HeLa-A and P1073 (one in VP1, one in 3C) and one in both P1073 and P1077 (located in 2C). Overall, we did not observe a significant difference in terms of the number of new mutations appearing in human and HeLa-OH cells.

Distribution of mutations along the ORF. We estimated the initial and final densities of minority mutations along the ORF with kernel functions using the ultra-deep sequencing data and were able to identify regions with a higher density of mutations (Fig. 3). The distribution and pattern of these spots along the ORF were similar between the inoculum and the viral populations analyzed after 5 days of replication in humans or cells. Most minority mutations are located within VP2, the second half of VP1, and the 2C genes (see filled and dashed curves, Fig. 3A, B, C). Interestingly, no minority mutations were detected within the VP1 drug-binding pocket [6], thus suggesting a relative stability of this structure, at least in the absence of specific pressure. This analysis also identified the presence of regions that were highly conserved: in VP4, from the start of 2A to the end of 2B; in 3A, 3B genes; and finally in 3D (the polymerase gene).

Comparison of the P1073, P1077 or HeLa-A sample sequences with that of the initial HRV-39 inoculum was then performed to determine both the presence of the new majority species (defined as a frequency change of over 50%) and the amplitude of all mutation frequency changes in the population after 5 days' infection (Fig. 3A, B, C, colored bars). This analysis shows the presence of three majority nucleotide changes along P1073 ORF,

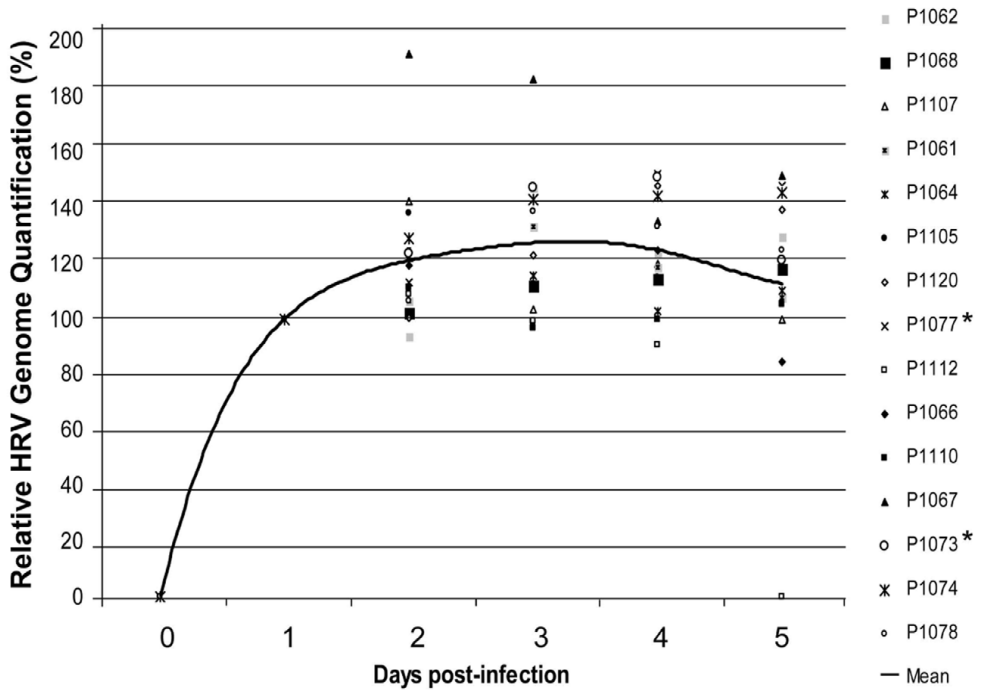


Figure 1. Kinetics of HRV-39 infection in inoculated patients. The course of infection is shown from days 0 to 5 for each patient (P). The relative HRV RNA values are expressed as $1/C_T$ and converted into % with 100% being arbitrarily set for each patient at day 1. The mean value is represented by the black line. Subjects analyzed further by ultra-deep sequencing are marked with an asterisk.
doi:10.1371/journal.pone.0010588.g001

none for P1077, and six for HeLa-A. Both the amplitude and the number of mutation (minority and majority mutations) frequency changes were similar between P1073, P1077 and HeLa-A and mainly located within regions previously identified as being enriched of mutations. Interestingly, all 14 minority mutations present in the four viral populations in Fig. 2 are mainly located within these highly variable regions (VP1 [4/14], VP2 [2/14], 2C [5/14], VP3 [2/14] or 3C [1/14]), but not in regions depleted of mutations identified above.

Majority species after five days of infection. We performed also Sanger sequencing in parallel on the entire ORF of the specimens analyzed above, as well on NW samples from three additional patients (P1062, P1120 and P1074) and cell supernatants from four additional experiments (HeLa-B to E). Again, consensus sequences after 5 days of viral replication were compared to the inoculum sequence (Fig. 4A). This analysis confirmed the ultra-deep sequencing as all the nucleotide changes identified previously for P1073, P1077 and HeLa-A samples are strictly those identified by the Sanger sequencing. A total number of four non-synonymous (two non-conservative, present both in P1073 at the full-length positions 2647 in VP1 and 4831 in 2C) and seven synonymous mutations were found in samples from human subjects. Similar to the ultra-deep analysis, most mutations are located within the viral capsid genes VP1 (3/11), VP2 (2/11) and VP3 (3/11), while VP4 (0/11), 2A (0/11), 2B (0/11), 3A (0/11), 3B (0/11) and 3D (1/11) regions were confirmed to represent relatively conserved regions with few or no mutations identified.

Fourteen non-synonymous (six non-conservative: one present in HeLa-D at the HRV-39 full-length position 2464 in VP1, and one present in the five HeLa-OH experiments at position 4831 in 2C) and 13 synonymous mutations were found in cells. Again, most mutations (15/27) are located within the viral capsid genes VP2, VP3 and VP1, while no mutations were present in VP4, 2A, 3A, 3B, and only one in 2B and 3D. None of the mutations observed both in vivo and in vitro is located in HRV domains with known functions [15–21].

Finally, taking into account the number of patients or cells analysed and the length of the ORF, we estimate that the occurrence of major mutations after 5 days of HRV-39 infection in human subjects is of 3.4×10^{-4} mutations/nucleotides (total of 11 mutations/[5 subjects ×6443 nucleotides analysed]) versus 8.4×10^{-4} in cells (total of 27 mutations/[5 HeLa-OH assays × 6443 nucleotides analyzed]).

In vivo and in vitro VP1 sequence analysis

The capsid protein VP1 is responsible for receptor binding and is also a target for certain antivirals. In addition, this protein is the most exposed viral protein at the capsid surface and the main inducer of neutralizing antibody. As this gene was identified as a region with a high frequency of mutations in our previous analysis, we assessed by classical sequencing the frequency of point mutations in natural HRV hosts by comparing the VP1 consensus sequences at day 5 post-HRV-39 infection versus the initial sequence present in the HRV-39 inoculum for all 15 subjects.

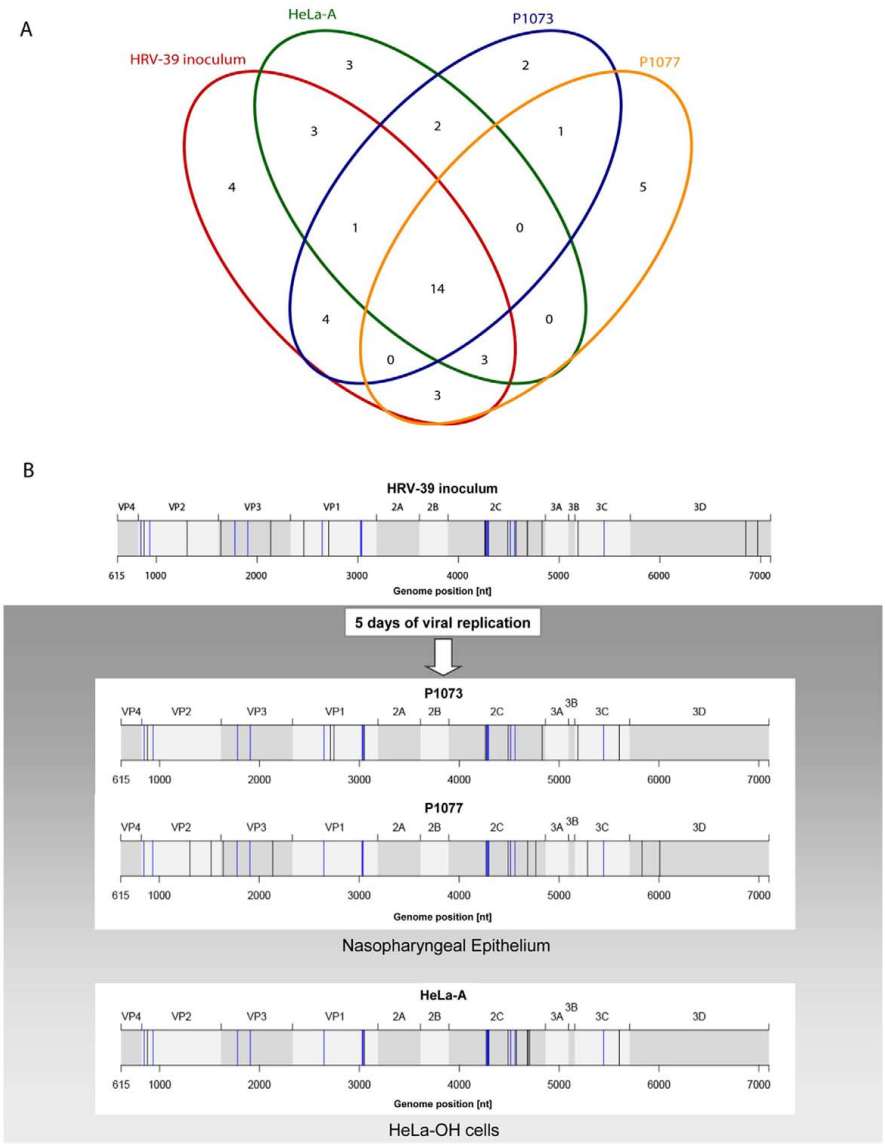


Figure 2. Representation of all specific or common minority mutations to be found in the same loci for the four samples analysed by ultra-deep sequencing. (A) Venn plot showing the minority mutations present initially in the HRV-39 inoculum and those present in HeLa-A, P1073 and P1077 after 5 days of infection. (B) Minority mutations present in the initial inoculum and in subjects P1073 and P1077 after 5 days of viral replication in the nasopharyngeal epithelium are represented by black and blue at their respective positions along the HRV-39 ORF. Blue bars represent the 14 minority mutations present in the Venn plot in the four viral populations.
doi:10.1371/journal.pone.0010588.g002

Again, the experiments were repeated in duplicate to exclude any mutation introduced by RT-PCR or sequencing. The sequence comparison revealed the occurrence of a single point mutation in three of the 15 subjects, one synonymous at HRV-39 full-length position 2711 for two subjects (P1105 and P1073) and one non-synonymous/non-conservative at position 3162 for one subject (P1062) (Table 1). In addition, a mixed population was found at position 2647. Thus, we can estimate that the VP1 mutation rate

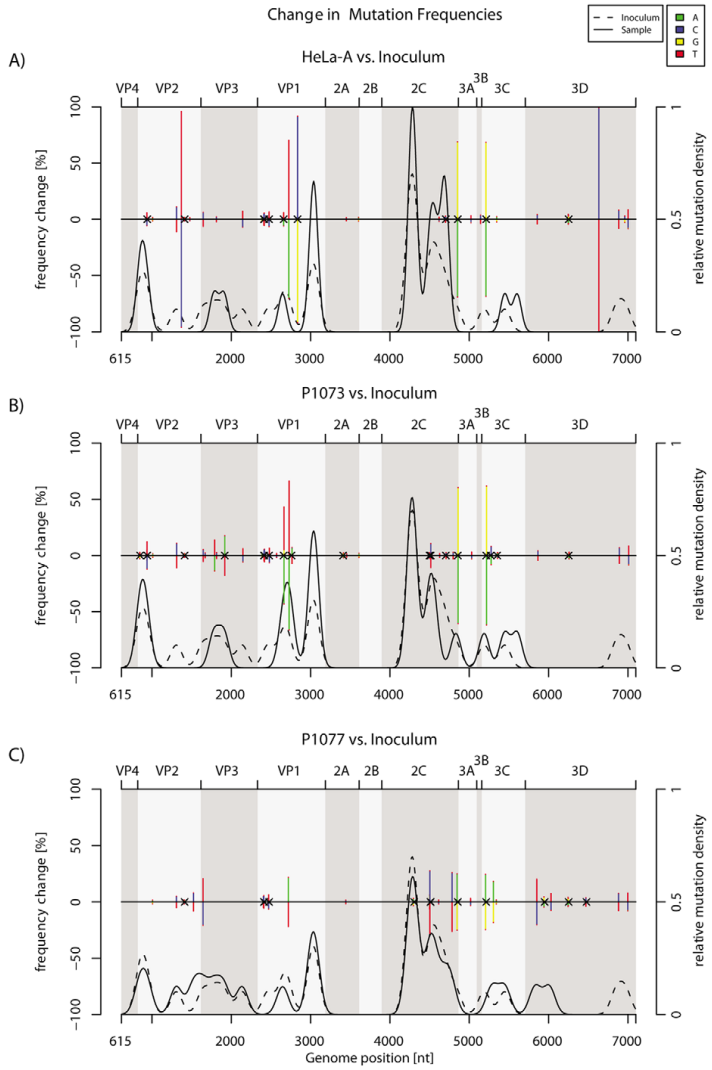


Figure 3. Change in mutation frequencies (minority and majority mutations, colored bars) and minority mutation densities (curves). Colored bars represent the difference in proportions of each nucleotide between the inoculum and the final (5 days' post-infection) sample in HeLa (A), subjects P1073 (B) and P1077 (C). As each gain in proportion by one nucleotide must be a loss by another, the changes sum to zero. Crosses indicate non-synonymous mutations. Curves indicate minority mutation densities (estimated by a Gaussian kernel function), including mutations whose nucleotide proportions did not change between the inoculum and the final sample.
doi:10.1371/journal.pone.0010588.g003

at day 5 of HRV-39 infection in human subjects is of approximately 3.1×10^{-4} mutations/nucleotides (total of 4 mutations/[15 subjects ×854 nucleotides analyzed]).

