Analysis of in situ pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing

Corioni, Margherita; Antih, Nicolas; Tanackovic Abbas-Terki, Goranka; Zavolan, Mihaela; Kraemer, Angela

1868–1879 *Nucleic Acids Research, 2011, Vol. 39, No. 5* *Published online 9 November 2010*
doi:10.1093/nar/gkq1042

Analysis of *in situ* pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing

Margherita Corioni[1], Nicolas Antih[1], Goranka Tanackovic[1], Mihaela Zavolan[2] and Angela Krämer[1,*]

[1]Department of Cell Biology, Faculty of Sciences, University of Geneva, 30 quai Ernest-Ansermet, CH-1211 Geneva 4 and [2]Biozentrum of the University of Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, CH-4056 Basel, Switzerland

Received April 22, 2010; Revised October 9, 2010; Accepted October 11, 2010

ABSTRACT

The conserved pre-mRNA splicing factor SF1 is implicated in 3′ splice site recognition by binding directly to the intron branch site. However, because SF1 is not essential for constitutive splicing, its role in pre-mRNA processing has remained mysterious. Here, we used crosslinking and immunoprecipitation (CLIP) to analyze short RNAs directly bound by human SF1 *in vivo*. SF1 bound mainly pre-mRNAs, with 77% of target sites in introns. Binding to target RNAs *in vitro* was dependent on the newly defined SF1 binding motif ACUNAC, strongly resembling human branch sites. Surprisingly, the majority of SF1 binding sites did not map to the expected position near 3′ splice sites. Instead, target sites were distributed throughout introns, and a smaller but significant fraction occurred in exons within coding and untranslated regions. These data suggest a more complex role for SF1 in splicing regulation. Indeed, SF1 silencing affected alternative splicing of endogenous transcripts, establishing a previously unexpected role for SF1 and branch site-like sequences in splice site selection.

INTRODUCTION

Pre-mRNA splicing is critical for the expression of genetic information in most eukaryotes, and alternative splicing events largely contribute to proteome diversity in metazoans (1,2). On average, human pre-mRNAs contain eight exons separated by introns that vary in length between <100 and >100 000 nt (3), but signals that mark 5′ and 3′ splice sites are short and degenerate (4). Precise juxtaposition of cognate exons for intron removal is accomplished by dynamic interactions between the pre-mRNA, five small nuclear ribonucleoprotein particles (snRNPs) and more than 100 non-snRNP proteins (1). With few exceptions, splice sites are defined at the onset of spliceosome assembly. At this time U1 snRNP binds the 5′ splice site and the 3′ splice site is recognized by three proteins: splicing factor 1 (SF1, or mammalian branch point binding protein, mBBP) and the two subunits of the U2 snRNP auxiliary factor, U2AF65 and U2AF35. SF1 specifically binds the intron branch point sequence (BPS; 5,6), which is degenerate in mammals (YNCURAY; N = any nt, R = A or G, Y = C or U) but almost invariant in yeast (UACUAAC; 4). The underlined adenosine acts as the nucleophile in the first catalytic step of splicing (1). U2AF65 interacts with the polypyrimidine (Py) tract, located downstream of the BPS (7). U2AF35 recognizes the conserved AG dinucleotide that marks the intron 3′-end (8). SF1 and U2AF65 interact *in vitro* and *in vivo* and cooperatively bind the pre-mRNA (9–12). Recruitment of the U2 snRNP, which involves base pairing of the U2 snRNA with the BPS and binding of U2 snRNP proteins at and adjacent to the BPS, displaces SF1 from the spliceosome (13).

A hnRNP K homology/Quaking 2 (KH/QUA2) domain in the N-terminal half of SF1 (Figure 3A) contacts the bases of the BPS and buries the BPS-adenosine in a hydrophobic pocket of the KH-fold, which is thought to facilitate the formation of the BPS-U2

*To whom correspondence should be addressed. Tel: +41 22 379 6750; Fax: +41 22 379 6727; Email: angela.kraemer@unige.ch
Present addresses:
Nicolas Antih, Friedrich Miescher Institute for Biomedical Research, PO Box 2543, CH-4002 Basel, Switzerland.
Goranka Tanackovic, Department of Medical Genetics, Faculty of Biology and Medicine, University of Lausanne, 27 Rue du Bugnon, CH-1005 Lausanne, Switzerland.

snRNA helix (6). U2AF65 binds to the Py tract through two central RNA recognition motifs (RRMs; Figure 3A; 14,15) and an arginine–serine-rich N-terminal region contacts the BPS in a sequence-independent manner (16,17). A third, non-canonical RRM of U2AF65 (or U2AF homology motif, UHM) interacts with the N terminus of SF1 (9–11). UHMs are also found in other proteins, which engage in networks with ligand proteins and coordinate constitutive and alternative splicing (18–20).

These results implied an important role for SF1 in early spliceosome assembly, emphasized by a requirement of SF1 for embryonic development in mice and *Caenorhabditis elegans* and viability in human cells and yeast (10,21–24). However, depletion of SF1 from yeast or human splicing extracts slowed the kinetics of early splicing complex formation without compromising splicing outcome, suggesting a kinetic role for SF1 in splicing (13,25). Moreover, splicing defects were not apparent after SF1 silencing (24), suggesting SF1 is only required for the splicing of a subset of pre-mRNAs in human cells, as reported for yeast SF1 (26), or plays another essential role in mammalian cells. SF1 has been implicated in changes in alternative splicing mediated by the β-catenin/TCF4 complex involved in colorectal carcinogenesis, but it is not clear whether this function is direct (27). In addition, an increased susceptibility of $Sf1^{(+/-)}$ mice to colon cancer may relate to a function in alternative splicing (23). Finally, a mutation in SF1 in fission yeast leads to exon skipping (28). Other findings suggested roles for SF1 in nuclear pre-mRNA retention in yeast (16,29) and as a repressor of transcription activation and elongation in human cells (30,31).

To clarify the role of SF1 in splicing or other aspects of RNA biogenesis we exploited its RNA-binding activity to isolate cognate RNA targets from HeLa cells. We used the crosslinking and immunoprecipitation (CLIP) method, which combines UV crosslinking in live cells with immunoprecipitation of short RNA fragments bound to a protein of interest (32,33). Consistent with a function for SF1 in mRNA maturation, the majority of SF1 target sequences map to protein-coding genes. Of these, 77% are found in introns and the remaining exonic targets are preferentially located in 3′ terminal exons. We validated selected RNAs as SF1 substrates and confirmed the function of SF1 in splicing of these targets. Our results demonstrate that SF1 is not a constitutive splicing factor, but drives alternative splice site choices.

MATERIALS AND METHODS

Crosslinking and immunoprecipitation

CLIP was performed as described by Ule *et al*. (32) with the following modifications. HeLa cells were grown in Dulbecco's Modified Eagle Medium (DMEM; Invitrogen) containing 10% fetal calf serum (Sigma-Aldrich). Cells (4×10^7) were rinsed in PBS and irradiated at 400 mJ/cm^2 in a Model 2400 Stratalinker UV Crosslinker (Stratagene) in 10 ml ice-cold Hank's balanced salt solution. Digestion with RNase T1 (Roche) was performed at a final concentration of 15 U/ml. SF1 and crosslinked RNA were precipitated by incubation of the cell lysate for 1 h at 4°C with mAb24D1 (directed against a region common to all SF1 isoforms; Z. Rafi and A.K., unpublished data) bound to Dynabeads Protein A (Dynal Biotech). Linker RL5 (Supplementary Table S5) was ligated to RNA and adducts were purified from a denaturing 10% polyacrylamide gel followed by ligation of linker RL3. RT–PCR was performed with primers DP5 and DP3. DNA primers for the concatamerization step were DP5EcoR1, DP3EcoR1 and DP3. Concatamers were cloned into pCR2.1-TOPO (Invitrogen) and sequenced by Sanger sequencing.

Tag annotation

CLIP tag sequences were mapped to the human genome assembly version that was available for BLAST-like alignment tool (BLAT) searches on the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgBlat), hg18, and to the ENSEMBL genome and transcript databases (34).

Determination of a weight matrix describing the sequence preferences of SF1

To determine the sequence preference of SF1, we used a formalism that has been initially developed in statistical mechanics. Briefly, it can be shown that if a system can take on a certain set of states, each of them associated with an energy, then at equilibrium, each of the states will be visited in proportion to $e^{-\beta E}$, where E is the energy of a state, and β is the inverse temperature. In our case, the states correspond to all the ways one or more proteins can be bound to the RNA. In the simplest example there is only a single type of protein and a short RNA fragment that can be bound in one way only. Let $\boldsymbol{S}$ be the concentration of the RNA, $\boldsymbol{P}$ the concentration of the protein and $\boldsymbol{C}$ the concentration of the complex. RNA–protein binding can be described by $S+P \underset{k_r}{\overset{k_f}{\Leftrightarrow}} C$ with k_f and k_r being the forward and reverse rate constants. Let B and U be the measured intensities of bound and unbound RNA, with a scaling factor β_0 relating the signal intensity to the RNA concentration and P_T be the total protein. Finally, let us assume that we can write the dissociation rate of the protein from an RNA as $K = (k_r / k_f) = k_0 \exp(-E(s))$, where E(s) is the sequence-dependent energy of binding between the RNA and the protein. After rewriting the rate equations in terms of these quantities, we derive the following relationship: $\frac{1}{\mathrm{U}} = \exp[\mathbf{E(s)}]/\boldsymbol{k}_0\left[\boldsymbol{P_T}/\mathbf{B} - \boldsymbol{\beta}_0\right]$, which relates the intensity of unbound RNA to the total protein concentration and the intensity of bound RNA. With the assumption that the energy of interaction between the RNA and protein is composed of the individual energy contributions of individual positions in the binding site, and that these contributions depend only on the nucleotide that is present at that position in the binding site, the proportionality factor is $\exp[\mathbf{E(s)}]/\boldsymbol{k}_0$ depends only on the sequence of the RNA. That is, assuming a weight matrix model, $E(s)$ can be written as $E(s) = \sum_{i=1}^{L} E(s_i)$, where L is the length of the binding site (and the number of columns in the

weight matrix), and s_i is the nt at position i of the binding site. The probability to observe nt s_i at position i of the binding site is proportional to $e^{-E(s_i)}$ as mentioned above. Because we have three measurements for each individual RNA sequence, corresponding to three different total protein concentrations, we can fit a straight line to the three data points and the slope of this line corresponds to $\exp[\mathbf{E(s)}]/\boldsymbol{k}_0$. Furthermore, because $E(s)$ can be decomposed into the binding energies of individual positions in the binding site, the relative binding energies of the mutational variants of a given position in the binding site only differ in the relative contribution of their respective nucleotides at that position. That is, the entry w_i^B in the weight matrix, which represents the expected frequency of nucleotide B at position i in a binding site of the protein is given by

$$\frac{\exp[E(B)]}{\sum_{B' \in \{A,C,G,U\}} \exp[E(B')]} = \frac{\exp\left[\sum_{j \neq i} E(s_j) + E(B)\right]}{\sum_{B' \in \{A,C,G,U\}} \exp\left[\sum_{j \neq i} E(s_j) + E(B')\right]},$$

which can be derived from the slopes fitted above.

Determination of sequence motifs overrepresented in SF1 CLIP tags

We compared the frequency of tetrameric motifs in isolated SF1 CLIP tags to randomized sequences with the same nucleotide composition, which were obtained by shuffling the tags. Motifs, whose frequency in the real and the randomized sets were significantly different, were identified as follows. Given the frequencies of the motifs and the total number of tetramers in the real and randomized sets, we computed the posterior probability for the model that assumes that the frequency of a motif is different as opposed to being the same between the two datasets, as described (35). Motifs that were found enriched in introns and exons are shown in Supplementary Table S3. We also compared the frequency of tetrameric motifs in exonic/intronic tags with those found genome-wide in exons or introns as follows. We downloaded mRNAs sequences to Genbank (www.nlm.nih.gov) and, based on the Entrez Gene database, we chose the longest mRNA corresponding to each annotated gene in the genome. We mapped the mRNAs to the hg18 release of the human genome assembly, which we obtained from the University of California at Santa Cruz (genome.cse.ucsc.edu) with the GMAP alignment program (36). This mapping defined exonic and intronic genomic regions. We then sampled for each data set (exonic or intronic tags) 100 randomized data sets. Each of these datasets was generated by choosing with uniform probability the same number and length of regions as present in the CLIP sets from the appropriate type of region in the genome. For each of the 100 randomized sets, we computed the frequency of all possible tetramers. Supplementary Table S4 shows the frequency of the tetramers in the CLIP sets and the z-value of this frequency with respect to the frequencies computed in the randomized datasets.

In vitro transcription and RNA purification

RNAs were synthesized in the presence of [α-^{32}P]UTP with the T7-MEGAshortscript kit (Ambion) and gel-purified. Additional G's were added to the 5′-end of the RNAs to optimize transcription efficiency. The BPS RNAs (GGGGAGUAUACUAACAAGUUGAAUU; the BPS is underlined) are based on RNAs used by Berglund *et al.* (5), but contained four 5′ terminal G's instead of a C.

Protein expression and purification

The SF1 KH/QUA2 domain was expressed and purified as described (6). His_6-SF1-C4 and His_6-SF1-C4/W22A (10,11) were purified on Ni-NTA columns (Qiagen) under native conditions. GST-U2AF65Δ1-94 (16) was purified on glutathione-agarose (Sigma). Purified proteins were dialyzed against 20 mM HEPES-KOH pH 7.9, 100 mM KCl, 20% glycerol, 0.2 mM EDTA and 0.5 mM dithiothreitol.

Electrophoretic mobility shift assays

SF1-KH/QUA2 was incubated for 15 min at room temperature in 10-µl reactions in the presence of [α-^{32}P]UTP-RNA (~1 pmol for Figure 2A, ~50 pmol for Figures 3C and 4B), 16 mM Tris–HCl pH 8.0, 80 mM NaCl, 4 mM imidazole, 2 mM β-mercaptoethanol, 5 µg tRNA and 7 U rRNasin (Promega). Reaction products were resolved in native 8% polyacrylamide gels (acrylamide:bisacryamide = 80:1) in 1 × TBE at 4°C for 3 h at 100 V.