In comparison, eight VP1 mutations were identified in the five HeLa-OH cell experiments after 5 days of infection. The synonymous mutation at position 2711, previously observed in vivo for two subjects, was present in all HeLa-OH cells. Each of the three remaining non- synonymous mutations (the one at HRV-39 full-length position 2400 representing a mixed population) was present in only one of the five cell experiments. Based on these data, we estimate that the occurrence of VP1 mutation after 5 days of HeLa-OH cell infection with HRV-39 is of 1.9×10^{-3} mutations/nucleotides (total of 8 mutations/[5 HeLa-OH assays ×854 nucleotides analyzed]). Regarding the complete ORF, VP1 mutation frequency is higher in vitro under our conditions than in vivo in experimentally infected individuals with a reference

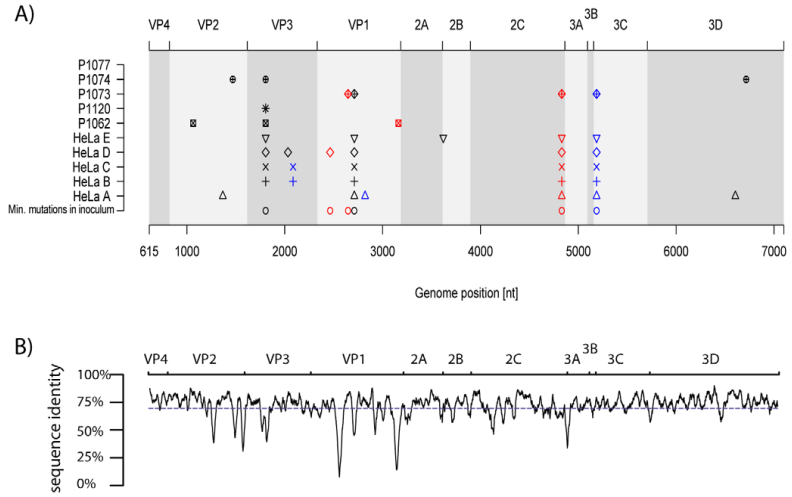


Figure 4. Comparison of ORF consensus sequences in HeLa-OH and human volunteers. (A) Five HeLa-OH (A to E) and samples from five human volunteers (P1062, P1120, P1073, P1074 and P1077) collected 5 days' post-infection were analysed by classical sequencing method. All mutants present after 5 days of infection, as well as those already present as minority mutations in the initial HRV-39 inoculum, are shown at their respective genome position. Synonymous, non-synonymous and non-conservative mutations are represented by black, and blue and red symbols, respectively. (B) The conservation plot was calculated based on an alignment of 99 rhinovirus serotypes as previously published [9]. The average sequence identity for the ORF was 69.9% (blue dashed line).
doi:10.1371/journal.pone.0010588.g004

strain. Of note, none of the mutations observed both in vivo and in vitro is located in neutralizing immunogens or within the VP1 drug-binding pocket. No mutations were found in the cis-acting replication element, which is located in the 2A gene for human rhinovirus A [22].

Discussion

The study of HRV genome evolution is limited by the lack of animal models and the short duration of most human infections. To bypass this limitation, we analyzed the HRV genome evolution in experimentally infected human volunteers and compared our results with those obtained in an in vitro culture model. Our analysis based on 15 HRV-39 inoculated subjects first confirms that the peak viral load is reached at days 2 or 3 post-infection, as shown previously in HRV-39- or HRV-16-inoculated subjects. This correlates also with the kinetic of common cold symptoms [14,23,24]. Relatively high levels of viral RNA are still found 5 days' post-infection, although substantial inter-individual variations were observed and suggest that factors including the host innate immune responses (all individuals presented a serum neutralizing antibody titre of 1:4 or less against the reference HRV-39 strain suggesting a minor effect of the adaptive-immune responses) could modulate viral replication. Most importantly, our analysis was conducted during a period of time that corresponds to the main period of transmissibility [25] and during a high rate of replication cycles.

By using ultra-deep sequencing technology, we were able to pinpoint HRV evolution at the level of a quasispecies population both in vivo and in vitro. Our data illustrate the ability of

Table 1. Analysis of VP1 consensus sequences.

Mutations	Nucleotide	HeLa-OH					Subjects			Amino Acid
(FL position)	in Inoculum	A	B	C	D	E	P1062	P1073	P1105	
2400	C		***C:T*** (50:50)							***VP124 H→Y***
2464	T				***C***					***VP145 I→T***
2647	A							***A:T*** (50:50)		***VP1106 Q→L***
2711	A	T	T	T	T	T		T	T	VP1127 (P)
2821	G	**C**								**VP1164 G→A**
3162	T						***C***			***VP1278 S→P***

Majority mutations present in VP1 after 5 days of HRV-39 infection in five independent HeLa-OH cell experiments (A to E) and in 15 inoculated human volunteers are represented. Non-synonymous mutations are in bold and non-conservative in bold and italic. The nucleotide and amino acid positions are indicated relative to the HRV-39 reference sequence (Genbank accession # AY751783), complete genome and VP1 protein, respectively.
doi:10.1371/journal.pone.0010588.t001

rhinoviruses to produce several new variants as rapidly as 5 days' post-infection. This represents a minimal estimate of replicating minority variants since mutations present at frequencies lower than the ultra-deep sequencing limit of detection are not represented here.

The transmission of viral populations between hosts and the subsequent mucosal infection could lead to a bottleneck effect that selects for a limited number of variants. Indeed, for other RNA viruses, such as HIV, infection might result from the selection of very few clones and it seems likely that as few as 1 to 5 viral particles could initiate an infection [26]. This effect was not observed for HRV since our ultra-deep sequencing analysis shows that most minority mutations present in the initial HRV-39 inoculum were transmitted and replicated in both human subjects and HeLa-OH cells (Fig. 2). Thus, based on our experimental conditions, it appears that HRV inter-host transmission is not associated with a strong bottleneck effect. We observed simultaneous co-infection with several minority variants together with the dominant population.

To the best of our knowledge, our study allowed to define for the first time a virus mutational map on the entire ORF in vivo. Both classical and ultra-deep sequencing methods revealed the presence of mutation hot spots along the entire coding sequence in the viral capsid genes VP1, VP2 and VP3, whereas cold spots were found in VP4, 2A, 2B, 3A, 3B and 3D. Interestingly, this is consistent with previous published data that compared the RNA genome sequence of prototype HRVs serotypes (Fig. 4B) [9]. A previous analysis based on the study of 35 HRV full-length sequences has already demonstrated that the HRV genome was under purifying selective pressure with islands of diversifying pressures located in VP1, VP2, VP3, 3C and 3D genes [27]. Our results obtained after 5 days of viral replication in human subjects and HeLa-OH cells confirm that VP1, VP2, VP3 and 3C, but also 2C genes, appeared to be under a diversifying pressure compared to others parts of the ORF (Fig. 2, 3, 4). A similar evolutionary pattern was observed in experimentally infected humans as well as in vitro, thus suggesting an evolutionary pattern common to the HRV species. Whether the presence of these mutation hot spots are linked to host immune pressure and/or to intrinsic replication constraints of the virus remains to be elucidated. However, knowledge of hot and cold spots may contribute to the identification of stable targets for new antiviral and vaccine therapies. This also contributes to our understanding of the diversity of one of the most frequent agents infecting humans.

When analyzing the occurrence of mutations in VP1 gene in both inoculated human subjects and HeLa-OH cells, we found a lower mutation frequency in vivo in humans (3.1×10^{-4}) than in vitro in HeLa-OH cells (1.9×10^{-3}). The mutation frequency estimations based on the ORF sequences strengthened these results (3.4×10^{-4} and 8.4×10^{-4} in humans and HeLa-OH cells, respectively) and are consistent with published data that estimate an error rate of picornavirus RNA polymerases ranging between 10^{-3} and 10^{-4} errors/nucleotide/cycle of replication [4,5]. HRV is one of the most frequent viral agents in humans with more than 100 serotypes co-circulating that have emerged through repeated infections of short duration. The mutation frequency observed in our study is rather limited and at the lower end of what might have been expected, thus suggesting that HRV serotypes (at least for HRV-39) have already evolved over a prolonged period of time. Furthermore, in the absence of any knowledge of the number of replication cycles, we should be careful to draw any definitive conclusion between in vitro and in vivo conditions. Still, these differences could be explained by the fact that the HRV-39 inoculum might be more human-adapted as it was previously cultured only twice in WI-38 diploid fibroblasts. Second, human volunteers and HeLa-OH cells were not inoculated with the same $TCID_{50}$/mL (10^3 and 10^2 $TCID_{50}$/mL, respectively) as HeLa-OH cells did not sustain 5 days of infection as with the one used to infect human subjects. Third, the relative percentage of infected cells is likely to be higher in vitro.

The Sanger sequencing analysis performed on the entire ORF of 5 HeLa-OH and 5 NW samples showed a similar ratio of synonymous and non-synonymous mutations mainly located in hot spot regions identified by the relative mutation density analysis. Both in HeLa-OH and in human subjects, half of the non-synonymous mutations were non-conservative and were all located in VP1 and 2C genes. Interestingly, the change observed at position 2711 seems to have a beneficial effect at the genome level (synonymous) in HeLa-OH cells since this mutation emerged in all five independent experiments in HeLa-OH. This was observed only twice among the 15 human cases. Only functional studies could determine whether this specific mutation could provide a phenotypic advantage in vitro. Our data demonstrate the intrinsic ability for an already human-adapted HRV to generate new mutations in the VP1 region during the course of an acute infection in vivo. This latter point is of importance as VP1 contains sites for both antigen recognition and antiviral drug targets, such as pleconaril [6,7]. Interestingly, after 5 days of infection, no mutations arose within the VP1 drug-binding pocket both in vivo and in vitro, but in the absence of any drug pressure.

Importantly, we did not find any mutations by both ultra-deep or classical Sanger sequencing methods within any previously known HRV functional domains (that are expected to tolerate a minimal number of mutations) and these were largely located in viral structural genes. This suggests that ultra-deep sequencing might represent a powerful tool to identify previously unknown functional domains that should be evidenced as cold spot regions. In addition, this approach is extremely useful for the study of any viral populations and quasispecies that cannot be grown in cells. Finally, this approach performed on different HRV serotypes could bring information on their respective evolutionary status as viruses having frequently circulated in humans should be more adapted and thus less susceptible to variations during the course of a human infection.

In summary, we took advantage of samples from experimentally inoculated volunteers to characterize HRV genome evolution over a 5-day course of infection, which represents the maximal window period of transmissibility. Ultra-deep sequencing analysis on minority mutation frequency and distribution allowed to identify hot spot and cold spot regions along HRV-39 ORF present both in vivo and in vitro. Our experiments suggest that HRV inter-host transmission is not associated with a strong bottleneck effect. Continued efforts to improve our understanding of HRV evolution in its natural host are essential as they represent the basis for the design of future potential vaccines and antiviral drugs.

Materials and Methods

Ethics statement

Written informed consent was obtained from all individuals prior to study participation. The study was approved by the Institutional Review Boards (IRB) of the University of Virginia, Charlottesville, Virginia, and the Medical University of South Caroline, Charleston, South Carolina.

Study participants

Subjects were enrolled in a clinical study after informed consent and following review of the IRB at the University of Virginia,

Charlottesville. The study aimed to assess the effect of an oral antiviral compound and only placebo-treated subjects were considered for the present study. Individuals were required to be previously healthy, between 18 to 65 years of age, and to present a serum neutralizing antibody titre of 1:4 or less against the reference HRV-39 strain. Exclusion criteria were a history of allergic disease or nonallergic rhinitis, abnormal nasal anatomy or mucosa, or a clinically diagnosed respiratory tract infection in the previous two weeks. Pregnant or lactating women or women not taking medically approved birth control were also excluded.

HRV-39 inoculum

The HRV-39 strain, commonly used for human inoculation studies, was initially recovered from a volunteer and cultured twice in WI-38 diploid fibroblast cultures according to standard recommendations to obtain a sufficient stock. The inoculum was tested to be safe for human in vivo usage [28].

We analyzed nasopharyngeal specimens of 19 consecutive subjects inoculated with viral stocks of ~1000 50% tissue culture infective dose ($TCID_{50}$)/mL administered as drops in two inocula of 250 µl per nostril given approximately 15 minutes apart while subjects were supine. NW samples were collected daily for 5 days by instillation of 5 mL of 0.9% saline into each nostril and stored at −80°C for subsequent assay.

Quantitative real-time RT-PCR

HRV-39 RNA was TRIzol-extracted (Invitrogen, Carlsbad, CA, USA) from 190 µl of NW samples collected daily or HeLa-OH infected supernatant according to the manufacturer's instructions. As an internal control, 10 µl of standardized Canine Distemper Virus (CDV) of known concentration were added to each sample before extraction. Extracted RNA was used as a template for the synthesis of cDNA with random hexamers (Roche, Indianapolis, IN, USA) or oligo-dT primers (Roche) at 42°C using the reverse transcriptase (RT) SuperScript II (Invitrogen) according to manufacturer's instructions. cDNA was then amplified and detected in a TaqMan real-time PCR reaction using a validated human picornavirus combination of primers and probes named "Panenterhino" [29] and the CDV assay (primers: CDV-fwd: 5′-gctacccaagaaaccgtcattg-3′, CDV-rev: 5′-gcatggcagggacgagtt-3′, probe: CDV-probe: 5′-VIC-cgttcagggagtccaggactacgtcaac-TAMRA-3′), respectively, under the following cycling conditions: 50°C for 2 min; 95°C for 10 min; 55 cycles of 95°C for 15 s and 60°C for 1 min in a 7500 Applied Biosystems thermocycler. Results were analyzed using the SDS version 1.4 program (Applied Biosystems, Foster City, CA, USA). Quantitative assays were run using a 10-fold dilution series of a titrated HRV-39 stock (ATCC) that was used as a reference quantitative curve for each run. In addition, we obtained Ct values for the CDV assay in each run to control for intra- and inter-assay variability.

Cell culture and infection

HeLa-OH cells were grown in Eagle's Minimum Essential Medium (EMEM; Lonza, Wokingham, UK) supplemented with 2 mM L-glutamine, 1 µg/mL amphotericin, 100 µg/mL gentamicin, 20 µg/mL vancomycin, and 10% fetal calf serum (FCS) at 37°C in a 5% CO_2-containing atmosphere. The original HRV-39 inoculum (10^6 $TCID_{50}$/mL) used to inoculate the human subjects was diluted 10^4-fold in 1 mL (10^2 $TCID_{50}$/mL) of McCoy's 5A Medium–2% FCS to infect, with this 1 mL, 80% confluent HeLa-OH in 33 mm wells for 2 h at 33°C. Cells were washed twice with PBS (Ca^{2+} and Mg^{2+} free). Finally, 1 mL of fresh McCoy's 5A Medium–2% FCS was added in each well and cells were incubated at 33°C. Supernatants were recovered and extracted 5 days' post-infection for real-time PCR analysis and sequencing. This procedure was repeated in five independent experiments.

VP1 gene and full-length genome amplification by PCR

VP1 amplicon was obtained by amplification of two overlapping PCR fragments (primers $VP1_{2195}$/$VP1_{2860}$ and $VP1_{2626}$/$VP1_{3325}$, Figure S1). The HRV-39 ORF sequence (nucleotide 615 to 7058) was obtained by amplification of eight overlapping PCR products with primers HRV39-1 to 8 forward/reverse (Figure S1). For ultra-deep sequencing, RT was performed in duplicate and each PCR in quadruplicate to discard any mutations introduced by polymerase errors. We used Pfx50 DNA polymerase for PCRs according to the manufacturer's instructions. All amplicons were purified with the microcon columns (Millipore, Zug, Switzerland) before sequencing. PCR quadruplicates originating from the same RT reaction were pooled at equimolar concentrations and used for ultra-deep sequencing.

Both VP1 and ORF classical sequencing were performed in duplicate directly on PCR products with the same primers used for each individual PCR. Sequencing was performed with ABI Prism 3130XL DNA Sequencer (Applied Biosystems). Chromatograms were imported for proofreading with the vector NTI Advance 10 program (Invitrogen). Overlapping fragments were assembled with the contigExpress module of the vector NTI Advance 10.

Ultra-deep sequencing analysis

Samples. Libraries were prepared according to the manufacturer's protocol (Illumina, Inc., San Diego, USA) using bar-coded adapters designed by Fasteris. Each library consists of eight PCR products pooled at equimolar concentration and fragmented by nebulization. After end repair to generate blunt ends and the addition of one A at the 3'ends, fragments were ligated with a modified genomic adapter containing a four-base bar-code at its 3'end and purified on agarose gel to recover fragments of approximately 300 bp. PCR amplification was performed for 15 cycles using the Phusion polymerase. Libraries were purified and quality controlled by cloning a 1 µl aliquot into a pCR4Blunt-TOPO plasmid (Invitrogen) and capillary sequencing of eight clones to verify correct constructs and inserts. We quantified the libraries using BioAnalyzer (Agilent, USA) and Q-bit (Invitrogen) and diluted to 10 nM.

Genome analyzer run. Libraries were pooled and sequenced on an Illumina Genome Analyzer GAII single-read channel for 76 cycles using a version 3 sequencing kit. We performed base-calling using Illumina GAPipeline-1.3.2, which produced over four million pass filter reads or 329 Mb.

Bioinformatics data analyses. An average of 3.5×10^5 HRV-39 mapped reads was obtained for each sample analyzed with a mean coverage of 4100 readings per nucleotide. Short reads were first mapped with MAQ software version 0.7.1 accepting a maximum of two mismatches within the first 24 bases on the HRV-39 reference sequence (Genbank accession AY751783). Reads mapping on repeated regions were attributed randomly to one of the possible locations. The MAQ consensus sequences, including SNP detection, were generated for all regions with a minimum coverage of three bases. We conducted a statistical analysis of the counts of the number of mapped A/C/G/T to extract potential SNP positions for each position on the reference sequence. Counts are used to determine the 95% confidence interval of the probability of observing A/C/G/T at the position while assuming the probabilities to follow a beta distribution. When the confidence interval of two of the bases probability is above a 5% threshold, the position is considered as a statistically

significant SNP. De novo assembling was performed using Velvet version 0.7.31 and the resulting contigs were compared using dnadiff (MUMmer version 3.20) with the MAQ consensus sequence to validate minority mutations and discover insertions-deletions (indels). All reads were also mapped with DNASTAR NGen on the HRV-39 reference sequence and on capillary sequences obtained in our laboratories. We performed visualization and minority mutation detection using DNASTAR SeqMan version 8.0.

Venn diagram. Significant mutations showing common loci in the four different samples and their individual overlaps were visualized in a Venn diagram using the statistical package R [30] and the R script overLapper.R (http://faculty.ucr.edu/~tgirke/Documents/R_BioCond/My_R_Scripts/overLapper.R; accessed February 1, 2010).

Mutation analysis along the whole HRV-39 ORF. The variation in relative frequency of a given base at a specified position is expressed as "final relative frequency–initial relative frequency" for that base at that position. The relative frequency of a given base at a given position is expressed as the base's count divided by the count of bases at that position. The algebraic sum of all variations at a given position is always zero. Density of mutations, including those which did not exhibit change in base frequencies, was represented with a Gaussian kernel density function using a smoothing band width of 0.1 kernel standard deviation. The curves are shown at the same scale and normalized so that a value of 1 is the highest density found over all genomes. Graphs, including kernel estimates, were produced with the R statistical package.

HRV ORF conservation plot. We conducted an RNA sequence identify plot for the ORF of an alignment covering all known human rhinoviruses (pmid:19213880). The plot was drafted using R [30] and the packages bio3d [31] and zoo [32]. All-against-all identities were assigned for each column and average; sliding windows of 30 nucleotides with a step size of one nucleotide were used to compute a moving mean sequence identity percentage that was then plotted.

Supporting Information

Figure S1 Primers used to amplify and sequence the human rhinovirus 39 capsid protein VP1 and the entire open reading frame.

Found at: doi:10.1371/journal.pone.0010588.s001 (8.94 MB TIF)

Acknowledgments

We would like to thank F.G. Hayden for providing both samples and critical comments on the manuscript. We are also grateful to M. Vignuzzi and L. Perrin for providing helpful comments to improve the manuscript, Lara Turin and Chantal Gaille for technical assistance, and Rosemary Sudan for editorial assistance.

Author Contributions

Conceived and designed the experiments: SC LF CT LK. Performed the experiments: SC TJ DG FG. Analyzed the data: SC TJ DG FG EMZ CT LK. Contributed reagents/materials/analysis tools: BW. Wrote the paper: SC TJ DG LF CT LK.

References

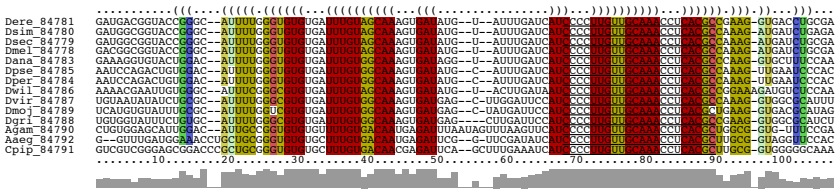
1. Denny FW, Jr. (1995) The clinical impact of human respiratory virus infections. Am J Respir Crit Care Med 152: S4–12.
2. Kaiser L, Aubert JD, Pache JC, Deffernez C, Rochat T, et al. (2006) Chronic rhinoviral infection in lung transplant recipients. Am J Respir Crit Care Med 174: 1392–1399.
3. Papadopoulos NG, Bates PJ, Bardin PG, Papi A, Leir SH, et al. (2000) Rhinoviruses infect the lower airways. J Infect Dis 181: 1875–1884.
4. Drake JW (1999) The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. Ann N Y Acad Sci 870: 100–107.
5. Harvala H, Simmonds P (2009) Human parechoviruses: biology, epidemiology and clinical significance. J Clin Virol 45: 1–9.
6. Ledford RM, Patel NR, Demenczuk TM, Watanyar A, Herbertz T, et al. (2004) VP1 sequencing of all human rhinovirus serotypes: insights into genus phylogeny and susceptibility to antiviral capsid-binding compounds. J Virol 78: 3663–3674.
7. Ledford RM, Collett MS, Pevear DC (2005) Insights into the genetic basis for natural phenotypic resistance of human rhinoviruses to pleconaril. Antiviral Res 68: 135–138.
8. Schmidtke M, Hammerschmidt E, Schuler S, Zell R, Birch-Hirschfeld E, et al. (2005) Susceptibility of coxsackievirus B3 laboratory strains and clinical isolates to the capsid function inhibitor pleconaril: antiviral studies with virus chimeras demonstrate the crucial role of amino acid 1092 in treatment. J Antimicrob Chemother 56: 648–656.
9. Palmenberg AC, Spiro D, Kuzmickas R, Wang S, Djikeng A, et al. (2009) Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. Science 324: 55–59.
10. Tapparel C, Junier T, Gerlach D, Van Belle S, Turin L, et al. (2009) New respiratory enterovirus and recombinant rhinoviruses among circulating picornaviruses. Emerg Infect Dis 15: 719–726.
11. Bushman FD, Hoffmann C, Ronen K, Malani N, Minkah N, et al. (2008) Massively parallel pyrosequencing in HIV research. AIDS 22: 1411–1415.
12. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res 17: 1195–1201.
13. Solmone M, Vincenti D, Prosperi MC, Bruselles A, Ippolito G, et al. (2009) Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. J Virol 83: 1718–1726.
14. Sanders SP, Proud D, Permutt S, Siekierski ES, Yachechko R, et al. (2004) Role of nasal nitric oxide in the resolution of experimental rhinovirus infection. J Allergy Clin Immunol 113: 697–702.
15. Rodriguez PL, Carrasco L (1993) Poliovirus protein 2C has ATPase and GTPase activities. J Biol Chem 268: 8105–8110.
16. Sherry B, Mosser AG, Colonno RJ, Rueckert RR (1986) Use of monoclonal antibodies to identify four neutralization immunogens on a common cold picornavirus, human rhinovirus 14. J Virol 57: 246–257.
17. Pfister T, Jones KW, Wimmer E (2000) A cysteine-rich motif in poliovirus protein 2C(ATPase) is involved in RNA replication and binds zinc in vitro. J Virol 74: 334–343.
18. Binford SL, Maldonado F, Brothers MA, Weady PT, Zalman LS, et al. (2005) Conservation of amino acids in human rhinovirus 3C protease correlates with broad-spectrum antiviral activity of rupintrivir, a novel human rhinovirus 3C protease inhibitor. Antimicrob Agents Chemother 49: 619–626.
19. Lewis-Rogers N, Bendall ML, Crandall KA (2009) Phylogenetic relationships and molecular adaptation dynamics of human rhinoviruses. Mol Biol Evol 26: 969–981.
20. Love RA, Maegley KA, Yu X, Ferre RA, Lingardo LK, et al. (2004) The crystal structure of the RNA-dependent RNA polymerase from human rhinovirus: a dual function target for common cold antiviral therapy. Structure 12: 1533–1544.
21. Petersen JF, Cherney MM, Liebig HD, Skern T, Kuechler E, et al. (1999) The structure of the 2A proteinase from a common cold virus: a proteinase responsible for the shut-off of host-cell protein synthesis. EMBO J 18: 5463–5475.
22. Gerber K, Wimmer E, Paul AV (2001) Biochemical and genetic studies of the initiation of human rhinovirus 2 RNA replication: identification of a cis-replicating element in the coding sequence of 2A(pro). J Virol 75: 10979–10990.
23. Turner RB (2001) Ineffectiveness of intranasal zinc gluconate for prevention of experimental rhinovirus colds. Clin Infect Dis 33: 1865–1870.
24. Tyrrell DA, Cohen S, Schlarb JE (1993) Signs and symptoms in common colds. Epidemiol Infect 111: 143–156.
25. Lessler J, Reich NG, Brookmeyer R, Perl TM, Nelson KE, et al. (2009) Incubation periods of acute respiratory viral infections: a systematic review. Lancet Infect Dis 9: 291–300.
26. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, et al. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U SA 105: 7552–7557.
27. Kistler AL, Webster DR, Rouskin S, Magrini V, Credle JJ, et al. (2007) Genome-wide diversity and selective pressure in the human rhinovirus. Virol J 4: 40.
28. Gwaltney JM, Jr., Hendley O, Hayden FG, McIntosh K, Hollinger FB, et al. (1992) Updated recommendations for safety-testing of viral inocula used in volunteer experiments on rhinovirus colds. Prog Med Virol 39: 256–263.