Cooperative binding was performed for 15 min at room temperature in 10-µl reactions containing GST-U2AF65Δ1-94 and His_6-SF1-C4 or His_6-SF1-C4/W22A, 50 pmol [α-^{32}P]UTP-RNA, 80 mM KCl, 16 mM Hepes-KOH pH 7.9, 16% glycerol, 2.4 mM $MgCl_2$, 0.4 mM DTT, 0.16 mM EDTA, 2.5 µg tRNA and 4 U rRNasin. Reaction products were resolved in native 5% polyacrylamide gels in 1 × TBE at 4°C for 3 h at 150 V.

Autoradiographs were quantified with the Molecular Imager FX (BioRad) and software Quantity One V 4.2.1 (BioRad).

RNA interference and semiquantitative RT-PCR of pre-mRNA targets of SF1

HeLa cells were grown in DMEM/high glucose containing 10% fetal calf serum (Sigma-Aldrich). Cells (2×10^5) were transfected in the presence of 73 nM SF1 or SF3a120 siRNAs (Dharmacon; Supplementary Table S5) and oligofectamine (Invitrogen) according to Dharmacon's instructions. Control transfections were performed in the presence of siRNA LUC, targeting firefly luciferase (LUC; 37) or in the absence of siRNAs. Duplicate samples were lyzed 48 h post-transfection for western blot analysis (24) and isolation of cytoplasmic RNA with the RNeasy Mini kit (Qiagen). RNA was treated with 2 U RQ-DNase (Promega) for 45 min at 37°C and

phenol–chloroform extracted. RT was done for 2 h at 40°C in the presence of 4 µg cytoplasmic RNA, 4 µg oligo $dT_{(12–18)}$ (Sigma-Aldrich), 40 U rRNasin and 200 U MMLV-RT (Invitrogen). PCR was performed with the Expand High-Fidelity PCR system (Roche) for 28 cycles. Control PCR reactions of 32 cycles indicated amplification in the linear range (not shown). Primers for amplification of SF1, SF3a120, H3F1 mRNAs have been described elsewhere (24). Other primers are listed in Supplementary Table S5. Forward primers were 5′-end-labeled with [γ-^{32}P]-ATP and T4 polynucleotide-kinase (Invitrogen). PCR products were separated in native 4% polyacrylamide gels, visualized by autoradiography and quantified by phosphoimaging as above. Alternatively, PCR products were separated in agarose gels and quantified with ImageJ 1.410 (National Institutes of Health). PCR products were gel-purified and sequenced.

RESULTS

Isolation of *in situ* RNA targets of SF1 from HeLa cells

We isolated SF1–RNA complexes from a whole-cell lysate of UV-irradiated HeLa cells with antibody mAb24D1 (see 'Materials and Methods' section). SF1 was efficiently removed from the lysate after immunoprecipitation in the presence, but not absence of mAb24D1 (Figure 1A). Non-specific binding of RNA to beads in the absence of antibody was not observed (I. Bagdiul and A. Krämer, data not shown). Following 5′-end-labeling of RNA, SF1–RNA complexes were resolved by PAGE and transferred to nitrocellulose. Autoradiography revealed a series of bands between 50 and 220 kDa superimposed on a smear of radioactivity (Figure 1B). SF1 isoforms detected by mAb24D1 in HeLa cells migrate between 70 and 85 kDa; SF1 crosslinked to RNA fragments of ≥30 nts are expected to migrate at least 10 kDa higher. Denaturing PAGE of RNA purified from consecutive slices of a membrane comparable to that shown in Figure 1B revealed an increase in RNA length with increasing size of the RNA–protein complexes (Figure 1C), suggesting that RNAs of increasing size are crosslinked to one protein rather than RNAs of similar size distribution being crosslinked to proteins of different molecular mass. As RNAs (CLIP tags) of 50–200 nt are best suited for analysis (32), we further processed SF1–RNA complexes of 80–100 kDa, followed by cloning and conventional sequencing.

Distribution of RNA targets in the human genome

We obtained ~360 SF1 CLIP tags and compared their sequences to the human genome by BLAT searches. Sequences with <90% match to the genome, multiple hits and exact duplicates were removed, resulting in a final set of 227 sequences (Supplementary Table S1). The majority of these (193) are located in protein-coding genes and 14 in genes of non-coding RNAs (Figure 1D). A total of 20 CLIP tags are present in intergenic regions, 7 of which map to the opposite strand of known genes. Intergenic CLIP tags were also found in other studies

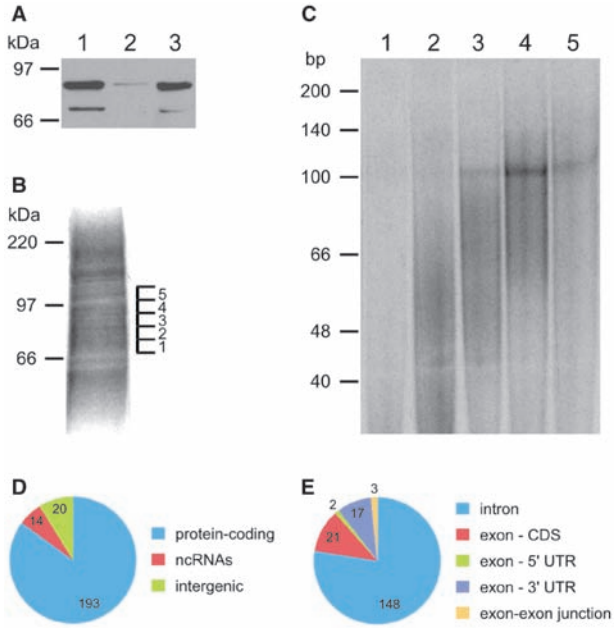

Figure 1. Isolation of SF1–RNA complexes from crosslinked HeLa cells. (**A**) Lysates of crosslinked HeLa cells (lane 1) were incubated with Dynabeads in the presence (lane 2) or absence (lane 3) of mAb24D1. Input (lane 1) and unbound material (lanes 2 and 3) were separated by 10% SDS–PAGE. After transfer to nitrocellulose SF1 was detected with mAb24D1. Protein size markers (in kDa) are indicated on the left. (**B**) SF1–RNA complexes eluted from mAb24D1-coupled Dynabeads were separated by 10% Bis–Tris NuPAGE and visualized by autoradiography. Protein size markers are shown on the left. Gel slices used for RNA analysis in panel C are marked on the right. (**C**) RNA extracted from nitrocellulose slices 1–5 in panel B was resolved in a denaturing 13% polyacrylamide gel. DNA size markers (in bp) are shown on the left. (**D**) Pie chart representing the number of SF1 CLIP tags in the protein-coding genes, genes encoding ncRNAs and in intergenic regions. (**E**) Pie chart representing the number of SF1 CLIP tags in introns or exons of protein-coding genes. Exonic CLIP tags were further divided into coding sequences (CDS), 5′- and 3′-UTRs and exon–exon junctions.

and may correspond to yet unannotated transcripts (38–41).

SF1 CLIP tags in protein-coding genes are largely confined to introns (Figure 1E and Supplementary Table S1). Despite SF1′s function in BPS binding (5,6), only 19 intronic tags are present at or close to 3′ splice sites, two map to 5′ splice sites and the remainder are located >40 nt from intron ends. Four intronic CLIP tags map to snoRNAs. One of these spans the intron–snoRNA border and is thus derived from the intron, which may also be true in the other cases.

Among 40 exonic CLIP tags, 21 are located in coding sequences, two in 5′-untranslated regions (UTRs) and 17 in 3′-UTRs (Figure 1E). Moreover, of the 43 tags located within exonic regions of known genes, 22 originated in 3′ terminal exons (Supplementary Table S1). Since human genes have on average 8–9 exons (3), this distribution indicates an enrichment of 3′ terminal exons among the exonic tags (*P*-value 9.9 x 10^{-10} in a binomial test assuming a probability to observe a 3′ terminal exon of 1/8). Additionally, in repeated sampling of 43 human exons from all human genes (dataset described in

'Materials and Methods' section), we only obtain an average of four 3′ terminal exons, indicating that targets in such exons are indeed overrepresented among SF1 CLIP tags. In addition, three CLIP tags span exon–exon junctions, suggesting that SF1 can bind mature mRNA. Finally, 42% of the CLIP tags in protein-coding genes are located in regions that are subject to alternative splicing (Supplementary Table S1).

Thus, SF1 mainly targets pre-mRNAs and binding sites are associated with constitutive and alternatively spliced regions of transcripts. An unexpectedly high percentage of SF1 target sites are located in 3′-UTRs and SF1 appears to bind mature mRNA as well. Compared with other studies that combined CLIP with high-throughput sequencing (38–41), our dataset only provides a glimpse at possible *in situ* SF1 binding sites and it is not surprising that, with a few exceptions containing two or more CLIP tags, only one SF1 CLIP tag was found per gene.

RNA binding preferences of SF1

SF1 specifically recognizes the BPS (5,6). Yeast SF1 selected the consensus BPS from a pool of randomized sequences (42), but information on binding preferences of human SF1 is limited. Mutational analyses and the NMR structure of the KH/QUA2 RNA-binding domain of human SF1 in complex with a UACUAAC-containing RNA showed that the URA in the 3′ half of the BPS is specifically recognized, whereas interactions with the 5′ half are less important (5,6,43). About 80% of the SF1 CLIP tags contained at least one copy of the URA motif; however, this sequence was not specifically enriched in our dataset compared to randomized sequences with the same overall length and nucleotide composition.

To precisely define binding preferences of human SF1, we tested 25-nt RNAs containing the consensus BPS or mutations at all BPS positions in electrophoretic mobility shift assays (EMSAs) for binding to recombinant SF1-KH/QUA2 protein (Figure 2A). The results were used to infer a weight matrix describing the sequence specificity of SF1 (Figure 2B and Supplementary Table S2). The derived motif resembles the human BPS consensus, except that SF1 strongly prefers an A at position −4 for optimal binding, whereas no nt bias is found at this position in the BPS consensus (4). The motif also comes close to the yeast BPS but lacks the preference for U at position −5. The information score at individual positions in the weight matrix is generally less than one (Supplementary Table S2), suggesting that human SF1 can bind a large variety of sequences. This is also reflected in only 4.4% of the CLIP tags containing the sequence ACUNAC (which we will refer to as the SF1 binding motif).

In contrast, Py-rich motifs were the most significantly enriched sequences in the CLIP tags relative to unrelated sequences with the same mononucleotide frequency. Approximately half of both intronic and exonic tags contained at least one occurrence of the motifs CCUG or UCCU (Supplementary Tables S3 and S4). CCUG repeats, which form mismatched stems, are known binding sites for MBNL proteins; the repeats sequester

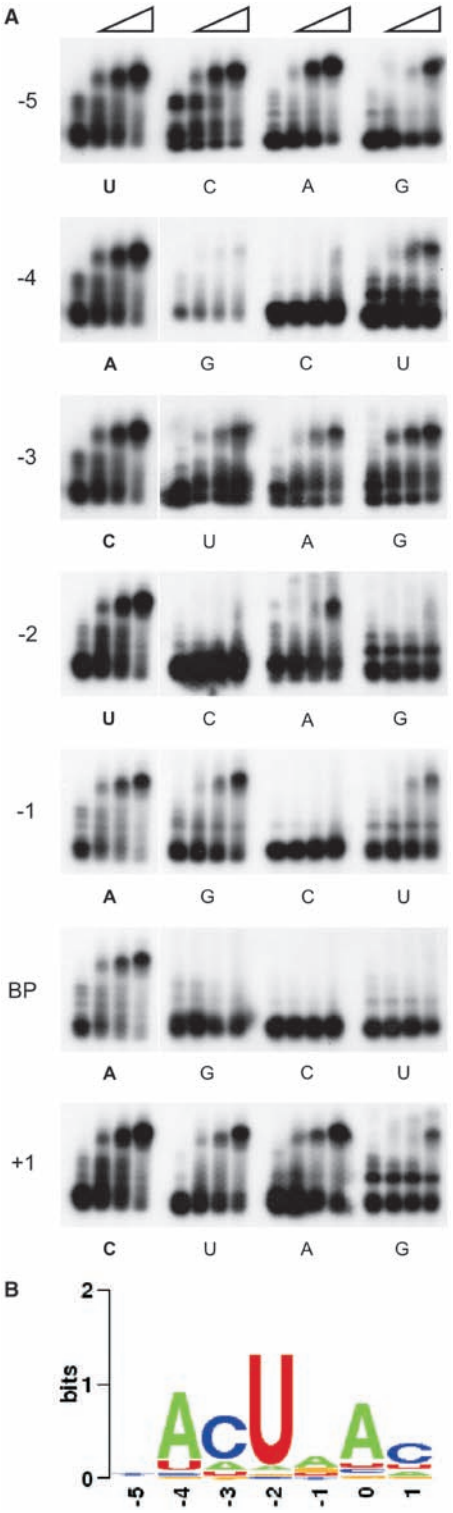

Figure 2. Determination of the optimal SF1–RNA-binding motif. (**A**) *In vitro*-transcribed RNAs containing the wild-type or mutant BPS were incubated with buffer or SF1-KH/QUA2 (5, 10 and 20 μM; indicated by triangles above the figure). Reaction products were separated by native PAGE and visualized by autoradiography. The same wild-type images were used for positions –5/–4/–3, –2/+1 and BP/–1. (**B**) Web logo representing the weight matrix for each position of the SF1 binding site derived from the quantification of the data presented in A.

the proteins, leading to splicing misregulation and myotonic dystrophy and other disorders (44). Such repeats are not observed in the CLIP tags (Supplementary Table S1). MBNL1 also binds a stem–loop structure flanked by two CCUG motifs and prevents U2AF65 binding upstream of a regulated 3′ splice site (45). That the CCUG motifs in the CLIP tags function in a similar way is unlikely, since only few are located close to splice sites. UCCU motifs (and to a lesser extent CCUG motifs) in SF1 CLIP tags are often embedded in longer Py-stretches, which may represent binding sites for U2AF65 (46). Py-rich motifs, particularly UCC, were also enriched with respect to random fragments chosen from the genome-wide set of exons or introns (Supplementary Table S4).