29. Tapparel C, Cordey S, Van Belle S, Turin L, Lee WM, et al. (2009) New molecular detection tools adapted to emerging rhinoviruses and enteroviruses. J Clin Microbiol 47: 1742–1749.
30. R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available: http://www.R-project.org. Accessed 3 February 2010.
31. Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS (2006) Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22: 2695–2696.
32. Zeileis A, Grothendieck G (2005) zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software 14(6) 1-27. Available: http://www.jstatsoft.org/v14/i06/. Accessed 3 February 2010.

B

miROrtho User's Guide

A computational pipeline for predicting putative microRNA genes in complete genomic sequences



Daniel Gerlach daniel.gerlach@unige.ch
Computational Evolutionary Genomics Group
University of Geneva, Switzerland
http://cegg.unige.ch/mirortho

B.1 INTRODUCTION

B.1.1 *Brief description*

The `miROrtho` pipeline identifies miRNA genes in genomic DNA sequences. It combines homology, comparative genomic and a state-of-the-art machine learning classifier. For a quick installation guide refer to Section B.2.1. To get started, use the files in the `Examples/` directory and follow the tutorial section on page 275.

B.1.2 *What is included in this package?*

This packages contains all the `Perl` scripts, example files and binaries need to run the `miROrtho` pipeline on genomic sequences. An installation of a recent version of the `Perl` interpreter is required (tested with v5.8.8).

B.1.3 *Web resources*

The `miROrtho` package was used to construct a comprehensive database of animal miRNA genes also named `miROrtho` which can be found here: `http://cegg.unige.ch/mirortho`.

B.2 INSTALLATION

B.2.1 *Quick installation instructions*

Several additional software package are required for the pipeline to work. Download, compile and install the following packages. The respective program version on which this pipeline was tested is shown in parentheses.

- Vienna RNA Secondary Structure Package (version 1.8.3) (Hofacker et al., 2002, 2004)
- `RNAshapes` (version 2.1.5) (Steffen et al., 2006)
- `MUSCLE` (version 3.7) (Edgar, 2004)
- `WU-BLAST` (version 2.0) (Gish, W. (1996–2004) `http://blast.wustl.edu`)
- `NCBI BLAST` (version 2.2.14) (Altschul et al., 1990)
- `libsvm` (version 2.89) (Chang and Lin, 2001)
- `WEKA` (version 3) (Hall et al., 2009)
- `CD-HIT-EST-2D` (version 2.0) (Li and Godzik, 2006)
- `ps2eps` (version 1.58)
- `convert` (part of `ImageMagick`, version 6.2.8)
- `Perl` modules
  - `BioPerl` (version 1.5.2) (Stajich et al., 2002)
  - `Algorithm::NaiveBayes` (version 0.04)
  - `RNA::Features` (part of `miROrtho`, version 0.061)
  - `Algorithm::LIBSVM` (part of `miROrtho`, version 0.022)

B.2.2 *More detailed installation notes*

B.2.2.1 *Vienna RNA secondary structure package*

Get the source tarball, unpack, compile and install like this:

```
$ tar xzvf ViennaRNA-x.x.x.tar.gz
$ cd ViennaRNA-x.x.x
$ ./configure
$ make
# su -c 'make install'
```

If you installed the Vienna Package on your system as root, the `Perl` package `RNA.pm` should be found in your Perl `@INC` path. To check this, running the following command should not print an error message:

```
$ perl -e "use RNA;"
```

Also make sure to copy the files in `ViennaRNA-x.x.x/Utils` to your `bin/` directory or add the path to your `PATH` variable in your `~/.bashrc` file like this if your are using the BASH shell.

```
PATH=/path/to/ViennaRNA-x.x.x/Utils:$PATH
```

B.2.2.2 *MUSCLE: A multiple sequence alignment program*

```
$ ./configure
$ make
# su -c 'make install'
```

B.2.2.3 *WU-BLAST*

Rights to BLAST 2.0 (WU-BLAST) have been acquired by Advanced Biocomputing, LLC. If you don't have a ancient version of WU-BLAST, try to modify the program `miRNAblast.pl` to use `NCBI's BLAST`.

Please note that the following executables have to be found in your system's path variable for the script to work correctly: `blastn`, `xdformat`, `xdget`, `dustmasker`.

B.2.2.4 *NCBI BLAST*

Install `NCBI's BLAST` package and move the executables `blastall` and `formatdb` to a directory in your path variable. This BLAST version is needed by `miRNAortho.pl`.

B.2.2.5 *libsvm*

Download the distribution it using the following commands:

```
$ tar xzvf libsvm-2.89.tar.gz
$ cd libsvm-2.89
$ make
$ cp svm-* ~/bin
```

B.2.2.6 *WEKA*

Weka is a collection of machine learning algorithms for data mining tasks. Note this software is not required for the basic `miROrtho` work flow.

```
$ unzip weka-3-6-1.zip
$ mv weka-3-6-1 ~/bin
% add these lines to ~/.bashrc
export WEKAHOME=~/bin/weka-3-6-1
export CLASSPATH=$WEKAHOME/weka.jar:$CLASSPATH
export CLASSPATH=~/bin/libsvm-2.89/java/libsvm.jar:$CLASSPATH
alias weka='java -Xmx8024m weka.gui.Main'
```

B.2.2.7 *CD-HIT-EST*

```
$ tar xzvf cd-hit-xx.tar.gz
$ cd cd-hit-xx
$ make
$ cp cd-hit-xx ~/bin
```

B.2.2.8 *Perl modules*

BIOPERL

Download and install the `BioPerl` package via the source tarball or use your system's package manger to install it. The later solution is favored. On a current Fedora system you can run the following command as root to install the `BioPerl` package:

```
# su -c 'yum install perl-bioperl'
```

RNA::FEATURES

The `Perl` module `RNA::Features` depends on the `RNA.pm` `Perl` module from the Vienna Package and implements some methods on RNA sequences and on folded RNA structures. It is required for the calculation of the features vectors for the classification. The installation follows the classical approach for installing `Perl` modules:

```
$ tar xzvf RNA-Features-x.xx.tar.gz
$ cd RNA-Features-x.xx
% to install without root permissions
$ perl Makefile.PL LIB=~/perl-modules/ \
  INSTALLSCRIPT=~/bin/
% otherwise
$ perl Makefile.PL
$ make
$ make test
# su -c 'make install'
```

ALGORITHM::LIBSVM

The `Perl` module `Algorithm::LIBSVM` implements an interface to the libsvm library. Installation:

```
$ tar xzvf Algorithm-LIBSVM-x.xx.tar.gz
$ cd Algorithm-LIBSVM-x.xx/libsvm-2.89
$ make
$ cp svm-* ~/bin
```

```
$ perl Makefile.PL
$ make
$ make test
# su -c 'make install'
```

THE REMAINING PERL MODULES

They remaining `Perl` modules can be downloaded and installed manually from `cpan.org` or using the cpan shell command. The preferred solution should be to use the system's package manger for installing new Perl modules.

```
% solution 1
$ perl -MCPAN -e shell
$ cpan -i Algorithm::NaiveBayes

% Solution 2
$ cpan -i Algorithm::NaiveBayes

% Solution 3 (Fedora)
# su -c 'yum install perl-Algorithm-NaiveBayes'

% Solution 3 (Ubuntu)
$ apt-cache search perl | grep digest
# sudo apt-get install libAlgorithmNaiveBayes
```

EXPORTING PATHS

After installation of the binaries and the `Perl` programs, add the directory in which you installed them to the `PATH` variable so that the shell can find them. To do so, add these lines to the `~/.bashrc` file if you are using the BASH shell:

```
export PATH=/path/to/programs/:$PATH
export PERL5LIB=/path/to/modules/:$PERL5LIB
```

B.3 TUTORIAL

This is a quick walk-through the pipeline. We will a) scan a sequence for hairpin-like structures (`stemloop-scan`) b) get numerical vectors of features for each sequence (`miRNAclassify`) c) apply a classifier to predict putative miRNA genes (`miRNAclassify`) and make a consensus set (`miRNAcluster`) d) search for orthologous precursors in closely related species and e) align and classify those orthologous groups as putative miRNA gene families (`miRNAorthoClassify`). All the predictions and family alignments can be visualized using the tool `miRNAvisualize`. We will use sample data files from the `examples/` sub-directory. The hereby described method was used to produce the data shown in the `miROrtho` database (Gerlach et al., 2009).

B.3.1 *Programs in the miROrtho package*

The core of the package consists of several `Perl` programs (Table 18 on the following page) which interact and produce results which can be extracted individually or used as an input for one of following programs. The basic data structure of the input files is the FASTA format and simple text files.

Table 18: Programs in the `miROrtho` package

miRNAblast	Performs a modified and constrained BLAST search for finding homologous miRNA genes
miRNAclassify	Computes numeric feature vectors for RNA sequences and classifies them using an SVM model
miRNAcluster	Clusters overlapping predictions, creating a non-redundant set of output sequences
miRNAortho	Groups predictions into orthologous groups
miRNAorthoClassify	Computes feature vectors for RNA alignments and classify them as miRNA or non-miRNA group
miRNAvisualize	Produces color-coded alignment figures for groups showing the conservation pattern and consistent and compensatory mutations
splitter-fasta	Splits a long genomic sequence into smaller overlapping junks
stemloop-scan	Scans a sequence for stem-loop-like structure resembling miRNA stem-loops
fasta-dustmasker.pl	Mask low-complexity regions in fasta sequences

Note that some of the programs use temporary files which are by default saved in `/tmp` on your file system. For using a different path for the temporary files, export the `TMP_PATH` variable in your `~/.bashrc` configuration file.

B.3.2 *File formats*

All programs use FASTA files as input or output. The SVM models can be found in the `models/` directory. They are simple text files.

B.3.3 *Files used in the tutorial*

The directory `examples/` in the `miROrtho` package contains several files which can be used for testing the package. The following files can be found:

MIRNA_SEQ_SAMPLE.FA A sample file containing a couple of actual miRNA sequences

MIRNA_ALN_SAMPLE.ALN A sample alignment file containing the mir-122 family alignment

Try to run `miRNAclassify.pl` on the FASTA sequences like this:

```
$ miRNAclassify.pl --type svm --classify \
  models/seq_model --scale models/seq_model.range \
   examples/mirna_seq_sample.fa

dme-mir-1 1 0.93912
dme-mir-3 1 0.971485
dme-mir-5 1 0.938752
dme-mir-4 1 0.652527
```

To classify an alignment you can run this command:

```
$ miRNAorthoClassify.pl --classify models/aln_model \
   --scale models/aln_model.range examples/

mirna_aln_sample.aln 1 0.93877
```

B.3.4 *splitter-fasta: Split a sequence into overlapping segments*

Theoretically you could directly use one big multiple fasta sequence file per genome and run the pipeline on it. However due to constrains in the implementation of the `RNALfold` (used by `stemloop-scan`) program and memory requirements it is probably a good idea to split the input sequence into smaller overlapping fragments. Use the following command to split large genomic sequences into segments which can than be easily distributed on a cluster. Be cautious with some genomes. The *Schistosoma japonicum* genome for example was in DOS-format and had to be converted to UNIX format using `dos2unix` before being scanned.

```
$ splitter-fasta.pl --size 5000000 --overlap 200 \
  --replace-ids Hsap_DNA --print-replaced-ids hsap.txt \
  --basename hsap --max-bp 50000000 genome.fa
```

This command basically does the following: If a single fasta sequence in the input file is larger than 5 Mb, split it into 5 Mb fragments with 200 bp overlap. A new identifier is created containing a four letter species abbreviation. The final identifiers which will be saved to a file called hsap.txt together with the original ids the file will have a format like this `Hsap_DNA_1, Hsap_DNA_2, etc` All the fragments are ordered by size and saved in individual fasta files named hsap1.fas, hsap2.fas, etc. not exceeding an upper limit of 50 Mb of total DNA sequence per file. The individual files can now be scanned for stable stem-loops in the next step. As miRNA genes are not supposed to exceed 200 bp in length, and overlap of 200 guarantees to not miss any miRNA at the borders of two segments. The required run time is a few minutes per genome. You will need a little bit more memory than the size of the input genome itself.

B.3.5 *stemloop-scan: Scan for miRNA-like stem-loop structures*

Run the following command on all the files from the previous section to scan for stem-loop structures which resemble the ones from already known miRNAs. The default parameters were estimated from all metazoan miRNAs stored in `miRBase 13.0`. The parameters were adapted to capture more than 95% of known `miRBase 13.0` miRNAs in this first step of the pipeline.

```
$ stemloop-scan.pl --min-length 50 --max-length 130 \
  --max-energy -13 --max-hairpin-size-multi 8 \
  --max-bulge-loop 5 --max-interior-loop 8 \
  --min-pairs 15 --max-pairs 70 --max-hairpins 2 \
  --max-continous-stack 30 --max-multi-loop 15 \
  --max-hairpin-loop 25 --min-ratio 0.35 --max-ratio 2.3 \
  --T 37 hsap1.fas > hsap1.fas.stemloops
```

The required run time for this program is about 6 hours on one CPU, using roughly 500 MB of memory. You can use larger single input sequences which also increases the run time and memory requirements. However, an upper limit of about 60 Mb per single sequence is set by constrains of the `RNALfold` implementation.

B.3.6 *miRNAclassify: Compute feature vectors and classify input sequences*

Get the features vectors for new data:

```
$ miRNAclassify.pl --features tab --header \
  hsap1.fas.stemloops > hsap1.fas.stemloops.features
```

Use the precalculated feature file and an SVM model to assign class labels to the input sequences. All classifications are stored in a csv file. Sequences with a miRNA class probability > 0.5 are written in a new fasta file. The model file was created using *svm-train* the range file (for scaling the features within 0, 1) was computed via *svm-scale*. Details on how the model was created and which training data was used can be found in Section B.4.1 on page 281.

```
$ miRNAclassify.pl --classify training.dat.scale.model \
  --scale training.dat.range  --type svm --feature \
```

```
hsap1.fas.stemloops.features hsap1.fas.stemloops \
--print-fasta --csv hsap1.fas.stemloops.svm.csv > \
hsap1.fas.stemloops.svm.fa
```

For smaller input files the feature and the SVM score calculation step can also be combined. Note however, the huge memory requirements. Using the *svm-score* option one can specify a minimum SVM cutoff score as a threshold.

B.3.7 *fasta-dustmasker: Remove low-complexity sequences*

As the SVM did not rigorously penalize low-complexity sequences enough (neither the positive nor the negative training set did contain many of those sequences) some of the prediction can have a low-complexity regions. To remove those sequences from the dataset, a low-complexity filter tool named `fasta-dustmasker` which is based on NCBI's `dustmasker` can be used. In this example we filter out all results containing more than 5% of filtered (low-complexity) regions.

```
$ fasta-dustmasker.pl --mask 5 \
  --fasta hsap1.fas.stemloops.svm.fa \
  > hsap1.fas.stemloops.svm.filtered.fa
```

B.3.8 *miRNAblast: Search for homologous miRNA in genomic sequences*

As the SVM-based procedure of the pipeline might miss some known candidates, one can use an extended BLAST search to "fill in the gaps". Run the following command to search for homologous miRNA sequences in your genomic input sequences.

```
$ miRNAblast.pl --E 10 --mfe -15 --shapes 0 --randfold 0.05 \
  --randomizations 100 --aln 20 --seed_cons 100 \
  --min_cons_mature 90 --min_cons_precursor 50 --cpus 4 \
  --query mirna.fa hsap*.fas > output.txt
```

The output is a tab separated file containing the following fields: hairpin_name, query_id, subject_id, e_value, N, Sprime, S, alignlen, nident, npos, nmism, pcident, pcpos, qgaps, qgaplen, sgaps, sgaplen, qframe, qstart, qend, sframe, sstart, send, uniq, seq, struc, extract_start, extract_end, cons_mature, shapes, mfe, randf, single_start, single_end.

To convert this output to a simple fasta file you could use the following command. Note that using homologous queries (e.g. let-7a, let-7b) the same locus can be found several times. The final output of `miRNAblast` has therefore to be purged from duplicates in a later step.

```
$ for x in *blast
  do awk \
  '{print ">"$3":"$18":"$27":"$28" BLAST:"$4" "$2" "$33"\
  -"$34"\n"$25}' $x > $x.fa
  done
```

Using all 10,122 metazoan miRNAs from `miRBase 14` the program has a run time of about three to eight hours on a 50 Mb fasta file used

as the database. Note the difference in run time arises from the number of sequences within the 50 Mb file and the complexity of the target sequences. The more database hits are found (e. g., low-complexity miRNAs like mir-466) the more post-BLAST checks have to be performed within the program to filter out the false positive hits.

B.3.9 *miRNAcluster: Create a non-redundant set of BLAST hits and SVM predictions*

This program can cluster the predicted putative miRNA sequences into a non-redundant set. It takes the hit with the lowest e-value as a representative sequence for a certain locus. For loci which are covered by an SVM prediction, it uses the prediction with the highest SVM score to represent the cluster. If an SVM prediction overlaps a BLAST hit, the BLAST sequences with the lowest e-value is used as the final output sequence for this region.

The input file format is just simple fasta with a header similar to this one:

```
>Hsap_2:1:10:110 SVM:0.990796 more_descriptive_text
# chr:strand:start:end svm_score description

>Hsap_2:-1:10:110 BLAST:7.4e-06 more_descriptive_text
# chr:strand:start:end blast_e_value description
```

As soon as the keyword *SVM* or *BLAST* appears in the fasta header the clustering is performed using the highest scoring SVM prediction or the BLAST hit with the lowest e-value as a single representative of any overlapping locus. The tool is generic and can be used to create a set of non-overlapping fasta sequences by simply choosing the longest sequence for each overlapping cluster. Run the tool like this using the internal identifiers as e. g., the human chr1 has been splitted several times and we must assure that two prediction lie on the same chromosome when clustering them.

```
$ miRNAcluster.pl --chr hsap.txt --ids int \
  *blast*.fa *.svm*.fa > hsap_mirna_predictions.fa
```

B.3.10 *miRNAortho: Group miRNAs in putative orthologous groups*

As soon as you have a set of putative miRNA which are non-redundant coming from the SVM or the BLAST procedure, you can group them. To assign orthologous groups a simple method using best-reciprocal BLAST hits coming computed by NCBI BLAST can be used. The program was tested on 115 files (one file per species) containing a total of 638,072 sequences. `miRNAortho` had a runtime of 4.6 days on six CPUs.

Run the following command on all your FASTA sequences. Note this might take some time and use a considerable amount of RAM depending on the number of input sequences.

```
$ cat *.fa | perl miRNAortho.pl
```

The different all-against-all BLAST queries span a complete graph which has n vertices (number of species) and $n(n-1)/2$ edges. As with the BLAST searches every vertice has an direction, the number of

BLAST queries to perform is actually $n(n-1)$. So for 115 species one has to compute 13,110 BLAST queries.

Orthologous groups are created with simple increasing numbers such as group1.fa, group2.fa, etc. To align those FASTA sequences you might want to use a program like R-Coffee which combines sequence based alignments with structural information. This is certainly important for the fairly diverse miRNA groups. The command is:

```
$ for x in *.fa; do t_coffee -mode mrcoffee -infile $x; done
```

B.3.11 *miRNAorthoClassify: Assign a miRNA family class probability score to an alignment*

A model built on alignments of miRNA and non-miRNA groups can be used to score new alignment as putative new miRNA families. Like the `miRNAclassify` program, `miRNAorthoClassify` uses an SVM model at its core. This time the model was trained with alignment specific features which should distinguish miRNA groups from those sequences which just align by chance and do not show the characteristics alignment properties known for conserved miRNA families.

The program has a run time of about 166 seconds for every 100 input alignments.

Prior to classification the groups have to be aligned. All alignment on which the model was trained were made using t_coffee with the mrcoffee switch for aligning structural RNAs.

```
$ for x in *.fa; do t_coffee -mode mrcoffee -infile $x; done
```

Run the following command for assigning a miRNA score to all alignments:

```
$ miRNAorthoClassify.pl --classify aln_svm.model \
  --scale aln_svm.range ortho_groups_alignments/ \
  > scores.txt
```

B.3.12 *miRNAvisualize: Draw secondary structure and alignment graphics*

Once the final groups are computed, `miRNAvisualize` allows for various ways to visualize the alignment and consensus secondary structure. All figures are color-coded according to consistent and compensatory base changes (see Fig. 2 on page 7).

Run the following command:

```
$ miRNAvisualize.pl group.fa --mature 1-22 --mature 40-61
```

The following figures are produced: group.ss.png showing the conserved secondary consensus structure and group.aln.png showing the color-coded alignment. (See also Figs. 38 to 39 on the next page)

B.4 DETAILS

B.4.1 *The SVM model*

The positive training data for the SVM model was extracted from a non-redundant set of miRNA genes from `miRBase 13.0` and a previous

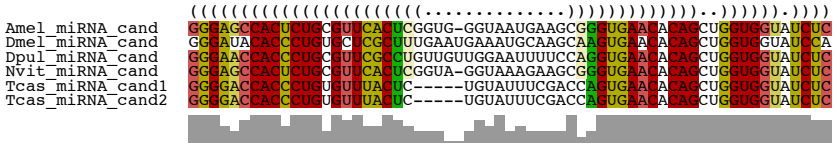


Figure 38: Alignment of a miRNA family showing the sequence conservation pattern at the bottom as gray bars and the secondary structure in dot-parenthesis notation on top of the alignment.

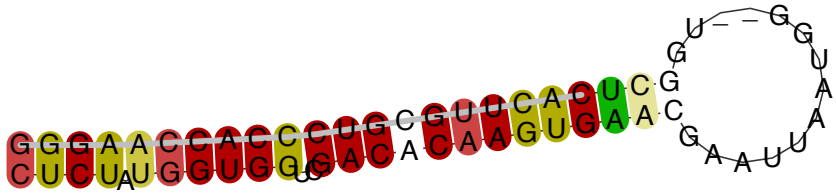


Figure 39: Conserved consensus secondary structure of a miRNA family showing many consistent mutations. The putative mature part is marked by a gray line overlapping the sequence.

run of the `miROrtho` pipeline. Stem-loop-like ncRNA sequences from `Rfam 9.1` and the randomly chosen sequences from the stemloop-scan procedure were used as negative training data. The model was then created using the following commands:

```
$ miRNAclassify.pl --features libsvm --pos pos.fa \
  --neg neg.fa > training.dat
$ svm-scale -l 0 -u 1 -s training.dat.range training.dat \
  > training.dat.scale
$ svm-train -b 1 training.dat.scale training.dat.scale.model
```

For the alignment SVM model the positive training data was taken from alignment of non miRNA groups whereas the negative training data came from other putative miRNAs which were aligned. Although there are certainly some true miRNA groups among this negative training set, the SVM should be flexible enough to handle this minor "noise". The model was created using the following commands:

```
$ miRNAorthoClassify.pl --features libsvm --pos \
  ortho-predictions-new10-pos/ \
  --neg ortho-predictions-new10-neg-400/ > training.dat
$ svm-scale -l 0 -u 1 -s training.dat.range training.dat \
  > training.dat.scale
$ svm-train -b 1 -c 8.0 -g 2.0 training.dat.scale \
  training.dat.scale.model
```

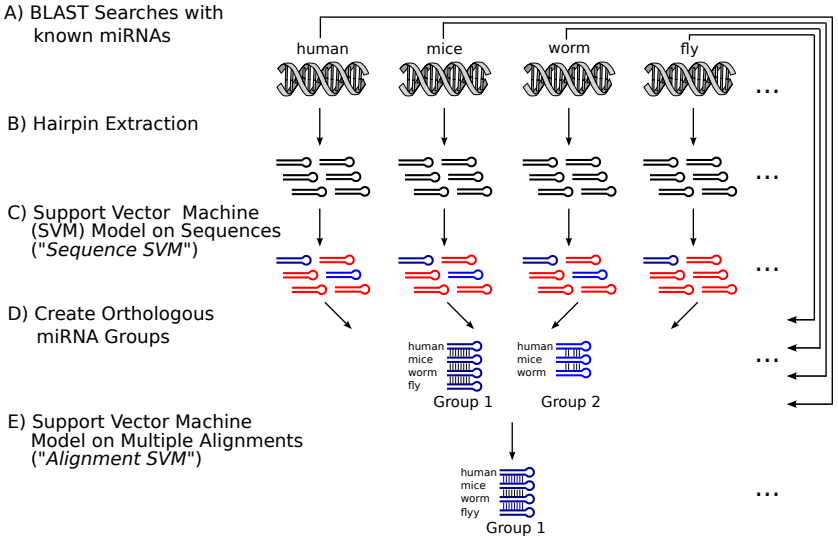


Figure 40: Overview of the `miROrtho` pipeline

B.5 MANUAL PAGES

B.5.1 *Programs*

B.5.1.1 *miRNAblast*

Search for homologous miRNA in a genome database

SYNOPSIS

```
miRNAblast.pl [options] --db genome --query mirna.fa
miRNAblast.pl [options] genome.fa --query mirna.fa

Options:

  --help                brief help message
  --man                 full documentation
  --verbose             more output details (default 0)
  --db                  name of preformatted BLAST database
  --query               a file containing the mirna
                        sequences
  --E                   e-value cutoff (default 10)
  --mfe                 mfe cutoff value (default -15)
  --shapes              RNAshapes cutoff value (default 0)
  --randfold            Randfold cutoff value
                        (default 0.05)
  --randomizations      shuffling steps for Randfold
                        (default 100)
  --aln                 min alignment length (default 20)
  --seed_cons           min conservation of the seed
                        (default 100)
  --min_cons_mature     minimum conservation mature part
                        (default 90)
  --min_cons_precursor  minimum conservation whole
                        precursor (default 50)
  --cpus                number of processors for blast to
                        use (default 1)
  --blast               detailed control over the blast
                        parameters
```

DESCRIPTION

miRNAblast searches for homologous pre-miRNA sequences in a genomic database. First a sensitive BLAST search is performed, and the results are aligned to the query sequence. The following filters are passed on the putative homologous sequence: Alignment length, sequence length, seed conservation, mature conservation, precursor sequence conservation, check for number of gaps in mature region, check for minimum-free folding energy, more than 30% of base pairs are paired, no multi-branching in mature region, RNAshapes, Randfold filter. A valid hit in printed if all of this criteria are met. All parameters can be adapted to account for more stringency or more sensitivity. The default parameters are mainly targeted for sensitivity to the cost of more computational time.