To further test SF1 binding preferences and validate CLIP tags as SF1 substrates, we analyzed the interaction of SF1 with *in-vitro*-transcribed RNAs representing 11 CLIP tags. SF1-KH/QUA2 bound 10 of the RNAs tested (Figures 3C and 4B; Supplementary Figure S1; data not shown). Mutations in potential SF1 binding sites introduced into six RNAs abolished the interaction (see below and data not shown). For example, SF1-KH/QUA2 efficiently bound RNA 2–50, which contains three partial matches to the SF1 binding motif and corresponds to a CLIP tag in intron 1 of cGMP-inhibited 3′,5′-cyclic phosphodiesterase 3A (PDE3A) pre-mRNA (Figure 3B and C). Mutation of all three motifs (M1) abolished SF1 binding. SF1 bound the 5′ most site (M3) with high affinity, whereas binding to the other two sites (M2 and M4) was weak. Binding to mutant RNAs containing two SF1 binding motifs (M5 and M6) was additive, suggesting that more than one molecule of SF1-KH/QUA2 can interact with RNA 2–50. This is also evident from the difference in migration of complexes of SF1-KH/QUA2 and RNAs with one or more binding sites.

These results indicate that SF1 binds RNA recognizing ACUNAC motifs. Variant ACUNAC motifs are bound with different affinities, suggesting that other factors, e.g. neighboring sequences or RNA conformation, may influence SF1 binding *in vitro*.

Cooperative interaction of SF1 and U2AF65 with RNA targets

Association of SF1 with branch sites likely reflects SF1 and U2AF binding to adjacent RNA sequences; in principle, SF1 and U2AF interactions with RNA could be competitive, neutral or cooperative. To address this, we performed EMSA with CLIP tag-derived RNAs and recombinant U2AF65Δ1-94 and SF1-C4 (Figure 3A), comprising the domains required for binding each other and RNA (11,16). U2AF65 and SF1-C4 alone bound RNA 2-50 (Figure 3D, lanes 2–6 and 7), containing an 11-nt Py tract following the 5′ most SF1 binding motif (Figure 3B). A ternary complex formed in the presence of both proteins, while the SF1-C4/RNA complex decreased (lanes 8–12). Consistent with cooperative RNA binding of SF1 and U2AF65, the ternary complex formed at a lower U2AF65 concentration when SF1 was present (compare lanes 2 and 3 with lanes 8 and 9). Cooperativity was also observed with mutant RNAs 2-50/M2, M3 and M4, which are less efficient substrates for SF1-KH/QUA2 and SF1-C4 than 2-50/WT, and with other CLIP tags (data not shown). Protein–RNA complexes migrated more slowly in the presence of high U2AF65 concentrations (Figure 3D), most likely due to binding of more than one U2AF65 molecule (14).

The importance of the U2AF65 interaction for SF1 RNA binding was further illustrated with CLIP tag 1-10, mapping to the 3′ splice site of exon 8 of the fibroblast growth factor receptor 1 oncogene partner (FGFR1OP) pre-mRNA. It contains a UCUCAC as the closest match to the ACUNAC motif and a 16-nt Py tract (Figure 4A). Neither SF1-KH/QUA2 nor SF1-C4 bound the corresponding RNA (1–10/WT), but mutation of UCU**C**AC to UCU**A**AC (RNA 1–10/M) converted the RNA into an efficient SF1 substrate (Figure 4B and C, lanes 6). Thus, at least in this sequence context, a C at position −1 is not compatible with SF1 binding, whereas a U at position −4 is tolerated, in agreement with the mutational analysis of the BPS (Figure 2). U2AF65 bound both RNAs with similar efficiency (Figure 4C, lanes 2–5) and ternary complexes in the presence of SF1-C4 formed more readily with RNA 1–10/M (lanes 7–10). No ternary complexes were apparent in the presence of SF1-C4/W22A, which does not interact with U2AF65 (11) but binds RNA 1-10/M (lanes 11–15). The more efficient formation of slow-migrating complexes with RNA 1–10/M can be explained by binding of SF1-C4/W22A and U2AF65 independently of one another. Thus, the ability of U2AF65 to increase SF1′s affinity for RNA 1–10/WT depends on the interaction between the two proteins. We conclude that U2AF65 can recruit SF1 to RNAs with suboptimal binding sites, explaining the isolation of CLIP tags with suboptimal or no SF1 binding motifs and containing Py-rich motifs.

Effects of SF1 depletion on splicing

A function of SF1 in splicing endogenous pre-mRNA targets was tested by RT–PCR after RNAi-mediated down-regulation of SF1 in HeLa cells. SF1 mRNA levels were reduced to ∼15% of the LUC or mock controls after transfection of two SF1 siRNAs, and SF1 protein was barely detectable (Figure 5A). Neither mRNA nor protein levels of the U2 snRNP-associated SF3a120 were affected after SF1 knockdown, as shown previously (24). In contrast, depletion of SF3a120 (used as a positive control, since it is thought to be a constitutive splicing factor) strongly reduced SF1 mRNA and protein levels. Negative effects on the expression of the intron-less H3F1 mRNA were not observed.

When we tested the splicing of pre-mRNAs containing CLIP tags, SF1 depletion modulated the ratio of inclusion/skipping of alternative cassette exons. We observed increased exon inclusion after SF1 silencing compared to the controls for FGFR1OP, TNFAIP3-interacting protein 1 (TNIP1) and procollagen-lysine 1, 2-oxoglutarate 5-dioxygenase 2 (PLOD2) pre-mRNA splicing (Figure 5B–D). The ratio of both, exon inclusion and skipping was changed in the case of splicing of the

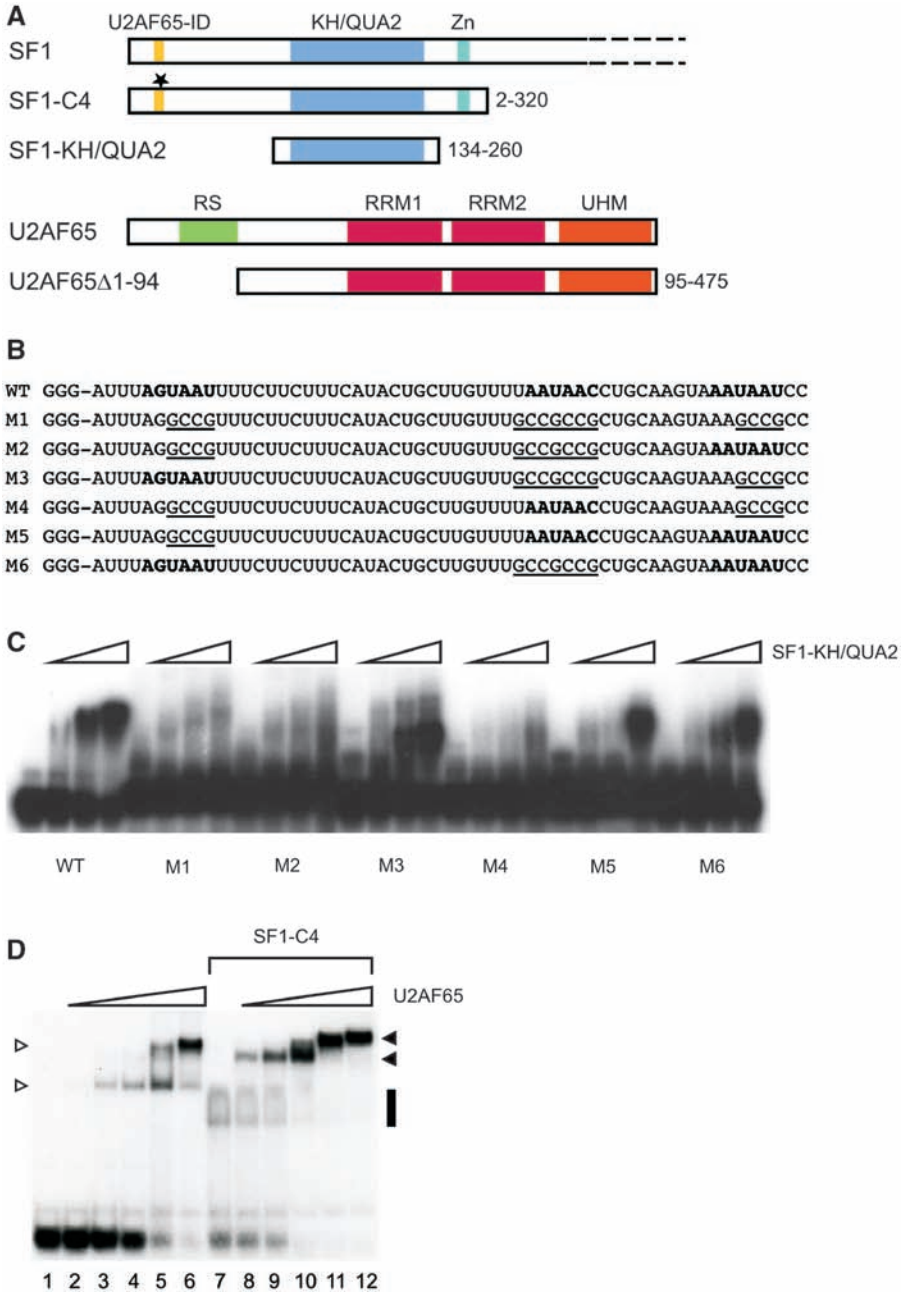

Figure 3. Cooperative binding of SF1 and U2AF65 to an endogenous SF1 target. (**A**) Scheme of SF1 and U2AF65 constructs used for EMSA. The U2AF65 interaction domain (U2AF65-ID), KH/QUA2 domain and the zinc knuckle (Zn) of SF1 are shown, as well as the arginine/serine-rich (RS) domain, RRMs 1 and 2 and the UHM of U2AF65. The star above the U2AF65-ID of SF1-C4 indicates the location of the W22A mutation. Numbers indicate amino acids of the truncated proteins. Variability in the length of SF1 isoforms is indicated by dashed lines. (**B**) RNAs corresponding to CLIP tag 2-50 (wild-type, WT, or mutants M1–M6) were transcribed *in vitro*. SF1 binding motifs are indicated in bold. Mutations are underlined. (**C**) Wild-type and mutant RNAs were incubated with buffer or SF1-KH/QUA2 (5, 10 and 20 μM; indicated by triangles). Reaction products were separated by native PAGE and visualized by autoradiography. (**D**) RNA 2–50 WT was incubated with buffer or U2AF65Δ1–94 (0.2, 0.5, 1, 2 and 4 μM; indicated by triangles) in the absence or presence of 6.6 μM SF1-C4 as indicated. Reaction products were separated by native PAGE and visualized by autoradiography. The migration of RNA 2–50 bound to U2AF65Δ1–94 (open arrowheads), SF1-C4 (vertical bar) or both (closed arrowheads) is indicated.

UPF3 regulator of nonsense transcript homolog A (UPF3A) pre-mRNA. The mRNA containing exons 1–5 was strongly reduced compared to the controls, whereas levels of mRNAs lacking exon 4 or exon 4 plus exon 2 or 3 (both of which are 107 nt in length) increased (Figure 5E). When we analyzed RT–PCR products from an independent experiment with SYBR Green, the 318 bp product from control cells was resolved into two closely migrating bands (Supplementary Figure S2B). PCR with primers in exons 2 or 3 and 5, as well as sequencing revealed that the slower migrating band corresponded to exons 1, 3 and 5, whereas the faster migrating band

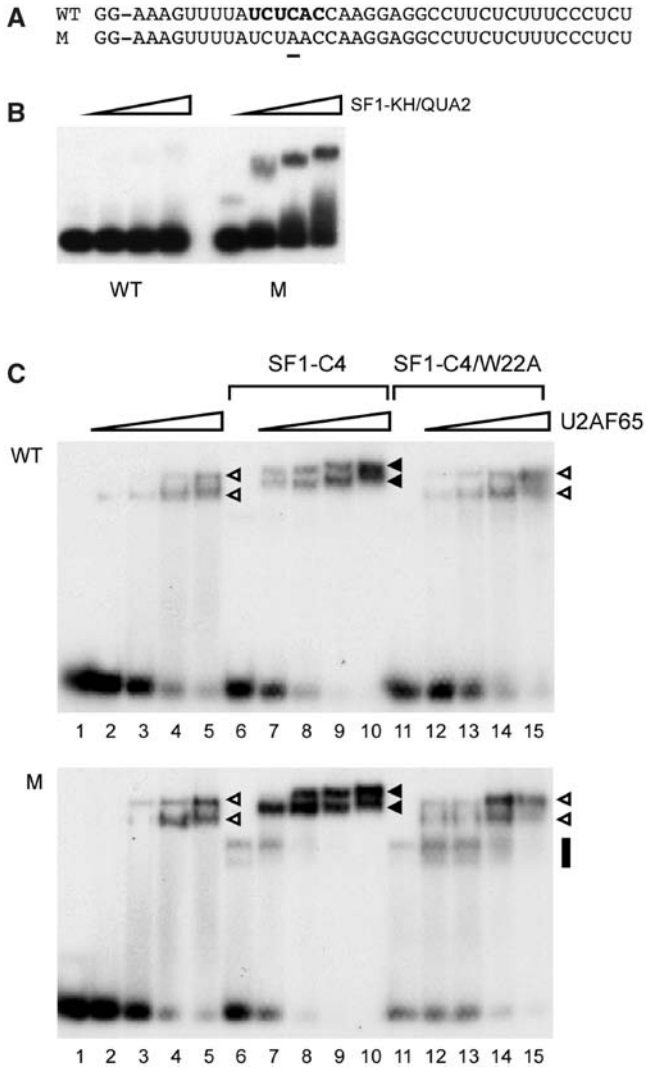

Figure 4. Cooperative binding of SF1 and U2AF65 to an endogenous SF1 target with a suboptimal SF1 binding site. (**A**) RNAs corresponding to wild-type (WT) and mutant (M) CLIP tag 1–10 were transcribed *in vitro*. A SF1 binding motif is indicated in bold. The mutation is underlined. (**B**) Wild-type and mutant RNAs were assayed as in Figure 3C. (**C**) Wild-type and mutant RNAs were tested as in Figure 3D in the absence or presence of 6.6 μM SF1-C4 or SF1-C4/W22A as indicated. The migration of RNAs bound to U2AF65Δ1–94 (open arrowheads), SF1-C4 (vertical bar) or both (closed arrowheads) is indicated.

contained exons 1, 2 and 5 (data not shown). SF1 silencing reduced the exon 1-2-5 mRNA to background levels, whereas the exon 1-3-5 mRNA was strongly increased. Together, these results show that SF1 can act as a negative or positive regulator of exon inclusion.