OPTIONS

--help

Print a brief help message and exit

--man

Print the full manual page

--verbose <INT>

Verbose level 0 prints only a tab separated output of the hits found; Verbose level 1 prints more detailed information including the alignments of the query and the subject sequences; Verbose level 2 prints detailed information about the individual BLAST hits and at which step they were discarded from being a new pre-miRNA.

--db <DB-NAME>

Name of a preformatted BLAST database.

--query <STRING>

A file from which the miRNA sequences as read as the query sequence

--e <FLOAT>

The e-value cutoff of the blast search (default 10)

--mfe <FLOAT>

The mfe value cutoff for the folded stem-loops (default -15)

--shapes <FLOAT>

The probability cutoff for folding into a simple stem-loop using RNAshapes. A perfect stem-loop would get a value of 1.0. Range 0.0 - 1.0 (default 0)

--randfold <FLOAT>

Calculates a p-value of a sequence to have a lower mfe than shuffled versions of the same sequence (default 0.05)

--randomizations <INT>

The number of shuffling steps performed by Randfold (default 100)

--aln <INT>

The minimum BLAST alignment length between the query and the subject sequence. (default 20)

--seed_cons <FLOAT>

The minimum required seed conservation in the alignment of the query and the subject sequence (default 100)

--min_cons_mature <FLOAT>

The minimum required conservation of the mature part.(default 90)

--min_cons_precursor <FLOAT>

The minimum required conservation of the whole precursor miRNA. (default 50)

--cpus <INT>

Number of processors for BLAST to use (default 1)

--blast <STRING>

Supply a string containing modified BLAST parameter to modify the default ones (default '-M 1 -N -1 -Q 3 -R 2 -W 7 -T 7 -wordmask seg -hspsepsmax 60 -topcomboN 15')

PREREQUISITES

WU-Blast, MUSCLE, RNAshapes, EMBOSS, Vienna Package, BioPerl, awk, Pod::Usage, Getopt::Long, File::Temp

CONFIGURATION AND ENVIRONMENT

Add the following environment variables to your ~/.bashrc file

```
export TMP_PATH=/tmp
export PATH=~/bin/ViennaRNA-1.8.3/Progs:$PATH
... plus all the other binaries (see prerequisites)
export PERL5LIB=ViennaRNA-1.8.3/Perl/blib/lib:$PERL5LIB
export PERL5LIB=ViennaRNA-1.8.3/Perl/blib/arch:$PERL5LIB
```

CAVEATS

The genomic database file should contain unique ids in the fasta headers. The format of the query mirna.fa file should be as follows:

```
>dme-miR-1 dme-mir-1 56-77
UUCAGCCUUUGAGAGUUCCAUGCUUCCUUGCAUUCAAUAGU ...
```

EXAMPLES

```
# Search for highly similar miRNA in several chromosomes
miRNAblast.pl chr*.fas -q mirnas.fa -E 1e-3 \
--min_cons_precursor 80

# Use a preformatted database
miRNAblast.pl --db /db/database --query seq.fa
```

AUTHOR

Daniel Gerlach

VERSION

```
0.43  Bug fix: Disregarding overlapping loci using start
      corrected fasta sequences did not work correctly
0.42  Add BLAST option -topcomboN 15
      Replace MAFFT by MUSCLE (faster)
      Only run the RNAshapes part if shapes_limit>0
      Replace Randfold filter by own implementation (faster)
0.41  Use BLAST with a preformatted database
0.4   Use environment variables for portability changing
      the default /tmp directory and the modules path
0.31  Bug with tmp file creation fixed
0.3   Modified version which alows a splitted BLAST
      databases with headers like Hsap_DNA_5000-10000 and
      multiple fasta input sequences
0.2   Modified the default parameter to be more sensitive
0.1   Original build
```

COPYRIGHT

This program is free software released under the terms of the BSD license.

B.5.1.2 *miRNAclassify*

Train a model for miRNA classification and use it on unseen data

SYNOPSIS

```
miRNAclassify.pl [options] [sequences.fa]
# See also examples below

Options:

  --help                brief help message
  --man                 full documentation
  --features            get features only (tab, libsvm,
                        weka) OR with --classify: use
                        precalculated feature file
  --print-fasta         also print fasta output when
                        using a precalculated feature
                        file instead of tab only
  --csv                 write the SVM scores to a file,
                        while printing sequence to stdout
                        using the --print-fasta option
  --header              include a header line
                        (only for features=tab)
  --plus-features       define additional features,
                        e.g. z-score
  --desc                dataset description for weka
  --train               filename for the model file
  --pos                 fasta input file, positive
                        training set
  --neg                 fasta input file, negative
                        training set
  --classify            use a previous trained model
                        file for classification
  --type                classifier to use ('svm' or 'nb')
  --scale               a scaling file for the testing
                        data (if type 'svm')
  --svm-score           svm cutoff score (default >= 0.5)
```

DESCRIPTION

miRNAclassify can train a Support Vector Machine (SVM) classifier using two input sets of data: A positive training class (known miRNA stemloop sequences) and a negative training class (non-miRNA stemloop sequences). After a model has been trained, it can be used on a new dataset to classify its member sequences as putative miRNA and non-miRNA sequences assigning them a class probability score.

OPTIONS

--help

Print a brief help message and exit

--man

Print the full manual page

--features <TAB|LIBSVM|WEKA>

Compute only the feature vectors. Prints a tab separated list of features ('tab') or features for the positive / negative class in the libsvm or weka format.

While using the –classify option, the –features option can be used to load a precalculated features file.to increase performance.

--print-fasta

Also print the fasta output for sequences that are classified as miRNAs while using precalculated feature files.

--csv <FILENAME>

Print the SVM scores to a file while the fasta sequence classifier as miRNAs are printed to stdout using the –print-fasta option.

--header

Print a header line for the features (only with –features tab)

--plus-features <Z-SCORE|SCI|SECEIGEN>

Calculate additional features: z-score, sci, secEigen (z-score and p-value, self containment index, second Eigenvalue of the Laplacian matrix). Note that using this additional feature will increase runtime refered to a base level of 0.033 sec per sequence: z-score (18x), sci (147x), secEigen (1.3x). Using all of those features at once will lead to a runtime of 4.93 sec per sequence for the features calculation.

--desc <DATASET-DESCRIPTION>

Description of the dataset. Used for the weka dataset format. Do not use white space or semicolons in this field.

--train <MODEL-FILENAME>

Invoke the training function and save the model file in the given filename.

--pos <FASTA-FILENAME>

Supply a fasta file containing positive training data sequences.

--neg <FASTA-FILENAME>

Supply a fasta file containing negative training data sequences.

--classify <MODEL-FILENAME>

Load the give model file and classify input data using the model file. Note that due to the implementation, scores using Naive Bayes tend to be very close to 0 or 1.

Use the –feature option to supply a precalculated feature file if using –type svm

The libsvm model should be scaled and calculated using the -b 1 option.

--type <SVM|NB>

Specify the classifier to use.

--scale <FILENAME>

Only to use with –type svm. Supply the same scaling file that was used for scaling the training data and on which the model was created.

--svm-score <0.5 - 1>

SVM cutoff score. Minimum score for SVM classification to assign the class "miRNA" to a sequence. The default value is 0.5. Higher values result in more stringent result sets, which however leads to a lower sensitivity. Note this score has no influence on the csv file output in which all raw classification score are written.

PREREQUISITES

Getopt::Long, Algorithm::NaiveBayes, RNA::Features, Pod::Usage, Bio::SeqIO, RNA, libsvm 2.89

CONFIGURATION AND ENVIRONMENT

Add the following environment variables to your ~/.bashrc file

```
export PERL5LIB=~/lib:$PERL5LIB
export PERL5LIB=ViennaRNA-1.8.3/Perl/blib/lib:$PERL5LIB
export PERL5LIB=ViennaRNA-1.8.3/Perl/blib/arch:$PERL5LIB
```

CAVEATS

The model generated by the Naive Bayes classifier is saved in binary format. To read it use the following command:

```
od -ca model_nb | less
```

Nevertheless this is not very informative.

If a model was trained with additional features, do not forget to load them also during the classification step. Otherwise the number of features do not match between the model and the classification step.

Some of the features are hard-coded. If you change the bulge or interior loop in the first stemloop-scan steps change these lines:

```
my %topo = $o->parse_structure($structure, 5, 8)
for my $i ( 1 .. 5 ) {
for my $i ( 2 .. 8 ) {
```

EXAMPLES

```
# Train a model using a Naive Bayes classifier
miRNAclassify.pl --train model_nb --pos test/mirna.fa \
--neg test/non-mirna.fa

# Classify unknown samples
miRNAclassify.pl --classify model_nb type nb test/mirna2.fa

# Generate a WEKA set
miRNAclassify.pl --features weka --pos pos.fa --neg neg.fa \
--desc dataset_1 > weka.arff

# Use libsvm to classify
miRNAclassify.pl --classify data.scale.model \
--scale data.range --type svm seq.fa
```

```
# Calculate a feature file
miRNAclassify.pl --features tab --header seq.fa > \
seq.fa.features

# Use libsvm in conjunction with a precalc feature file and
# print tab output
miRNAclassify.pl --classify data.scale.model \
--scale data.range --type svm --features \
seq.fa.features

# Use libsvm in conjunction with a precalc feature file and
# print fasta
# Save the scores for all predictions in a csv file
# Write only sequence with a SVM score > 0.99
# in the output file
miRNAclassify.pl --classify data.scale.model
--scale data.range --type svm --features \
seq.fa.features --print-fasta --csv scores.csv \
--svm-score 0.99 seq.fa > seq.svm.fa
```

AUTHOR

Daniel Gerlach

VERSION

```
0.41  Minor bug fix
0.40  New feature: entropy wordsize 3 (like in NCBI's
      dustmasker)
0.39  Bugfix --print-header printed one '\t' too much at
      the end of the line
      Suppress features stddev_M and S1 as they are
      dispensable
      Buffix - feature over_09 was always 1
0.38  Add a svm score cutoff parameter
0.37  "Classify using a SVM" extended and bug fix
0.36  Also print the scores for all the other sequences in a
      csv file
0.35  Add FASTA output to the LIBSVM classifier
0.34  Internal changes for feature output generation
      Classifier libsvm works on precalculated feature file
0.33  Add an option for weka dataset description
0.32  More features (size of bulges, stacks, interior loops)
0.31  Add WEKA output format
0.3   Add the libsvm classifier
0.2   Add the Naive Bayes classifier
0.1   Original build
```

COPYRIGHT

This program is free software released under the terms of the BSD license.

B.5.1.3 *fasta-dustmaker*

Mask low-complexity regions in fasta sequences

SYNOPSIS

```
fasta-dustmaker [options] file.fa
```

```
Options:
  --help                brief help message
  --man                 full documentation
  --mask                max. % of masked nt in input
                        sequence
  --reverse             print out the sequence which are
                        normally filtered
  --fasta               write a new fasta file ignoring
                        sequences
                        with more low-complexity regions
                        than x% (--masked)
  --tab                 tab separated output of %masked nt
                        per sequence
```

DESCRIPTION

fasta-dustmaker reads a FASTA file and removes entries with low-complexity regions defined by a threshold value.

OPTIONS

--help

Print a brief help message and exit

--man

Print the full manual page

--mask <0-100>

Maximum percentage of masked nt in output sequence (default 0).

--reverse

Print out the filtered sequences which did not pass the –mask criteria. So low complexity sequencs are printed to stdout.

--fasta

Output a fasta sequence with low-complexity entries removed.

--tab

Write a tab separated output with sequence ids and % masked nucleotides.

EXAMPLES

```
# Remove entries having more than 5% of their nt's masked
fasta-dustmasker.pl --fasta --mask 5 file.fa > file_small.fa
```

PREQUISITES

bioperl, NCBI's dustmasker

AUTHOR

Daniel Gerlach

VERSION

```
0.1   Original build
```

COPYRIGHT

B.5.1.4 *miRNAcluster*

Create a set of non-redundant fasta sequences from overlapping input sequences

SYNOPSIS

```
miRNAcluster.pl [options] file.fa [file.fa ...]

Options:
  --help                brief help message
  --man                 full documentation
  --chr                 contains a map of internal to
                        external ids
  --ids                 print internal or external ids
                        ('ext' or 'int')
```

DESCRIPTION

miRNAcluster clusters a fasta file of sequences and only returns non-overlapping sequences. For overlapping blocks the BLAST sequence with the lowest e-value is taken, for SVM sequences the one with the highest SVM score is taken. If a BLAST sequence overlaps an SVM predicted sequence, the BLAST one is printed out.

```
Input format:
>Hsap_2:1:10:110 SVM:0.990796 more_descriptive_text
>Hsap_2:-1:10:110 BLAST:7.4e-06 more_descriptive_text
chr:strand:start:end SVM_score|BLAST_e_value description
```

If there are no keywords like 'BLAST' or 'FASTA' in the header simply the longest sequence will represent a cluster of overlapping sequences.