Of note, although depletion of SF3a120 resulted in a general reduction of mRNA levels (as apparent by inspection of the autoradiographs), in some cases the ratio of spliced products was also changed compared to controls (Figure 5B–E). This result is not unexpected, since SF3a120 has been shown to be involved in alternative splicing in *Drosophila* (47).

PDE3A pre-mRNA was identified as a potential SF1 target by three CLIP tags (Figure 3 and Supplementary Table S1). SF1 and SF3a120 knockdown reduced the single endogenous PDE3A mRNA detected in HeLa cells to 51–53 and 19%, respectively, compared to the controls (Supplementary Figure S3A), suggesting that SF1 affects the maturation or stability of PDE3A mRNA. Similarly, the level of a single mRNA of the asparagine-linked glycosylation 8 homolog (ALG8) decreased to 50–54 and 39% after SF1 and SF3a120 depletion, respectively (Supplementary Figure S3B). Thus, SF1 knockdown negatively affects the expression of these pre-mRNAs, although to a lesser extent than depletion of SF3a120.

These data demonstrate that SF1 indeed functions in splicing. However, it does not appear to be a constitutive splicing factor. SF1 may be required for the splicing of all introns of PDE3A and ALG8 pre-mRNAs. In contrast, in other cases tested, SF1 downregulation changed the ratio of mRNA isoforms without decreasing total mRNA levels, as observed after depletion of SF3a120. We conclude that SF1 plays a role in alternative splicing events.

DISCUSSION

We have isolated RNAs bound to SF1 in HeLa cells to clarify SF1′s role in splicing and/or other steps of RNA biogenesis. As expected for a protein that functions in mRNA maturation, the majority of SF1 targets mapped to protein-coding genes. We determined the RNA binding preferences of SF1, validated selected CLIP tags as bona fide SF1 ligands and demonstrated a role for SF1 in the splicing of several pre-mRNAs identified as potential targets. Our data indicate that SF1 is not required for the splicing of all introns, but influences alternative splicing decisions.

A role for SF1 in alternative splicing

Silencing of SF1 changed the ratio of exon inclusion/skipping in alternatively spliced products of several endogenous pre-mRNAs identified as potential SF1 targets (Figure 5 and Supplementary Figure S2). Depending on the event analyzed, SF1 activated or repressed exon inclusion, which is not without precedent. Nova, Fox and several hnRNP proteins regulate alternative splicing in both directions (2). Moreover, CLIP combined with high-throughput sequencing (HTS) provided a genome-wide view of splicing factor binding sites and demonstrated that Nova, Fox2 and PTB regulate exon inclusion/skipping depending on their binding sites with respect to the regulated exon (38,39,41).

Previous, mostly indirect evidence is in line with a function of SF1 in alternative splicing. First, exon inclusion into mRNAs of the soluble histocompatibility leukocyte antigen DQβ correlated with the strength of SF1 binding to BPS RNAs carrying disease-associated mutations (48). Second, SF1 expression is negatively regulated by the β-catenin/TCF4 complex involved in colorectal carcinogenesis and SF1 is required for changes in alternative splicing mediated by this complex (27). A direct involvement of SF1 is not clear, since β-catenin is itself implicated

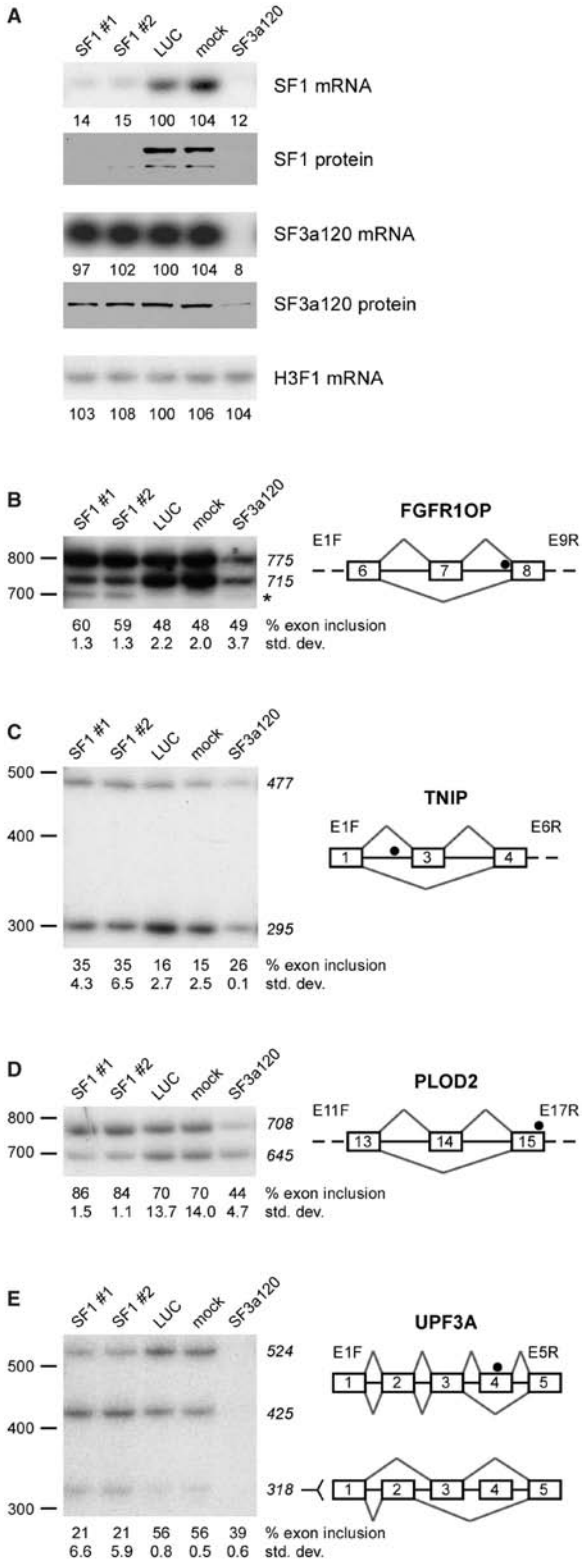

Figure 5. Effects of RNAi-mediated depletion of SF1 on the splicing of CLIP tag-containing pre-mRNAs. (**A**) cDNA from HeLa cells transfected in the absence (mock) or presence of siRNAs targeting SF1 (SF1 #1 and #2), luciferase (LUC) or SF3a120, as indicated on top of each panel, was PCR-amplified with primers specific for SF1, SF3a120 and H3F1 mRNAs, products were separated by PAGE and visualized by autoradiography. The numbers below the 'mRNA' panels indicate the percentage of mRNA normalized to the LUC control. Data represent the average of two experiments; the standard deviations were <1. HeLa cell lysates prepared in parallel were separated by SDS–PAGE and proteins revealed by western blotting. (**B, C, D** and **E**) The effect of SF1 depletion on the splicing of endogenous FGFR1OP (B), TNIP1 (C), PLOD2 (D) and UPF3A (E) pre-mRNAs was analyzed as in panel A. Sizes of DNA markers and spliced products are shown to the left and right of the images, respectively. Results of at least two experiments were quantified and are expressed as percent exon inclusion below the panels in addition to the standard deviation (std. dev.). The asterisk (panel B) indicates a PCR product that could not be identified by sequencing. The schemes depict pre-mRNA regions subject to alternative splicing (not to scale) and observed splicing patterns. Forward (F) and reverse (R) PCR primers including the exon number (E#) are shown above the schemes. Filled circles indicate the approximate location of CLIP tags. TNIP exon 2 is transcribed from an alternative promoter and not included in mRNAs transcribed from exon 1.

in splicing (49). Third, an increased susceptibility to colon cancer of Sf1 $^{(+/-)}$ mice could be caused by aberrant splicing due to reduced SF1 levels (23). Finally, a mutation in SF1 in fission yeast leads to exon skipping (28) and *Saccharomyces cerevisiae* SF1 is required only for the splicing of pre-mRNAs with suboptimal splice sites (26).

SF1 silencing also reduced the levels of single endogenous mRNAs (Supplementary Figure S3). Whether this is due to defects in splicing one or more introns is not clear. Lower mRNA levels rather than aberrantly spliced products may be due to faulty splicing upon SF1 depletion, which can introduce premature termination codons into mRNAs and cause mRNA degradation by nonsense-mediated decay (50).

SF1 is not an isolated case of a splicing factor thought to merely function in constitutive splicing but involved in alternative splice site choice. Other examples are U2AF65 and SF3a120 (47,51). It would be interesting to see whether these and possibly other proteins are required for all constitutive splicing events or, similar to SF1, also participate in the splicing of a subset of introns.

The SF1–BPS interaction is thought to facilitate base pairing of U2 snRNA with the BPS by pre-bulging the branch point adenosine (6). If SF1 is involved in the splicing of only certain introns, does another protein take over SF1′s role? The arginine–serine-rich domains of U2AF65 and SR proteins contact the phosphates of the BPS and could aid U2 snRNP in binding the BPS in the absence of SF1 (16,17,52). Moreover, the U2 snRNP-associated p14/SF3b14a, which specifically binds the branch point adenosine during A complex assembly (53,54), could facilitate base pairing.

SF1 binding preferences

An advantage of CLIP over other methods to identify targets of RNA-binding proteins is the possibility to locate approximate binding sites within targets and define consensus binding motifs (32,38–41,55). SF1 specifically recognizes the BPS (5,6,43). However, BPS-like sequences were not enriched in our dataset compared to random sequences with the same mononucleotide composition. Analysis of SF1 binding to a comprehensive set of BPS mutants defined ACUNAC as the consensus SF1 binding motif (Figure 2), which agrees in large part with previous limited mutational studies and the human

consensus BPS. The ACUNAC also resembles a BPS-like sequence identified downstream of exons specifically included in muscle (56). An exception is position −4, where SF1 favors an A, but no preference is evident in the consensus BPS (4), nor is the A specifically recognized in the SF1–BPS solution structure (6). In addition, a −4 A to G mutation did not affect SF1 function in previous studies (5,43). Another exception is position −1, where our data reveal the lowest information score for SF1 binding, although a −1 C can compromise SF1 binding (Figures 2 and 4, Supplementary Table S2). The SF1–BPS solution structure indicated coordination of a purine at this position (6); however, a −1 A to G mutation of an *in vivo* splicing reporter reduced SF1 function (43).

An information score of <1 at individual positions of the weight matrix (Supplementary Table S2) and <5% of the SF1 CLIP tags containing an exact match to the ACUNAC motif suggest a more relaxed RNA binding specificity of human SF1 compared to other splicing factors, which mirrors the variability of the mammalian BPS (4). This agrees with an RNA binding affinity of SF1, even for the consensus BPS, in the micro-molar range (5,6). In addition, the efficiency of SF1 binding to CLIP tag-derived RNAs did not always correlate with a perfect or suboptimal fit to the SF1 binding motif, suggesting that other parameters, such as neighboring sequences or RNA secondary structure, may influence the interaction. This may explain the variability in SF1 binding preferences observed in different experimental systems (see above).

The enrichment of Py-rich motifs in SF1 CLIP tags, many of which resemble U2AF65 binding sites (46; Supplementary Tables S1 and S3), is not unexpected, since SF1 binds RNA cooperatively with U2AF65 (9,43). Importantly, we not only observed cooperativity when SF1 alone bound RNA, but also when SF1–RNA binding was not evident (Figures 3D and 4C; data not shown). Thus, the interaction with U2AF65 increases the binding repertoire of SF1. Domains similar to the U2AF65 UHM, which contacts SF1, are present in other splicing factors engaging in networks with UHM ligands, including SF1 (19,20; G. Gregorovic and A. Krämer, unpublished data). We therefore speculate that not only U2AF65, but also other UHM proteins direct SF1 to appropriate binding sites in a pre-mRNA and regulate its function in the selection of specific splice sites.

Potential additional functions of SF1

Yeast SF1 dissociates from the spliceosome upon U2 snRNP recruitment (13) and human SF1 is most likely displaced from the BPS by binding of SF3b14a and base paring of the U2 snRNA (53,54,57), consistent with the absence of SF1 from late spliceosomes or spliced mRNPs (1,58). Thus, it may be surprising that 3′ terminal exons were overrepresented among the SF1 CLIP tags and that SF1 bound RNA spanning exon junctions. The binding of SF1 to 3′ terminal exons could hint to a role in coupling splicing of the last intron to 3′-end processing, as observed for U2AF65 and the U2 snRNP (59–61). Interestingly, the alternative splicing factor Nova also binds 3′ UTRs and regulates alternative polyadenylation in the brain (38). Similar to other splicing factors SF1 shuttles between the nucleus and the cytoplasm (G. Moreau and A. Krämer, unpublished data) and may thus have an additional role in the cytoplasm. For example, shuttling SR proteins function in constitutive and alternative splicing, mRNA export and translation, U2AF65 and PTB associate with defined sets of mRNAs in the cytoplasm, and TIA-1 plays a role in alternative splicing and translation repression (62–65).

In conclusion, our data confirm a function of SF1 in splicing, possibly of a subset of introns, and demonstrate a role in alternative splicing events. Future studies combining CLIP with HTS should yield information regarding SF1 binding with respect to regulated exons, which will help elucidate the mechanism of SF1 action. The resemblance of the SF1 binding motif to the mammalian consensus BPS moreover suggests that the SF1–BPS interaction plays an important role in alternative splicing decisions. Furthermore, the identification of an unexpectedly high number of SF1 target sites in 3′ terminal exons hints to additional functions of SF1 in other steps of mRNA biogenesis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Jernej Ule for helpful suggestions on the CLIP method, Dr Juan Valcárcel for the U2AF65 plasmid, Sophie Houngninou-Molango for the initial analysis of SF1 binding to the mutant BPS and Flore Mulhaupt and Mireille Quentin for technical assistance.

FUNDING

Swiss National Science Foundation (3100-068239 to A.K.); the European Science Foundation, EUROCORES Programme EuroDYNA (ERAS-CT-2003-980409 to A.K.); the European Commission (EURASNET-LSHG-CT-2005-518238 to A.K. and M.Z.); the Fondation MEDIC and the Canton of Geneva (to A.K.). Funding for open access charge: University of Geneva.

Conflict of interest statement. None declared.

REFERENCES

1. Wahl,M.C., Will,C.L. and Lührmann,R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701–718.
2. Chen,M. and Manley,J.L. (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.*, **10**, 741–754.
3. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
4. Schwartz,S., Silva,J., Burstein,D., Pupko,T., Eyras,E. and Ast,G. (2008) Large-scale comparative analysis of splicing signals and

their corresponding splicing factors in eukaryotes. *Genome Res.*, **18**, 88–103.
5. Berglund,J.A., Chua,K., Abovich,N., Reed,R. and Rosbash,M. (1997) The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell*, **89**, 781–787.
6. Liu,Z., Luyten,I., Bottomley,M., Messias,A., Houngninou-Molango,S., Sprangers,R., Zanier,K., Krämer,A. and Sattler,M. (2001) Structural basis for recognition of the intron branch site by splicing factor 1. *Science*, **294**, 1098–1102.
7. Zamore,P.D., Patton,J.G. and Green,M.R. (1992) Cloning and domain structure of the mammalian splicing factor U2AF. *Nature*, **355**, 609–614.
8. Moore,M. (2000) Intron recognition comes of AGe. *Nature Struct. Biol.*, **7**, 14–16.
9. Berglund,J.A., Abovich,N. and Rosbash,M. (1998) A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes Dev.*, **12**, 858–867.
10. Rain,J.-C., Rafi,Z., Rhani,Z., Legrain,P. and Krämer,A. (1998) Conservation of functional domains involved in RNA binding and protein–protein interactions in human and *Saccharomyces cerevisiae* pre-mRNA splicing factor SF1. *RNA*, **4**, 551–565.
11. Selenko,P., Gregorovic,G., Sprangers,R., Stier,G., Rhani,Z., Krämer,A. and Sattler,M. (2003) Structural basis for the molecular recognition between human splicing factors $U2AF^{65}$ and SF1/mBBP. *Mol. Cell*, **11**, 965–976.
12. Rino,J., Desterro,J.M., Pacheco,T.R., Gadella,T.W. Jr and Carmo-Fonseca,M. (2008) Splicing factors SF1 and U2AF associate in extra-spliceosomal complexes. *Mol. Cell. Biol.*, **28**, 3045–3057.
13. Rutz,B. and Séraphin,B. (1999) Transient interaction of BBP/ScSF1 and Mud2 with the splicing machinery affects the kinetics of spliceosome assembly. *RNA*, **5**, 819–831.
14. Sickmier,E.A., Frato,K.E., Shen,H., Paranawithana,S.R., Green,M.R. and Kielkopf,C.L. (2006) Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. *Mol. Cell*, **23**, 49–59.
15. Banerjee,H., Rahn,A., Davis,W. and Singh,R. (2003) Sex lethal and U2 small nuclear ribonucleoprotein auxiliary factor (U2AF65) recognize polypyrimidine tracts using multiple modes of binding. *RNA*, **9**, 88–99.
16. Valcárcel,J., Gaur,R.K., Singh,R. and Green,M.R. (1996) Interaction of $U2AF^{65}$ RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA. *Science*, **273**, 1706–1709.
17. Shen,H. and Green,M.R. (2004) A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. *Mol. Cell*, **16**, 363–373.
18. Thickman,K.R., Swenson,M.C., Kabogo,J.M., Gryczynski,Z. and Kielkopf,C.L. (2006) Multiple U2AF65 binding sites within SF3b155: thermodynamic and spectroscopic characterization of protein–protein interactions among pre-mRNA splicing factors. *J. Mol. Biol.*, **356**, 664–683.
19. Corsini,L., Bonnal,S., Basquin,J., Hothorn,M., Scheffzek,K., Valcárcel,J. and Sattler,M. (2007) U2AF-homology motif interactions are required for alternative splicing regulation by SPF45. *Nat. Struct. Mol. Biol.*, **14**, 620–629.
20. Corsini,L., Hothorn,M., Stier,G., Rybin,V., Scheffzek,K., Gibson,T.J. and Sattler,M. (2009) Dimerization and protein binding specificity of the U2AF homology motif of the splicing factor Puf60. *J. Biol. Chem.*, **284**, 630–639.
21. Mazroui,R., Puoti,A. and Krämer,A. (1999) Splicing factor SF1 from *Drosophila* and *Caenorhabditis*: presence of an N-terminal RS domain and requirement for viability. *RNA*, **5**, 1615–1631.
22. Abovich,N. and Rosbash,M. (1997) Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell*, **89**, 403–412.
23. Shitashige,M., Satow,R., Honda,K., Ono,M., Hirohashi,S. and Yamada,T. (2007) Increased susceptibility of $Sf1^{(+/-)}$ mice to azoxymethane-induced colon tumorigenesis. *Cancer Sci.*, **98**, 1862–1867.
24. Tanackovic,G. and Krämer,A. (2005) Human splicing factor SF3a, but not SF1, is essential for pre-mRNA splicing in vivo. *Mol. Biol. Cell*, **16**, 1366–1377.
25. Guth,S. and Valcárcel,J. (2000) Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. *J. Biol. Chem.*, **275**, 38059–38066.
26. Rutz,B. and Séraphin,B. (2000) A dual role for BBP/ScSF1 in nuclear pre-mRNA retention and splicing. *EMBO J.*, **19**, 1873–1886.
27. Shitashige,M., Naishiro,Y., Idogawa,M., Honda,K., Ono,M., Hirohashi,S. and Yamada,T. (2007) Involvement of splicing factor-1 in beta-catenin/T-cell factor-4-mediated gene transactivation and pre-mRNA splicing. *Gastroenterology*, **132**, 1039–1054.
28. Haraguchi,N., Andoh,T., Frendewey,D. and Tani,T. (2007) Mutations in the SF1–U2AF59–U2AF23 complex cause exon skipping in *Schizosaccharomyces pombe*. *J. Biol. Chem.*, **282**, 2221–2228.
29. Galy,V., Gadal,O., Fromont-Racine,M., Romano,A., Jacquier,A. and Nehrbass,U. (2004) Nuclear retention of unspliced mRNAs in yeast is mediated by perinuclear Mlp1. *Cell*, **116**, 63–73.
30. Zhang,D., Paley,A.J. and Childs,G. (1998) The transcriptional repressor ZFM1 interacts with and modulates the ability of EWS to activate transcription. *J. Biol. Chem.*, **273**, 18086–18091.
31. Goldstrohm,A.C., Albrecht,T.R., Sune,C., Bedford,M.T. and Garcia-Blanco,M.A. (2001) The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. *Mol. Cell. Biol.*, **21**, 7617–7628.
32. Ule,J., Jensen,K.B., Ruggiu,M., Mele,A., Ule,A. and Darnell,R.B. (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science*, **302**, 1212–1215.
33. Ule,J., Jensen,K., Mele,A. and Darnell,R.B. (2005) CLIP: a method for identifying protein–RNA interaction sites in living cells. *Methods*, **37**, 376–386.
34. Curwen,V., Eyras,E., Andrews,T.D., Clarke,L., Mongin,E., Searle,S.M. and Clamp,M. (2004) The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–950.
35. Berninger,P., Gaidatzis,D., van Nimwegen,E. and Zavolan,M. (2008) Computational analysis of small RNA cloning data. *Methods*, **44**, 13–21.
36. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
37. Elbashir,S.M., Harborth,J., Weber,K. and Tuschl,T. (2002) Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods*, **26**, 199–213.
38. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X. *et al.* (2008) HITS–CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
39. Yeo,G.W., Coufal,N.G., Liang,T.Y., Peng,G.E., Fu,X.D. and Gage,F.H. (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat. Struct. Mol. Biol.*, **16**, 130–137.
40. Sanford,J.R., Wang,X., Mort,M., Vanduyn,N., Cooper,D.N., Mooney,S.D., Edenberg,H.J. and Liu,Y. (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.*, **19**, 381–394.
41. Xue,Y., Zhou,Y., Wu,T., Zhu,T., Ji,X., Kwon,Y.-S., Zhang,C., Yeo,G., Black,D.L., Sun,H. *et al.* (2009) Genome-wide analysis of PTB–RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol. Cell*, **36**, 996–1006.
42. Garrey,S.M., Voelker,R. and Berglund,J.A. (2006) An extended RNA binding site for the yeast branch point-binding protein and the role of its zinc knuckle domains in RNA binding. *J. Biol. Chem.*, **281**, 27443–27453.
43. Peled-Zehavi,H., Berglund,J.A., Rosbash,M. and Frankel,A.D. (2001) Recognition of RNA branch point sequences by the KH domain of splicing factor 1 (mammalian branch point binding protein) in a splicing factor complex. *Mol. Cell. Biol.*, **21**, 5232–5241.
44. Lee,J.E. and Cooper,T.A. (2009) Pathogenic mechanisms of myotonic dystrophy. *Biochem. Soc. Trans.*, **37**, 1281–1286.
45. Warf,M.B. and Berglund,J.A. (2007) MBNL binds similar RNA structures in the CUG repeats of myotonic dystrophy and its pre-mRNA substrate cardiac troponin T. *RNA*, **13**, 2238–2251.

46. Singh,R., Valcárcel,J. and Green,M.R. (1995) Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science*, **268**, 1173–1176.
47. Park,J.W., Parisky,K., Celotto,A.M., Reenan,R.A. and Graveley,B.R. (2004) Identification of alternative splicing regulators by RNA interference in *Drosophila*. *Proc. Natl Acad. Sci. USA*, **101**, 15974–15979.
48. Kralovicova,J., Houngninou-Molango,S., Krämer,A. and Vorechovsky,I. (2004) Branch site haplotypes that control alternative splicing. *Hum. Mol. Genet.*, **13**, 3189–3202.
49. Sato,S., Idogawa,M., Honda,K., Fujii,G., Kawashima,H., Takekuma,K., Hoshika,A., Hirohashi,S. and Yamada,T. (2005) β-catenin interacts with the FUS proto-oncogene product and regulates pre-mRNA splicing. *Gastroenterology*, **129**, 1225–1236.
50. McGlincy,N.J. and Smith,C.W. (2008) Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem. Sci.*, **33**, 385–393.
51. Hastings,M.L., Allemand,E., Duelli,D.M., Myers,M.P. and Krainer,A.R. (2007) Control of pre-mRNA splicing by the general splicing factors PUF60 and U2AF65. *PLoS ONE*, **2**, e538.
52. Shen,H. and Green,M.R. (2006) RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. *Genes Dev.*, **20**, 1755–1765.
53. Will,C.L., Schneider,C., MacMillan,A.M., Katopodis,N.F., Neubauer,G., Wilm,M., Lührmann,R. and Query,C.C. (2001) A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *EMBO J.*, **20**, 4536–4546.
54. Query,C.C., Strobel,S.A. and Sharp,P.A. (1996) Three recognition events at the branch-site adenine. *EMBO J.*, **15**, 1392–1402.
55. Sanford,J.R., Coutinho,P., Hackett,J.A., Wang,X., Ranahan,W. and Caceres,J.F. (2008) Identification of nuclear and cytoplasmic mRNA targets for the shuttling protein SF2/ASF. *PLoS ONE*, **3**, e3369.
56. Sugnet,C.W., Srinivasan,K., Clark,T.A., O'Brien,G., Cline,M.S., Wang,H., Williams,A., Kulp,D., Blume,J.E., Haussler,D. *et al.* (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput. Biol.*, **2**, e4.
57. Das,R., Zhou,Z. and Reed,R. (2000) Functional association of U2 snRNP with the ATP-independent spliceosomal complex E. *Mol. Cell*, **5**, 779–787.
58. Merz,C., Urlaub,H., Will,C.L. and Lührmann,R. (2007) Protein composition of human mRNPs spliced in vitro and differential requirements for mRNP protein recruitment. *RNA*, **13**, 116–128.
59. Vagner,S., Vagner,C. and Mattaj,I.W. (2000) The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3′-end processing and splicing. *Genes Dev.*, **14**, 403–413.
60. Kyburz,A., Friedlein,A., Langen,H. and Keller,W. (2006) Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3′ end processing and splicing. *Mol. Cell*, **23**, 195–205.
61. Millevoi,S., Loulergue,C., Dettwiler,S., Karaa,S.Z., Keller,W., Antoniou,M. and Vagner,S. (2006) An interaction between U2AF 65 and CF I(m) links the splicing and 3′ end processing machineries. *EMBO J.*, **25**, 4854–4864.
62. Izquierdo,J.M., Majós,N., Bonnal,S., Martinez,C., Castelo,R., Guigó,R., Bilbao,D. and Valcárcel,J. (2005) Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol. Cell*, **19**, 475–484.
63. Anderson,P. and Kedersha,N. (2002) Visibly stressed: the role of eIF2, TIA-1, and stress granules in protein translation. *Cell Stress Chaperones*, **7**, 213–221.
64. Zhong,X.Y., Wang,P., Han,J., Rosenfeld,M.G. and Fu,X.D. (2009) SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. *Mol. Cell*, **35**, 1–10.
65. Gama-Carvalho,M., Barbosa-Morais,N.L., Brodsky,A.S., Silver,P.A. and Carmo-Fonseca,M. (2006) Genome-wide identification of functionally distinct subsets of cellular mrnas associated with two nucleocytoplasmic-shuttling mammalian splicing factors. *GenomeBiol.*, **7**, R113.