OPTIONS

--help

Print a brief help message and exit

--man

Print the full manual page

--chr <ID-MAP-FILE>

A file containing three fields - an internal id, an external id, and a description field

--ids <INT|EXT>

Print internal or external ids from the id-map-file

EXAMPLES

```
# Create a set of non-overlapping sequences
miRNAcluster --chr ids.txt --ids int *.fa
```

PERQUISITES

bioperl, File::Temp, Number::Interval

AUTHOR
Daniel Gerlach

VERSION
```
0.11  Small bugfix for internal ids
0.1   Original build
```

COPYRIGHT
This program is free software released under the terms of the BSD license.

B.5.1.5 *miRNAorthoClassify*

Calculate features on an alignment and predict it as miRNA or non-miRNA family

SYNOPSIS
```
miRNAorthoClassify.pl [classify|features] [options] dir/

Options:

  --help                brief help message
  --man                 full documentation
  --desc                dataset description for weka
  --features            get features only (tab, libsvm,
                        weka)
  --pos                 directory with clustal alignments,
                        positive training set
  --neg                 directory with clustal alignments,
                        negative training set
  --classify            use a previous trained model file
                        for classification
  --scale               a scaling file for the testing data
```

DESCRIPTION
miRNAorthoClassify uses a Support Vector Machine to classify an alignment of RNA sequences as a miRNA family or a non-miRNA family alignment. It also allows to individually compute features for a sample of alignments to train the SVM model.

Acknowledgement: Some of the ideas used in this program follow the paper Hertel J, Stadler PF. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics. 2006 Jul 15;22(14):e197-202.

OPTIONS
--help

Print a brief help message and exits

--man

Prints the manual page and exits

--desc

A short data set description. Only used for 'WEKA'. Do not use white space or any strange characters.

--features<TAB|LIBSVM|WEKA>

Get features for a set of alignments in TAB, LIBSVM, or WEKA format.

--pos<DIR>

Directory containing positive training sample alignments of real miRNA groups (Only with features libsvm|weka)

--neg<DIR>

Directory containing negative training sample alignments of non-miRNA groups (Only with features libsvm|weka)

--classify<SVM MODEL FILE>

A precalculated SVM model file

--scale<SVM MODEL SCALING FILE>

A scaling file that was used for scaling the training data and on which the model was created

PREREQUISITES

Bioperl, Vienna Package, libsvm 2.89, Algorithm::LIBSVM

CONFIGURATION AND ENVIRONMENT

Add the following environment variables to your ~/.bashrc file

```
# general for user modules
export PERL5LIB=~/lib:$PERL5LIB
# for the RNA.pm module
export PERL5LIB=ViennaRNA-1.8.3/Perl/blib/lib:$PERL5LIB
export PERL5LIB=ViennaRNA-1.8.3/Perl/blib/arch:$PERL5LIB
```

Add the following applications to your PATH variable: RNAfold, RNAalifold, refold.pl

EXAMPLES

```
# Get features for alignments
miRNAorthoClassify.pl --features libsvm --pos dir1/ \
--neg dir2/

# Classify alignments
miRNAorthoClassify.pl --classify model_file --scale \
scale_file dir_to_aln/
```

CEVEATS

Note that the first 4-letters of the FASTA header should contain the species code. This is important for the calculation of the branch-length-score.

Also note that the alignment should be in clustalW format with upper case letters. This is actually required by RNAalifold which fails on lower case alignment letters.

AUTHOR

Daniel Gerlach

VERSION

```
0.21  Bug fix for species names
0.2   Implement classification step
0.12  Bug fixes handling directories containing more than
      1024 alignment files
0.11  Bug fixes and design improvements
0.1   Original build, ported from miRNAalign.pl
```

COPYRIGHT

This program is free software released under the terms of the BSD license.

B.5.1.6 *miRNAortho*

Group predictions into orthologous groups

SYNOPSIS

```
cat *.fa | miRNAorthoy.pl [options]

Options:

  --help                brief help message
  --man                 full documentation
  --log                 use existing ultimate.LOG file
```

DESCRIPTION

miRNAortho uses proteinortho_v2.1.pl to group sequences into orthologous groups using NCBI BLAST. All groups are written into individual FASTA files - one file per group. Use any alignment algorithm to align the groups. E.g.

```
for x in *.fa; do t_coffee -mode mrcoffee -infile $x; done
```

OPTIONS

--help

Print a brief help message and exits

--man

Prints the manual page and exits

--log

Use an existing ultimate.LOG (from protein_ortho_v2.1) file from a previous run

PREREQUISITES

Bioperl, proteinortho_v2.1.pl, NCBI BLAST

EXAMPLES

```
cat *.fa | miRNAortho.pl
```

CAVEATS

All sequences are appended to their groups and groups get a simple running number. Note that if you run miRNAortho a second time the groups just get appended to the existing ones. So in case you already run miRNAortho in the same directory, delete any group*.fa files before rerunning it.

AUTHOR
Daniel Gerlach

VERSION
```
0.2   Bug fix, uses now ultimate.LOG to parse,
      uses precalculated *.bla and *.fa files
0.1   Original build
```

COPYRIGHT
This program is free software released under the terms of the BSD license.

B.5.1.7 *miRNAvisualize*

Align gene families and compute color-coded alignments and consensus secondary structures

SYNOPSIS
```
miRNAvisualize.pl [options] file.fa

Options:
  --help                brief help message
  --man                 full documentation
  --mature              mark mature positions
  --png                 also create *.png versions
  --svg                 also create *.svg versions
  --aln                 compute alignments on the
                        input files
```

DESCRIPTION
miRNAvisualize aligns a gene family and calculates a consensus structure. Graphic files are produced showing the conserved secondary structure, the alignment, and the base pair probability dot-plot. All figures are color-coded according to consistent / compensatory base changes in the alignment in relation to the consensus structure.

The following output files are generated:

```
file.aln              the alignment file
file.aln.(eps|png)    color-coded alignment figure
file.ss.(eps|png)     color-coded consensus secondary
                      structure figure
file.dot.(eps|png)    color-coded-dot plot of pair
                      probabilities
file.alifold          raw RNAalifold output
```

OPTIONS

--help

Print a brief help message and exit

--man

Print the full manual page

--mature <INT-INT>

Mark mature positions with a light gray line. Supply a range including the start and end position like '–mature 1-22' to mark

the first 22 base pairs in the secondary structure. Several mature region can be marked by repeatedly applying the option for different regions.

--png

Generate additionally *.png versions for all graphics files.

--svg

Generate additionally *.svg versions for all graphics files. NOT YET IMPLEMENTED!

--aln

Compute alignments if unaligned FASTA sequences are used as input.

EXAMPLES

```
# Create a color-coded secondary structure figure and
# alignment file with annotated mature regions
# Also create *.png versions of the *.eps graphics
miRNAvisualize.pl bantam.fa --mature 2-20 --mature 40-61 \
--png
```

PERQUISITES

bioperl, Vienna Package (including its scripts), ps2eps, convert (ImageMagick)

AUTHOR

Daniel Gerlach

VERSION

```
0.2   Bug fix wrong eps bounding box
      Add 'aln' option
0.11  Added the png support
0.1   Original build
```

COPYRIGHT

This program is free software released under the terms of the BSD license.

B.5.1.8 *splitter-fasta*

Splits large FASTA files into overlapping fragments

SYNOPSIS

```
splitter-fasta.pl [options] file.fa

Options:
  --help                brief help message
  --man                 full documentation
  --size                final fragment size (default 10 Kb)
  --overlap             overlapping nucleotides
                        (default 0 bp)
  --replace-ids         replace ids with a text and a
                        running number
  --print-replaced-ids  print a file with replaced and
```

```
                        original ids
   --basename           basename for the output files
   --max-bp             maximum number of bp per output
                        file (default 50 Mb)
```

DESCRIPTION

splitter-fasta reads a fasta file and splits it into overlapping fragments

OPTIONS

--help

Print a brief help message and exit

--man

Print the full manual page

--size <INT>

Size of an individual fragment (default 10 Kb)

--overlap <INT>

Size of overlapping region (default 0 bp)

--replace-ids <STRING>

Replace ids with a custom string plus a running number. E.g. 'Hsap_DNA' results in 'Hsap_DNA_1', 'Hsap_DNA_2', etc. Do not use any white space characters in your custom id string.

--print-replaced-ids <FILENAME>

Print a file with the new ids, the original ids and the original description text of the fasta header separated by *tab* characters

--basename <STRING>

Uses the specified string as the basename for the output filenames. If given all splitted sequences are printed to files e.g seq1.fas, seq2.fas and no standard output is printed.

--max-bp <INT>

Restrict the number of nucleotides per output file to N bp. (default 50 Mb)

PREREQUISITES

Getopt::Long, Pod::Usage

CAVEATS

Without any options splitter-fasta.pl behaves similar like the EMBOSS tool splitter.

The program dies if any non unique identifiers are found in the sequence headers. Any gap characters are replaced by N's and a warnings message is shown. In case anything else than a gap or a nucleotide character if found in the sequence, the program dies with a warning message.

EXAMPLES

```
# Split a file into 100 Kb fragments with 100 bp overlap,
# replace the ids, and write the new and old ids to a text
# file:
splitter-fasta.pl -s 100000 -o 100 -r ID -p \
new-old-ids.txt sequences.fa
```

AUTHOR

Daniel Gerlach, with parts from O'Reilly's BLAST book (segment subroutine)

VERSION

```
0.2   Check for unique fasta identifiers;
      Replace gap characters by N's and issue warning
0.11  Bug in --size and --max-bp options fixed
0.1   Original version
```

COPYRIGHT

This program is free software released under the terms of the BSD license.

B.5.1.9 *stemloop-scan*

Scan Genomic Sequences for Putative pre-miRNAs

SYNOPSIS

```
stemloop-scan.pl [options] file.fa

Options:
  --help                   brief help message
  --man                    full documentation
  --top                    search only the top strand
  --bottom                 search only the bottom strand
  --min-length             min length (default 50)
  --max-length             max length (default 130)
  --max-energy             max mfe (default -13 kcal/mol)
  --min-pairs              min number of pairs in the
                           "longest run" (default 20)
  --max-pairs              max number of pairs in the
                           "longest run" (default 70)
  --max-hairpins           max number of hairpin loops
                           (default 2)
  --max-hairpin-size-multi max size of hairpins if two
                           or more hairpins (default 5)
  --max-interior-loop      max size of interior loops
                           (default 7)
  --max-bulge-loop         max size of bulge loops
                           (default 3)
  --max-continous-stack    max size of a continous stack
                           in the stem (default 13)
  --max-multi-loop         max size of a multi-loop
                           (default 15)
  --max-hairpin-loop       max size of a hairpin loop
                           if single (default 16)
  --min-ratio              min ratio paired bp / unpaired bp
```

```
                          (default 0.35)
   --max-ratio            max ratio paired bp / unpaired bp
                          (default 2.3)
   --T                    adjust folding temperature
                          (default 37)
```

DESCRIPTION

stemloop-scan reads a fasta file and prints out regions of putative pre-miRNA precursors according to some structural constrains

OPTIONS

Help

--help

Print a brief help message and exit

--man

Print the full manual page

General Parameters

--top

Scan only the top strand of the input sequence

--bottom

Scan only the bottom strand of the input sequence

--t <FLOAT>

Rescale the folding energy parameters according to the specified temperature. (default 37)

Structure / Energy Restrictions

--min-length <INT>

The minimal length of a hit (default 50)

--max-length <INT>

The maximum length of a hit (default 130)

--max-energy <FLOAT>

The maximum minimum free folding energy (mfe) of a hit (default -13)

--min-pairs <INT>

The minimal number of pairs in the "longest run" of a hit. (default 20)

--max-pairs <INT>

The maximal number of pairs in the "longest run" of a hit. (default 70)

--max-hairpins <INT>

The maximal number of hairpin loops in a hit. Choose '1' if you only want to output simple stem-loops as hits. (default 2)

--max-hairpin-size-multi <INT>

The maximal size of a hairpin if the structure contains two or more hairpins (default 5)

--max-interior-loop <INT>

The maximal size of interior loops within the structure (default 7)

--max-bulge-loop <INT>

The maximal size of bulge loops within the structure (default 3)

--max-continous-stack <INT>

The maximal size of a continous uninterrupted stack in the stem region (default 13)

--max-multi-loop <INT>

The maximal size of a multi-loop region (default 15)

--max-hairpin-loop <INT>

The maximal size of a hairpin loop in a simple stem-loop structure (not multi-branching) (default 16)

--min-ratio <FLOAT>

The minimum value for the paired / unpaired bp ratio in the structure (default 0.35)

--max-ratio <FLOAT>

The maximum value for the paired / unpaired bp ratio in the structure (default 2.3)

PREREQUISITES

Vienna Package, Pod::Usage, Getop::Long, File::Temp

CONFIGURATION AND ENVIRONMENT

Add the following environment variables to your ~/.bashrc file

```
export TMP_PATH=/tmp
export PATH=~/bin/ViennaRNA-1.8.3/Progs:$PATH
export PERL5LIB=~/lib:$PERL5LIB
export PERL5LIB=ViennaRNA-1.8.3/Perl/blib/lib:$PERL5LIB
export PERL5LIB=ViennaRNA-1.8.3/Perl/blib/arch:$PERL5LIB
```

CAVEATS

Due to the implementation of RNALfold and memory issues there is an upper limit for the length of a single fasta sequence to scan. If you try to parse very large DNA sequences like e.g. the human chromosome 1, try to split your input sequences in junks of around 50 Mb (using e.g. EMBOSS' splitter tool).