SUPPLEMENTARY MATERIAL (Corioni et al.)

FIGURES

Figure S1

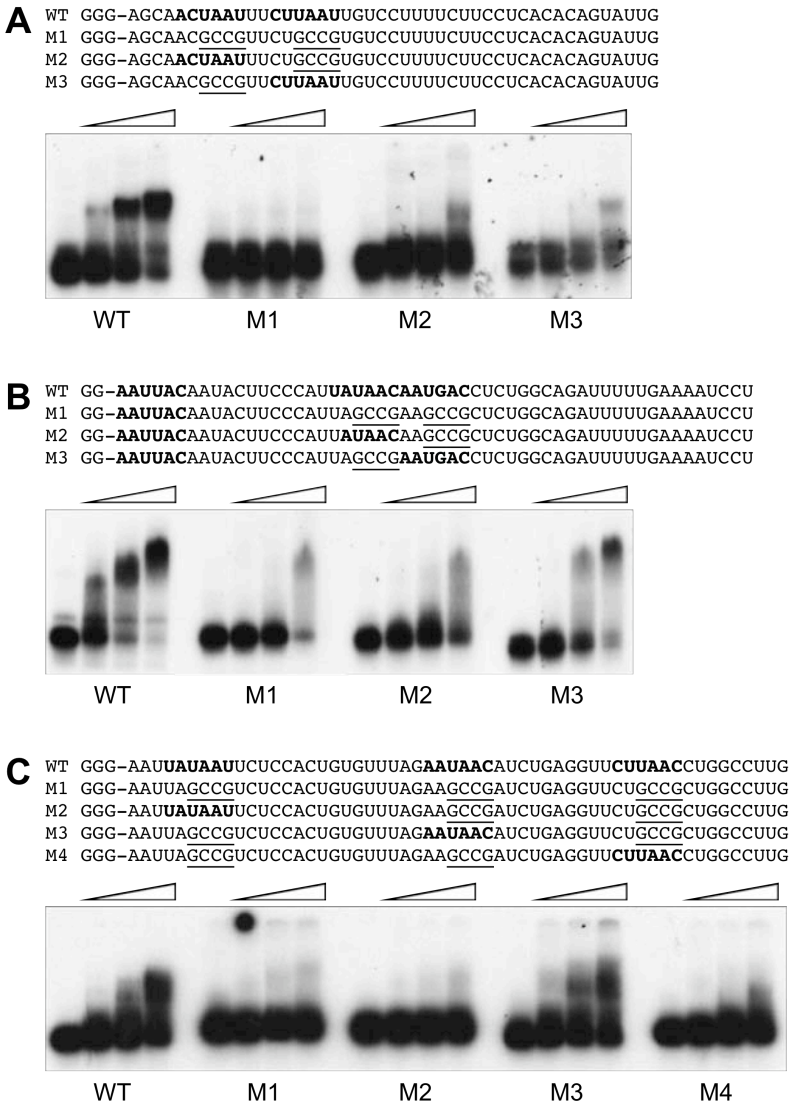

Figure S1. Binding of SF1-KH/QUA2 to different CLIP tag-derived RNAs. RNAs corresponding to wild-type or mutant CLIP tags 1-20a (**A**; overlapping the 3' splice site of exon 3 of the pre-mRNA encoding Erythrocyte membrane protein band 4.1 like 4A, EPB41L4A), 2-64 (**B**; in exon 15 of Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2 pre-mRNA, PLOD2) and 2-109 (**C**; in the 3' UTR of Ribosomal protein L37 pre-mRNA, RPL37) were transcribed *in vitro*. Top panels show the sequences of wild-type (WT) and mutant (M) RNAs. Potential SF1 binding motifs are indicated in bold; mutations are underlined. Bottom panels show results of experiments in which wild-type and mutant RNAs were incubated with buffer or SF1-KH/QUA2 (5, 10, and 20 mM; indicated by triangles). Reaction products were separated by native PAGE and visualized by autoradiography.

Figure S2

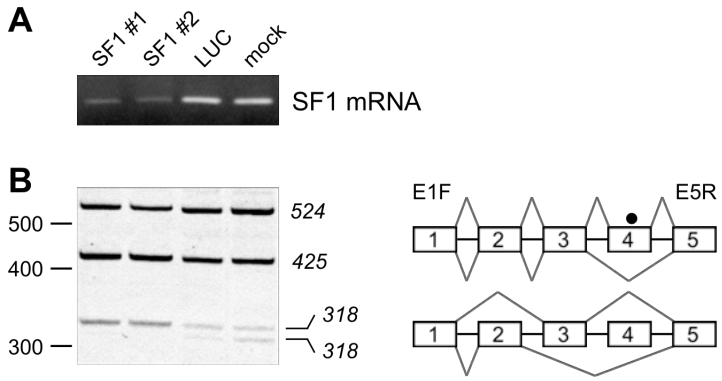

Figure S2. SF1 knockdown affects the alternative splicing of UPF3a pre-mRNA**.**

cDNA from HeLa cells transfected in the absence (mock) or presence of siRNAs targeting SF1 (SF1 #1 and #2) or luciferase (LUC), as indicated on top of panel A, was PCR-amplified with primers specific for SF1 (A) and UPF3A (B) mRNAs. (**A**) PCR products were separated in a 1% agarose gel and visualized by ethidium bromide staining. (**B**) PCR products were separated by 5% PAGE and visualized by staining with SYBR Green (Invitrogen). Sizes of DNA markers and spliced products are shown to the left and right of the image, respectively. The scheme depicts the amplified pre-mRNA region (not to scale) and observed splicing patterns. Forward (F) and reverse (R) PCR primers including the exon number (E#) are shown above the scheme. The filled circle indicates the approximate location of the CLIP tag.

Figure S3

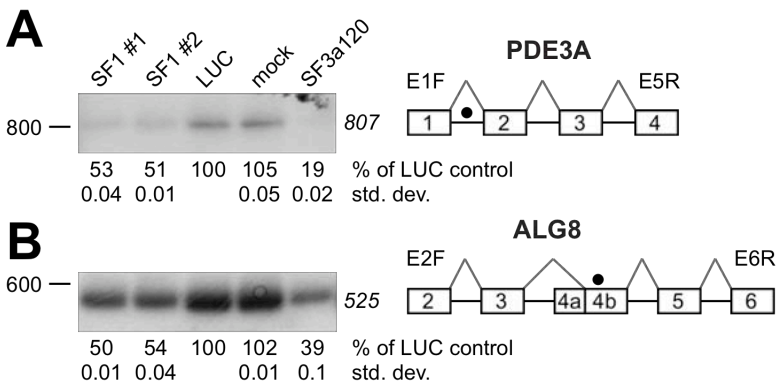

Figure S3. SF1 knockdown reduces PDE3A and ALG8 mRNA levels. cDNA from HeLa cells transfected in the absence (mock) or presence of siRNAs targeting SF1 (SF1 #1 and #2), SF3a120 or luciferase (LUC), as indicated on top of panel A, was PCR-amplified with primers specific for PDE3A (**A**) and ALG8 (**B**) mRNAs. The controls for SF1 knockdown are the same as those in Figure 5A. PCR products were separated by PAGE and visualized by autoradiography. Results of two experiments were quantified and are expressed as percentage of mRNA normalized to the LUC control. Exon inclusion (in % of the LUC control) is shown below the panels in addition to the standard deviation (std. dev.). Sizes of DNA markers and spliced products are shown to the left and right, respectively. The schemes depict the amplified pre-mRNA regions (not to scale) and observed splicing patterns. Filled circles indicate the approximate location of CLIP tags. PDE3A contains two SF1 CLIP tags in intron 1. ALG8 exon 4a is transcribed from an alternative promoter and not included in mRNAs transcribed from exon 1.

TABLES

Table S1: Sequences and genomic location of SF1 CLIP tags.

CLIP tag sequences were mapped to the human genome using BLAT (or in some cases NCBI BLAST) and locations were further defined by comparison to ENSEMBL entries, where possible.

CLIP ID	Sequence	Chr.	Coordinates[1]	Gene ID[1]	Description[1]	Location[2]	Features[3]
2-136 2-150 2-233	ACAUCAUCAUCUGCCUGGGGGGAGACGG GACGCUGCUGUACGCUUCCUCGCUUUUC CAGGGCAGCGUCCCUCCGGUCAUGGCCU UCCACCUGGGCUCCCUGGGCUUCCUGAC CCCAUUCAGCUUUGAG	1 -	1,676,707-1,677,775 and 1,677,557-1,677,615	NADK	NAD kinase	exon/exon junction	CDS
1-103	UUCGAAUAACCCUCUAGACAAGCAUACCA AACAAG	1 -	108,095,266-108,095,301	VAV3	Vav 3 guanine nucleotide exchange factor	exon	CDS
2-145	UUCAGGGAGGGAGACCCAUAAGGCUUCU UCCUCUCCUGUUCCUGGGAACUACCUGC GAGCUCUACUCUCAUCC	1 -	148,170,321-148,170,393	MTMR11	Myotubularin related protein 11	**intron**	**CDS**
2-227	AAGCUGUUCUCCUGCCUCAGCCUCCGAG UAGCUGGGAUUGCAGAAG	1 -	153,708,578-153,708,623	ASH1L	Ash1 (absent, small, or homeotic)-like (Drosophila)	intron	CDS
2-226	AAAGGGGAAUCUCACCCAGCACAGCACA GUCCCUUCCUAAGUCUGGUCUUUGAGCC UCAUUUACAAAGUCUUGGGCUCAAG	1 -	154,765,163-154,765,243	IQGAP3	IQ motif containing GTPase activating protein 3	intron	CDS
2-181a 2-218b	UUAAUUUUGCAUUGCUUUUAAUUGAAAU CACUACC	1 -	178,609,734-178,609,770	ACBD6	Acyl-Coenzyme A binding domain containing 6	**intron**	**CDS**
2-83	AAUAAAGUCAGAGGGGUGGCAUAACUCA GCCAAGGUCGCAUGGCUGGCCAGCAUCA UCU	1 -	19,951,835-19,951,893	TMC04	transmembrane and coiled-coil domains 4	**intron**	**CDS**
2-200 2-250	UUAAAUAAAUACUGUUUCUUUCUUUGCU UUG	1 -	191,265,096-191,265,124	UCHL5	Ubiquitin carboxyl-terminal hydrolase L5	intron	CDS
2-49	ACAAAAAUUAGCUGUGCAUGGUGGUGGG CGCCUGUAAUCCCAGCUACUUAGGAGGC UGAGG	1 -	223,868,497-223,868,557	ENAH	Enabled homolog (Drosophila)	intron	CDS
2-195	AUGCCUUCUUAGGAAGCAGCUAUGAAUC CAUUGUCCUUGUAGUUUCUUCCCUCCUG UUCUCUG	1 -	25,099,179-25,099,241	RUNX3	Runt-related transcription factor 3	3' exon	3' UTR
1-80b 1-76b	AAUUCACCAAGCGUUGGAUUGUUCACCC ACU	1 -	91,625,390-91,625,420	HFM1	HFM1, ATP-dependent DNA helicase homolog (S. cerevisiae)	**intron**	**CDS**
2-53	UUGUGUGGGAUUGAUGGUGGUGCCGAG GCAUGAAAGGCUAGUAUGAGCGAG	1 +	11,006,082-11,006,132	TARDBP	TAR DNA binding protein	**3' exon**	**3' UTR**
2-84	UUGUCAUCUUCACUUCCAUCCCUGUCAC UGUUAAUCCUUCCUACCUUCCUAAUUGU ACCUGGUGGCAG	1 +	11,942,003-11,942,070	PLOD1	Procollagen-lysine 1, 2-oxoglutarate 5-dioxygenase 1	**intron**	**CDS**
1-107b 1-66rev1	AUUUCUUCCUUCCUAGCGAGAGUAAGAC CAUG	1 +	162,915,880-162,915,911	PBX1	Pre-B-cell leukemia homeobox 1	intron	CDS
1-66b	AAUUAUGAUAGAUAGAUCCAAGGCCCCUA UCCUGUCCUUUUUCACUUG	1 +	165,577,284-165,577,333	POU2F1	POU class 2 homeobox 1	**intron**	**CDS**
2-199 2-264	AAAUAAAAACAUUGGUUCAAGUAGAGGUG UUCCUAGUGAAACACUUUGUUUUCUCCU CAAAAUG	1 +	213,864,969-213,865,032		on opposite strand of USH2A	intergenic	
2-60	AGGAUGGUCUUGAUCUCUUGACCUCGUG AUCCACCCACUUUUGCCUCCCAAA	1 +	46,147,986-46,148,037	MAST2	Microtubule associated serine/threonine kinase 2	**intron**	**CDS**
2-119	CACAUUUUUUUUCUUUUCUUUUCUUUAG UUGGAGAACAUCAAUUAACA	1 +	56,912,612-56,912,661	PRKAA2	Protein kinase, AMP-activated, alpha 2 catalytic subunit	intron/3 'SS	CDS
2-111	CUUUGCUUGCUUUUUUGAGACAGAAUGU CCCUCUGGUCUCAGCUCACUGCCAUUUC CACUUCCCGGGUUCAAGUGAUUCUCUUG CCUCAGCCUCCCAAG	1 +	6,865,567-6,865,671	CAMTA 1	calmodulin binding transcription activator 1	intron	CDS
1-80c 1-76a	AUAUAUGUCUUAUCCCUUGGAAGUAACG	1 +	62,680,150-62,680,177	USP1	Ubiquitin specific peptidase 1	intron	CDS
2-11	AGAAGACUUUAUCUUUUUUUUCUGUAGC AACCAU	10 -	89,078,010-89,078,044	CR614919		**intron/3' SS**	
2-113	CAAGUCAAUCACAAGAGUCUUUCCUCAG GACGGCUGCCAUACUUCAG	10 +	104,798,455-104,798,501	CNNM2	Cyclin M2	intron	CDS
2-76b	UUUCUCCUAUAUUUCUUUUUCUGUGUCU UCAGGUUCUGGAGAUAUAUCUUC	10 +	14,984,379-14,984,429	SUV39H2	Suppressor of variegation 3-9 homolog 2 (Drosophila)	intron/3' SS	CDS
2-268	AAAACCUGAUUGAAGAGAGAAUUAAGGAG GCUAGUCCCCACCCUGAGGCUUGGCUUU G	10 +	60,213,688-60,213,746 and 60,212,470-60,212,528	BICC1	Bicaudal C homolog 1 (Drosophila)	intron	CDS

1-68	ACUUUUUCCAGCUUCCCAGUGAUUAUAUGUUG	10 +	93,729,958-93,729,989	BTAF1	BTAF1 RNA polymerase II, B-TFIID transcription factor-associated, 170kDa (Mot1 homolog, S. cerevisiae)	intron	CDS
2-14	AAUUUAUGUGAUGGUUAACUUAAACAUUUUCUCUUCCAGCUUGAGU	10 +	96,332,670-96,332,717	HELLS	Helicase, lymphoid-specific	**intron/3' SS**	**CDS**
2-55	UUUAUUCUUUUUUCCUUUUUUCAGAUCACCAUCUAAGUUACAUCUUUAGCUCAGGUCCAUCCUUCUCAAGAUCUCCUUCUUAGCCCCCCAGCCCCUGGUG	10 +	99,467,281-99,467,380	MARVELD1	MARVEL domain containing 1	3' exon	3' UTR
1-6a	CAACUACCUUCCUUGGAUGUCUGAGUGAC	11 -	122,433,994-122,434,022	HSPA8	Heat shock 70kDa protein 8	**intron**	**CDS**
2-261	AAGAGAAUCCAUAUGACCUUGAUGCUUG	11 -	33,120,032-33,120,059	CSTF3	Cleavage stimulation factor, 3' pre-RNA, subunit 3, 77kDa	**exon**	**CDS**
2-197a	CCAGGGGAGUCAAUGUGUGACCUCAACACUCCAGGAAGGAGAGUGGGCAAGUGGGAUGCCAUGGACAGAGCCACAGAAAAACAGA	11 -	44,908,503-44,908,586	TP53I11	Tumor protein p53 inducible protein 11	intron	CDS
2-121a 2-232a	UCUAAAUUCAAGUUUGUUGCUUUGUUUUACAGUUUCCUUUCACUCCUCCCCCAUCCCCUCAUGAGUCAUAAG	11 -	47,501,882-47,501,955	CUGBP1	CUG triplet repeat, RNA binding protein 1	intron	5' UTR
1-140	UCAAAACCAGCAGGCUCAGAUGGUGAGCGGCGGGGGGUAAAGAGACAGCGGGAUGAGAAG	11 -	62,248,392-62,248,451	HNRPUL2	heterogeneous nuclear ribonucleoprotein U-like 2	**exon**	**CDS**
2-75	UUGGAAAUAGGAGAUGUAAUAGGAGCUUGCUCCGUCCACUCCACGCAUCGACCUGGUAUUG	11 -	62,365,695-62,365,755	U2 snRNA	U2 snRNA	ncRNA	
1-36	GGAGUCUCACUCUGUUGCCCAGGCUGGAAUGCGGUUUCAUGAUCUUGG	11 -	66,470,615-66,470,662	PC	Pyruvate carboxylase	intron	5' UTR
2-234b	UAAAUGCAUUGAUGGAAAAAAAGUGGGUAAAGAACUUACAG	11 -	77,509,822-77,509,862	ALG8	Asparagine-linked glycosylation 8 homolog (S. cerevisiae, alpha-1,3-glucosyltransferase)	**exon**	**CDS**
1-96 1-121	UUUUCAUACUCUCAUGACUGAUGAUAUU	11 -	77,569,291-77,569,318	KCTD21	Potassium channel tetramerisation domain containing 21	intron	5' UTR
2-257	UUUUGGCCUCACAUAUGCUUGUAUCAGCUCUUAGCCGAGACUCAUGUGCAUGCUCCUGCAGAUUUCG	11 -	8,549,310-8,549,376	STK33	Serine/threonine kinase 33	**intron**	**5' UTR**
1-137	AUAACAUGUCAAUUUUGUAGCCCACCAGAAAAGUUUGCAAUG	11 +	107,607,974-107,608,015	ATM	Ataxia telangiectasia mutated	intron	CDS
2-181b 2-218a	UCUUAGUCUAUAUUUCUGUACUCACCUUUCUUUAUCCACUCAUCAG	11 +	64,956,079-64,956,124	NR_002802/ NEAT1 (Men e/b)	Homo sapiens nuclear paraspeckle assembly transcript 1 (non-protein coding) (NEAT1), non-coding RNA	ncRNA	
2-63	AGUGGUUGGUAAAAAUCCGUGAGGUCGGCAAUAG	11 +	65,024,803-65,024,835	MALAT1	Homo sapiens metastasis associated lung adenocarcinoma transcript 1 (non-protein coding) (MALAT1), non-coding RNA	ncRNA	
1-120b	AAUAUGUUGUUUUUCUGGAACUUACUUAUG	11 +	65,024,832-65,024,861	MALAT1	Homo sapiens metastasis associated lung adenocarcinoma transcript 1 (non-protein coding) (MALAT1), non-coding RNA	ncRNA	
1-126	GAAGGAAAGUAUUGAACUGGGGGUUGGUCUGGCCUACUGGGCUGACAUUAACU	11 +	65,026,215-65,026,267	MALAT1	Homo sapiens metastasis associated lung adenocarcinoma transcript 1 (non-protein coding) (MALAT1), non-coding RNA	ncRNA	
2-142 2-188	UUAGUCUUUUCCAGAUGCAACUUUAAAAUCAG	11 +	65,028,691-65,028,722	MALAT1	Homo sapiens metastasis associated lung adenocarcinoma transcript 1 (non-protein coding) (MALAT1), non-coding RNA	ncRNA	
1-20b	AAAAAAUAGCCUAUUUACUUUAAAUAAACCA	11 +	65,028,921-65,028,951	MALAT1	Homo sapiens metastasis associated lung adenocarcinoma transcript 1 (non-protein coding) (MALAT1), non-coding RNA	ncRNA	
2-172 2-180	AACCACUAGUUCUUUCAGAUGGUAUUCUUCA	11 +	65,028,981-65,029,011	MALAT1	Homo sapiens metastasis associated lung adenocarcinoma transcript 1 (non-protein coding) (MALAT1), non-coding RNA	ncRNA	
2-246	ACUUGGCCAAGCUAGCAUCUUAGCGGAAGCUGAUCUCCAAUGCUCUUCAG	11 +	65,030,012-65,030,061	MALAT1	Homo sapiens metastasis associated lung adenocarcinoma transcript 1 (non-protein coding) (MALAT1), non-coding RNA	ncRNA	
1-108b	UACCAUUCUGUUUUUCCUACUUGGAGCAUUCCAUUUGUAAAAG	12 -	107,619,284-107,619,326	CORO1C	Coronin, actin binding protein, 1C	**intron**	**5' UTR**
1-94	AGAAUUCCCCAAAGCUGCUUUUGUGUCUG	12 -	119,067,355-119,067,383	GCN1L1	GCN1 general control of amino-acid synthesis 1-like 1 (yeast)	**intron**	**CDS**

1-133	AACGACUGACUGCCUGGGCUGCCCCCCG	12 -	120,413,215-120,413,238	FBXL10	F-box and leucine-rich repeat protein 10	intron	CDS
1-92	AAANUUAGGAUAUUUGGGAUAUANCAGUGGCCUUCCUUGUUGUGUAUG	12 -	45,045,157-45,045,164 and 45,044,712-45,044,752	SLC38A2	Solute carrier family 38, member 2	**exon/exon junction**	**CDS**
1-109	AAAUUCUCUAAAAGCAGUUUGUCUGCAAUUUUG	12 -	78,733,647-78,733,674	PPP1R12A	Protein phosphatase 1, regulatory (inhibitor) subunit 12A	**intron**	**CDS**
2-138	AAAAUUCUUCUUUGCCCAGGUACCUCAUACAUACAUGGACUAUACUACUUG	12 +	20,460,974-20,461,023	PDE3A	Phosphodiesterase 3A, cGMP-inhibited	intron	CDS
2-50	AUUUAGUAAUUUUCUUCUUUCAUACUGCUUGUUUUAAUAACCUGCAAGUAAAUAAUCC	12 +	20,504,510-20,504,566	PDE3A	Phosphodiesterase 3A, cGMP-inhibited	intron	CDS
2-81	AUUUCUCCCAUUUGGAAUGAUUGUAUUUACCCAGUGCCUGUACCAUCAUUGUAUCUGGGAAGUAACUAACUUGCUUUUGAUUUUGCAGGUUUAUAGG	12 +	20,707,840-20,707,936	PDE3A	Phosphodiesterase 3A, cGMP-inhibited	intron	CDS
2-57	UACCUCUGUCACUCCAUGACUUGUGUGCAUUCAAUCUUUCUAUUGUAACU	12 +	22,699,117-22,699,166	ETNK1	Ethanolamine kinase 1	**intron**	**CDS**
1-62	AAUUGUUAAUGCAGUUAAGGAUCCUGAUGCCAAUGGGAAAUCCUUUG	12 +	4,649,188-4,649,234	NDUFA9	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 9, 39kDa	exon	CDS
2-220	CUUUUUCUCCUGGCUUUGGGGUCUUUCCCGCCCCACCCCCGUGCCUGUGCCUGGGUGGUUUUUUCG	12 +	52,689,856-52,689,921	HOXC8	Homeobox C8	intron	CDS
2-143b	AUACUGGCUCACUGCAACCUCCAACCUCCCGGGUUCAAGUGAUUCUCUUGC	12 +	63,331,470-63,331,520	RASSF3	Ras association (RalGDS/AF-6) domain family 3	**intron**	**CDS**
1-50 2-77	AUGUUCCAUGAUGGACGUUUAUCCCAUAAGACGUCUCCCUCCUGAUAGAUGUCCGUAUAUGCAGAUUAAACUACCA	13 -	110,331,317-110,331,393	ANKRD10	Ankyrin repeat domain 10	**intron**	**CDS**
1-80a 1-76c	CUUUUCUGAAUAGUGGCCAUCCUG	13 -	113,169,891-113,169,913	DCUN1D2	DCN1, defective in cullin neddylation 1, domain containing 2 (S. cerevisiae)	**intron**	**CDS**
1-146	UAAGGUGUCACUGUAUUUAACUGUUGCACUU	13 -	43,905,693-43,905,723	TBC22D1	TSC22 domain family, member 1	3' exon	3' UTR
2-38	AAAUUACAAAUGCAAUUUCAUUUAAACUUCUAG	13 -	76,516,922-76,516,954	MYCBP2	MYC binding protein 2	3' exon	3' UTR
2-37	GUCUGGUUCUAAAAUGCCCCUGUAAAAGCGGUCGGCGUCCUGCACCGGCGACAGCA	13 -	87,128,175-87,128,230		on opposite strand of SLITRK5	intergenic	
1-112	AAUUAUUUAAAGUGAAUACCUACCCACUCUAGUUUCCUGUUCUG	13 +	100,918,148-100,918,191	ITGBL1	Integrin, beta-like 1 (with EGF-like repeat domains)	intron	CDS
2-153 2-269	AGUUUGCUCCAUUCCAGAAGAUAGCCAA	13 +	114,069,903-114,069,930	UPF3A	UPF3 regulator of nonsense transcripts homolog A (yeast)	**exon**	**CDS**
1-66a	UAGUGUAGUUGUUUCUGAUUGUAAAAAUAGUCAUA	13 +	31,809,714-31,809,748	BRCA2	Breast cancer 2, early onset	exon	CDS
2-214	UAAGGUUUCAUAAGCCCAAACUGCUGUUCUUGGGAUUUAUGAAAUAACUGCUUGCAUUGUCUUCUGGCCCUUAGCUCUCAGACUUG	13 +	40,931,035-40,931,119	C13orf15	chromosome 13 open reading frame 15	**intron**	**CDS**
1-38b	GAGCUGGAAGGUAUGGUGCUAUCUGCGGAUUUGGCCUUUCCUUGAUCAAAUGGAUCCUU	13 +	79,015,752-79,015,809	NDFIP2	Nedd4 family interacting protein 2	**exon**	**CDS**
2-166	AUAUUCUCUGCUAGUAACUUUCUAUCUGCUGAUAGUCCCAUCUCCAAUCUGUGGCUACAA	13 +	91,667,758-91,667,817	GPC5	Glypican 5	intron	CDS
1-65	CAUUUGUGUCCUGAUGUGUCUUGAUGACAUGCU	13 +	95,561,210-95,561,242	HS6ST3	Heparan sulfate 6-O-sulfotransferase 3	intron	CDS
2-151a 2-158b 2-187b 2-219a	AAAAGAAUCUCAGCAAUACCUUCCAGUGGUUAUGGGGCCUUUAAUGAAGACG	13 +	97,460,183-97,460,234	IPO5	Importin 5	exon	CDS
2-4	GGACAGUCUUUGUGGUCCAUUUANCANAUACUCUAUGUACCUG	14 -	20,634,817-20,634,860	ZNF219	Zinc finger protein 219	intron	5' UTR
2-44	AACAGAUUCCAUUACCUCGCUGAACCUGUCCCCUUAUAUG	14 -	90,204,747-90,204,786	TTC7B	Tetratricopeptide repeat domain 7B	intron	CDS
2-95	UUAUUGUUCUAUUCUGCAAACUUGAAUACAGAUCUGCUUUUAUGCUGUCCUUGUAUUAAGUAACAGUG	14 +	56,819,262-56,819,330	C14orf108	Chromosome 14 open reading frame 108	intron	CDS
1-125	AAUACUAAUUUGCUGAACAUCCCUUUGUAACG	14 +	66,648,901-66,648,932	GPHN	Gephyrin	intron	CDS
1-147	AUUAAACACUUGGUGAUCAAGGACUCCAAACUGAACAACACCAACAUAAG	14 +	69,530,893-69,530,942	SMOC1	SPARC related modular calcium binding 1	exon	CDS
2-78	UACAAAGUUUUGAAGACAGAGGCAAACCAUAAUUCUUCCAUG	14 +	70,554,472-70,554,513	PCNX	Pecanex homolog (Drosophila)	intron	CDS