The directory for temporary files is defined as TMP_PATH which contains the one line fasta sequences. If the program terminates correctly the temporary files are automatically deleted.

EXAMPLES

```
# Generate very stable stemloops at high temperatures:
stemloop-scan.pl --max-energy -30 --max-hairpins 1 \
-T 40 sequences.fa
```

AUTHOR
Daniel Gerlach

VERSION
```
0.61  Second bug fix for RNALfold 1.83
0.6   Use environment variables for portability changing
      the default /tmp directory and the modules path
0.53  Bug fix for RNALfold 1.83 giving wrong positions
      in the id line (1 nt shifted) for substructures
      which stick to the 3' end of the input sequence
0.52  Fixed a bug with the temporary file creation
0.51  Suppress printing of mfe in fasta output file
0.5   Save RNALfold first in a file and then parse this
      file. Solves some memory issues and CPU load
      problems on large files
0.4   Change default parameter, derived from Metazoa
      miRBase 13.0 miRNAs
0.3   Change the design for memory efficiency
0.2   Modified the regexp for parsing the RNALfold output
0.1   Original build
```

COPYRIGHT

B.5.2 *Modules*

B.5.2.1 *RNA::Features*

Calculate some features on RNA sequences

VERSION
Version 0.061

SYNOPSIS
```
use RNA::Features;
my $seq = "CGCAGGGAUACCCGCG";
my $o = RNA::Features->new();
my ($structure, $mfe, $pair_table_ref) = $o->fold($seq);
my @table = @$pair_table_ref;
```

DESCRIPTION
Provides some basic functionality to calculate features on RNA sequences and characteristics of folded RNA sequences

PUBLIC METHODS
new Creates a new RNA::Features object and initializes some default fold parameter like RNA::dangles=2, and RNA::noLonelyPairs=1.

```
Example:
my $o = RNA::Features->new();
```

fold Compute the structure, the minimum free folding energy, and a reference to an array of base pair positions in the mfe. The array of base pair positions is symmetric, so $i=5, $j=10 eq $i=10, $j=5.

```
Example:
my ($structure, $mfe, $pair_table) = $o->fold($seq);
# length of the structure
print $pair_table->[0];
# get the base j paired with i=1, if unpaired j is 0
my $j = $pair_table->[1]
# dereference
my @table = @$pair_table;
```

monomers Compute monomer frequencies, GC, and AU content

```
Example:
my %monomers = $o->monomers($seq);
# get GC content and frequency of U
print "$monomers{'GC_content'}, $monomers{'U'}";
```

dimers Compute dimer frequencies

```
Example:
my %dimers = $o->dimers($seq);
print $dimers{'AA'};
```

rnaAnalysis Returns a hash with the following values:

```
* number of base pairs in mfe 'bp'
* minimum free folding energy (mfe) 'mfe'
* free energy of ensemble 'energy'
* centroid fold energy 'centroid_en'
* mean bp distance of the centroid structure to the mfe
  structure 'centroid_bp_dist'
* frequency of mfe structure in ensemble 'mfe_in_ensemble'
* sum-of-entropy (Shannon entropy) 'Q'
* mean base pair distance within the ensemble 'D'
* mean bp probability in the mfe structure
  'mean_bp_prob_mfe'
* fraction of bp in the mfe having a probability
  over 0.9 'over_09'
* a reference to an array containing the max. mean base
  pair probability for a 18 base pairs window, the mean base
  pair probability of the surrounding region, and the start
  and end base i for the window with the highest mean base
  pair probability 'window_bpp'
* a reference to an array of array containing base pair
  probabilities 'bpp'
* a reference to a hash containing the number of A-U
  G-C, and G-U pairs 'types_base_pairs'

Example:
my %hash = $o->rnaAnalysis($seq);
# get the mfe
print $hash{'mfe'};
# get the base pair probabilities
my $bpp = $hash{'bpp'};
foreach my $i (1..(length($seq)-1)) {
    foreach my $j ($i+1..length($seq)) {
```

```
        print "$i $j $bpp->[$i][$j]\n" if ($bpp->[$i][$j]);
    }
}
# get the highest mean probability for a 18 base pairs
# window
print $hash{'window_bpp'}->[0] \
 if ($hash{'window_bpp'}->[0]);
# get number of GC pairs
$hash{'types_base_pairs'}->{GC}
```

dinuc_shuffle Shuffles a sequence keeping the dinucleotide frequencies stable. Based on code from the Altschul-Erikson dinucleotide shuffle algorithm implemented in perl by Stanley NG Kwang Loong.

```
Example:
my $shuffled_seq = $o->dinuc_shuffle($seq);
```

z_score Calculates a z-score for the minimum free folding energy of a sequences related to dinucleotide shuffled versions of the same sequence. Tests if a sequence has a lower mfe than its shuffled versions. Also reports a p-value - the probability of finding a shuffled sequence which has a lower mfe than that one of the original sequence. 100 shuffling steps should be a good trade off between speed and statistic significantly results.

```
Example:
my ($z, $p) = $o->z_score($seq, 100);
```

mean Calculates the mean value for an array of numbers

```
Example:
my $mean = $o->mean(4, 4, 5, 3);
```

stddev Calculates the standard deviation for an array of numbers

```
Example:
my $stddev = $o->stddev(4, 4, 3, 2);
```

max Returns the maximum value in an array of numbers

```
Example:
my $max = $o->max(10,2,3,44,2,5);
```

min Returns the minimum value in an array of numbers

```
Example:
my $min = $o->min(10,2,3,44,2,5);
```

median Returns the median value from an array of numbers

```
Example:
my $median = $o->median(10,2,3,22,2,5);
```

amfe Adjusted minimum free folding energy (mfe / (length(seq) x 100))

```
Example:
my $amfe = $o->amfe($seq, $mfe);
```

mfei The minimum free folding energy index (amfe / gc_content)

```
Example:
my $mfei = $o->mfei($seq, $mfe);
```

sc_index Calculates the self containment index described by Lee & Kim, 2008. It measures the robustness of RNA structures to changes in the surrounding sequence context, which we hypothesize to be a hallmark of structural modularity. SC values range from 0.0 (no self containment) to 1.0 (completely self contained).

```
Example:
my $sci = $o->sc_index($seq, "/path/to/bin/selfcontain.py");
```

entropy Calculate the sequence entropy using a given wordsize. Using a wordsize of 1, the maximal information content (related to entropy) you get is 2 bit.

```
Example:
my $a = $o->entropy($seq, 1);
```

parse_structure Calculates different statistics on folded RNA sequence. Takes a dot-bracket formated sequence structure, a max bulge loop size, and a max interior loop size as arguments

```
* Hairpin loop sizes 'H' (array reference)
* Bulge loop sizes 'B' (array reference)
* Interior loop sizes 'I' (array reference)
* Multiloop loop sizes 'M' (array reference)
* Stack sizes 'S' (array reference)
* Number of hairpins 'hairpins' (scalar)
* Stacks in "longest run" 'stacks_longest' (scalar)
* Paired / Unpaired bp 'ratio' (scalar)
```

```
Example:
# a structure in dot-bracket format ((..)), max bulge size
# max interior loop size
my %hash = $o->parse_structure($structure, 5, 8);
print $hash{'ratio'},"\n";
my @bulges = @{$hash{'B'}};
my $max_stack = max(@{$hash{'S'}});
```

rnaspectral Computes the second (Fiedler) eigenvalue dF. An RNA structure is represented as a tree-graph G, where vertices represents loops, and edges represent stems. The tree graph can be mathematical represented by a Laplacian matrix L(G). The second eigenvalue of L(G) measures the compactness of the tree-graph.

```
Example:
my $secEigen = $o->rnaspectral($seq, $structure,
    "path/to/bin/rnaspectral");
```

neu Use EvoRSR to calculate the neutrality (neu) of an RNA sequence based on all 3 x length(sequence) one-mutants. The base-pair distance between all one-mutants (Hamming distance 1) and the original sequence is calculated. Thus <n> represents the average fraction of the structure that remains unchanged after a mutation occurs. Instead of the mfe dG the ensemble free-energy and the distance between two sets of ensemble structures is used.

Note: For one 100 nt sequence it takes around 2 sec for the mfe, and 9 sec for the ensemble option to calculate the sequence neutrality. The maximal sequence length is 500 nt.

```
Example:
my $neu = $o->neu($seq, 'ensemble', 'path/to/bin/neu');
```

rnaforester Use RNAforester to calculate a the relative similarity between two sequences based on their sequence and secondary structure. The RIBOSUM85-60 scoring matrix will be used. The score maximum is '1' - meaning the sequences and structures are identical.

```
Example:
my $forester = $o->rnaforester($seq1, $seq2,
    'path/to/bin/RNAforester');
```

INTERNAL METHODS

Internal methods are usually preceded with a "_".

_throw Throws an exception and prints an error message.

```
Example:
$o->_throw(Error in sequence: $seq) unless \
 ($o->_validate_seq($seq));
```

_warn Prints a simple warnings message.

```
Example:
$o->_warn(Warn about s.th.: $seq) unless \
 ($o->_validate_seq($seq));
```

_validate_seq Check if input is a valid sequence

```
Example:
$o->_validate_seq($seq) || die;
```

_shuffle Based on the code of Algorithm-Numerical-Shuffle-2009040301. Performs a one pass, fair shuffle on a list. If the list is passed as a reference to an array, the shuffle is done in situ. The subroutine returns the list in list context, and a reference to the list in scalar context.

```
Example:
@shuffled = _shuffle (1, 2, 3, 4, 5, 6, 7);
```

_regional_bpp Computes the mean base pair probabilities for a 18 bp region having the highest mean base pair probability plus its surrounding counterpart. All calculations are performed on the base pairs formed by the mfe structure and only if the structure contains at least 21 base pairs.

```
Example
my ($max_window_bpp, $surrounding_bpp, $start_i, $end_i)
   = _regional_bpp();
```

AUTHOR

Daniel Gerlach, <daniel.gerlach at unige.ch>

SUPPORT

You can find documentation for this module with the perldoc command.

```
perldoc RNA::Features
```

TO DO

COPYRIGHT & LICENSE

B.5.2.2 *Algorithm::LIBSVM*

A basic Perl interface to libsvm-2.89

VERSION

Version 0.022

SYNOPSIS

```
use Algorithm::LIBSVM;
my $o = Algorithm::LIBSVM->new();
```

PUBLIC METHODS

new Creates a new Algorithm::LIBSVM object

```
Example:
my $o = Algorithm::LIBSVM->new(
   Model => 'sample.scale.model',
   Scale => 'sample.range',
);
```

predict Classify a new input object by supplying a precalculated feature file or by passing two references, the features reference and the labels reference.

```
Example:
# Supply a reference to an array of arrays for the features
# and a reference to an array for the labels
my $predictions = $o->predict($feat, $labels);
foreach my $el (keys %$predictions) {
    print "$el => $predictions->{$el}{'class'} ";
    print "$predictions->{$el}{'p'}\n";
}
```

```
# Supply a precalculated feature file (--feature tab
# --header)
my $predictions = $o->predict("features.txt");
```

read_tab Reads a tab separated file and returns a reference to an array of arrays for the features and a reference to an array for the corresponding ids. Don't use this method if you have large input files as this is very memory inefficient.

```
Example:
my ($ids, $features) = $o->(file.txt);
my $first_id = $id->[0];
my $first_features = $features->[0][0];
```

AUTHOR

Daniel Gerlach, <daniel.gerlach at unige.ch>

SUPPORT

You can find documentation for this module with the perldoc command.

```
perldoc Algorithm::LIBSVM
```

COPYRIGHT & LICENSE

C

NOTES

COLOPHON

This thesis was typeset with LaTeX 2ε using Hermann Zapf's *Palatino* and *Euler* type faces (Type 1 PostScript fonts *URW Palladio L* and *FPL* were used).

The typographic style was inspired by Bringhurst's genius as presented in *The Elements of Typographic Style* (Bringhurst, 2002). It is available for LaTeX via CTAN as "`classicthesis`".

Final Version as of June 1, 2010 at 14:23.

DECLARATION

I hereby declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research. Whenever used, the published work of others has been clearly attributed. Parts of this thesis have been published by myself and colleagues in the following papers: Tapparel et al. (2007), Cordey et al. (2008), Gerlach et al. (2009), Tapparel et al. (2009), Elsik et al. (2009), Gatfield et al. (2009), Werren et al. (2010), Cordey et al. (2010), Kirkness et al. (2010)

Geneva, 2010

Daniel Gerlach