2-45	UCAAAACUCCCGUGCUGAUCAGUAGUGG GAUCGCGCCUGUGAAUAGCCACUGCACU CCAGCCUGAGCAACAUAGCGAGACCCCG UCUCUU	14 + --- 14 -	49,123,258-49,123,347 --- 49,399,018-49,399,107	SRP_RNA	7SLRNA/srpRNA/srpRNA	ncRNA	
2-127	UUAUUUAGCAUUAGGUAUAUCUCCUAAU GCUAUCCCUCCCCGGCCCCCCCACCCUA CAACAGUCCCCG	15 +	41,473,044-41,473,111	TUBGCP4	Tubulin, gamma complex associated protein 4	intron	CDS
1-9a	UUGAAAUGUCCAUCAGGAGUGAUUAAAU GAAUUAU	15 +	62,223,144-62,223,179	SNX1	Sorting nexin 1	3' exon	3' UTR
2-116a	AUUACUAAAAUUGGUUUGCCAUUCUCAAU UCUCUUCACUUGUACUA	15 +	90,323,213-90,323,258	SLCO3A1	Solute carrier organic anion transporter family, member 3A1	intron	CDS
1-142	AAAAAGCAGGAGAUGUAAGGAAGGGCAG GCCUUG	15 +	92,694,402-92,694,435	MCTP2	Multiple C2 domains, transmembrane 2	**intron**	**CDS**
2-193 2-194	UAGUAUCUGUUNUUNUCAGUUUAAUAUN UG	15 +	94,090,070-94,090,099	U2-like snRNA	pseudogene	ncRNA	
1-16a	AUUAAAAUUCCUUGUAGGCGUCUCCUUC AAUGGCUACGUGACACUCCAUCUAG	15 +	99,332,270-99,332,319	LRRK1	Leucine-rich repeat kinase 1	intron	CDS
1-144	UUUUAAAUUGCAUCGUUUAGAAUACUAAG CAGUGUACCCUGCUGACAUUCACC	16 -	86,354,152-86,354,206	KLHDC4	Kelch domain containing 4	**intron**	**CDS**
2-46	AGGCUUUGCCUGUGUGUCCUUCACAGGG CAUGAGGUCCCGUCCUCGGGUGCCUCCC AGUCNGCAGCUNUNCAG	16 -	86,442,721-86,442,793	SLC7A5	Solute carrier family 7 (cationic amino acid transporter, y+ system), member 5	**intron/5' SS (37 nt)**	**CDS**
2-22	GGAUUCAUCUGUUUUCUUCACUUCCCUU UGCAUCUGAGAUCCUGCUGGAAACCACG GCAACCUGUAUCC	16 - ---- 16 +	18,418,737-18,418,805 ---- 16,296,047-16,296,115	NOMO2 ---- NOMO3	NODAL modulator 2 ---- NODAL modulator 3	**3' exon**	**CDS or 3' UTR**
2-225	CUUUCCUUUAAUCCGGCAGUGACCGUGU GUCAGAACAAUCUUGAAUCAUG	16 +	21,872,199-21,872,248	UQCRC2	Ubiquinol-cytochrome c reductase core protein II	5' exon	START
2-34	UAACCAUCCUGUGUGGUCUUCCUCAGCU AGAAGAUACCCUGAGCCUUG	16 +	66,657,941-66,657,991	DUS2L	Dihydrouridine synthase 2-like, SMM1 homolog (S. cerevisiae)	intron/3' SS	CDS
1-57	GGAGCAGAGUGGUCUCCUCUGUCCGCAC A	17 -	3,510,523-3,510,551		on opposite strand of CTNS	intergenic	
1-122	AAAUUGCUCCAAGAAGAAGGUGCUACUA UACAGUGGAUGUCAUCUG	17 -	33,067,514-33,067,561	DDX52	DEAD (Asp-Glu-Ala-Asp) box polypeptide 52	**exon**	**CDS or 3' UTR**
2-76a	UGGACUUCUGACCUCGGCCUCGGCGGGA GAGGACCCGGAACGCGGGCCUGGGCCC CAGG	17 -	37,682,339-37,682,398			intergenic	
1-23	UGGAGCCUUCCCUGUCUUUUUCUUUUCC UUCCGUUUUCUUCAUGCUGUAGUCUCUU C	17 -	51,967,302-51,967,358			intergenic	
2-35	GUUAGCUGUUAAGGUGGCGUGUUGCAGU GCAGAGUGCUUGGCUGUUUCCUGUUUUC UCCCGAUUGCUCCUGUGUAAAGAUGCCU UG	17 -	72,242,124-72,242,209	SFRS2	Splicing factor, arginine/serine-rich 2	**3' exon**	**3' UTR**
1-70	AACAAGCCUAAAGGAUUCCCACAGCU	17 -	76,056,994-76,057,019	NPTX1	Neuronal pentraxin I	**3' exon**	**3' UTR**
2-107	CUAUAACAUUUGCCUUUUUAUUGCAGGU UUUCCCUUGGAAUAUGGAUAAAUACACCA UGA	17 +	22,660,227-22,660,286	WSB1	WD repeat and SOCS box-containing 1	**intron/3' SS**	**CDS**
1-49	UUUUAAUCUCCUGACACUUGUGAUGUCU UCAAAGGAACCACUGAUG	17 +	24,071,713-24,071,758	RPL23A	Ribosomal protein L23a; U42B CD box snoRNA (complete overlap)	intron/ ncRNA	CDS
2-79	UCCUUCCCUUCUUCCUAUCCUCUGCCAC UUCCUCGCUGUUGCAGUCUGUGCAUACU GCCUCCUUCCCUG	17 +	35,747,571-35,747,639	RARA	Retinoic acid receptor, alpha	intron	CDS
1-95	UUCUUUAAAGAACUGCUCAAUAUUCAAUC UGUUG	17 +	55,126,254-55,126,287	CLTC	Clathrin, heavy chain (Hc)	3' exon	3' UTR
1-52	UAUCUGUCUUCAGAAAGUUUUUACGUCU GAUAUCCUUGUUGGUAAC	17 +	59,230,395-59,230,440	DDX42	DEAD (Asp-Glu-Ala-Asp) box polypeptide 42	**intron**	**CDS**
1-148	AGGUAGUGUUUUCUCCUGAGCGUGAAGC CGGCUUUCUGGCGUUGCUUG	17 + and -	several	U3 snoRNA		ncRNA	
2-30b	AAAAAUUAGUUGUUAGUUUUAUUGCUGU UCCCUUCCAGGUGAUAUG	18 -	21,895,519-21,895,564	SS18	Synovial sarcoma translocation, chromosome 18	**intron**	**CDS**
1-100	UUUCCUUUUCUGAAUGAAACUGCUAU	18 -	35,360,667-35,360,692	AK090603		intron	
2-128 2-253	AAAGUCUAUCUCAAAUUUUAUUUUUAAAC AUUACUAAAAAUAAACUUUAAAAUUACUG	18 -	65,953,166-65,953,223	RTTN	Rotatin	intron	CDS
2-204	CUAUAAUUCUAGACUCUAAAUCUUCUCUC CUGGUUACAGUGACCAUUUGAACACU	18 +	13,048,254-13,048,308	CEP192	Centrosomal protein 192kDa	**intron**	**CDS**
2-117	UCAUUUUACAUCUUGAGUCCUUAGUUUC UUCAUCUGUAAAAAUGAAG	18 +	2,602,173-2,602,220	KNTC2		intron	CDS

2-52	UUAAAGCAGACAUUUGUGCACUAAUCUG ACGCCCCUCCCACUGCUGGGACUCUUG	18 +	32,023,315-32,023,369	MOCOS	Molybdenum cofactor sulfurase	intron	CDS
1-118	AAGGACUCUUGAUUCUGACCCUAUCCAA AAUA	19 -	16,595,179-16,595,210	MED26	mediator complex subunit 26	**intron**	**CDS**
1-98 1-130	ACGUUGACCCUGGGCGAGGCUGACCACA ACCACUAUG	19 -	232,411-232,447	PPAP2C	Phosphatidic acid phosphatase type 2C	**3' exon**	**CDS**
1-43	ACACUCUGCUGUUUCCUGCCUAGGCGUG GCUGCAGCCAUGGCUAG	19 -	55,993,972-55,994,016	C19orf48	chromosome 19 open reading frame 48	**3' exon**	**5' UTR**
1-107a 1-66rev2	AAGAGCUCUCCCUUUGUGGAAAAAACUA GCUCUUCAUCUUCUCUG	19 -	8,918,374-8,918,418	MUC16	Mucin 16, cell surface associated	exon	CDS
2-72	GAAACUUCAGUCCUUUUCUGAUCUAUUG GAAAUAGCU	19 +	43,852,433-43,852,471	ACTN4	Actinin, alpha 4	intron	CDS
2-101	CUUCUCCUCUUGUCUCCAUAGCUUCCGC CCUGCGCUGCCAGUCUCCUACCUG	19 +	59,663,737-59,663,788	LENG8	Leukocyte receptor cluster (LRC) member 8	**intron/3' SS**	**CDS**
1-93	UAGCAGUUGGAGUAUUUUUCUUUCUUUU GUAUUUCUCUU	19 +	8,438,332-8,438,371	HNRPM	Heterogeneous nuclear ribonucleoprotein M	intron	CDS
1-106b	UACACUGGCAUUAUGCUUUGUUUAUCCU	2 -	127,964,133-127,964,160	IWS1	IWS1 homolog (S. cerevisiae)	**intron**	**CDS**
1-131	UCUUUGGGUUCCGAGGGGAGUAUGGUU G	2 -	132,729,068-132,729,095	BC045801		intron	
1-120c NEW	ANAUGUCUAGUAUACCANUUUUGGCCAA GCCAGAGUCCA	2 -	27,394,214-27,394,252	MPV17	MpV17 mitochondrial inner membrane protein	**intron**	**CDS**
1-44	UCUUUUCUUCUUCUCUCAAAGGUCUCAA ACAG	2 -	37,282,540-37,282,571	CEBPZ	CCAAT/enhancer binding protein zeta	intron/3' SS	CDS
2-221	AAUAUGAUAUACCUUUACAUUUAUUUAGA UGAUCUUUAUCUCAUCAAUAUGUUG	2 -	39,509,940-39,509,993	MAP4K3	Mitogen-activated protein kinase kinase kinase kinase 3	**intron**	**CDS**
1-108a	CUUCAUACAUUUUAGCUCCAUUAGGUAAA ACA	2 -	55,965,640-55,965,672	EFEMP1	EGF-containing fibulin-like extracellular matrix protein 1	**intron**	**CDS**
2-51	GUCCAGUAAAAUCUUCCUUGUCAGCUCU CAACAUGGAUAG	2 +	143,478,384-143,478,423	KYNU	Kynureninase (L-kynurenine hydrolase)	**intron**	**CDS**
2-152	ACUAGCAAUUUAUUCUUAGAGUGGCGAC UGAAUGUGCAUACCUCAAUG	2 +	189,581,825-189,581,872	COL3A1	Collagen, type III, alpha 1 (Ehlers-Danlos syndrome type IV, autosomal dominant)	**intron/3' SS (22 nt)**	**CDS**
2-82	AAUCUAAGUGAUCUGACUCAAUAUUCGU CACUACCACUGAGACA	2 +	202,849,435-202,849,478	NOL5	Nucleolar protein NOP5, NOP58; HBII-234 CD box snoRNA (complete overlap)	**intron/ ncRNA**	**CDS**
1-13 1-47	AAUGUUCUGCAAAAUGGUGCUAUGUCAU G	2 +	203,692,060-203,692,089	NBEAL1	neurobeachin-like 1	intron	CDS
2-122	UAGUUAUUGUAGCACAAUUACCGUAAACC UAAAUGGAACUUCUGUAUACUCUAACAUU ACUAUAUAUGUUAGUUUGUUUGGAAUUG UUUUUUCCCCAUAUG	2 +	211,240,430-211,240,532	CPS1	Carbamoyl-phosphate synthetase 1, mitochondrial	intron	CDS
2-7a	CAGACUCCCUCUUGGUAAAAUUGUGCAA GAAACGCACCACCUAGGUUUGUGCCUCC UGCCGUUGCUUACGAGGUUCUG	2 +	236,117,664-236,117,741	AGAP1	Arf-GAP, GTPase, ANK repeat and PH domain-containing protein 1	intron	CDS
1-123 1-127	GAAUUCUUCAAGUCCUGAUGAUAAAUUAU G	2 +	236,207,560-236,207,586	AGAP1	Arf-GAP, GTPase, ANK repeat and PH domain-containing protein 1	intron	CDS
1-138	UUCCAUAUAGAGUUCAUGCUCACCCCGA AGUCCUUCUUUGGCACUG	2 +	236,662,935-236,662,980	AGAP1	Arf-GAP, GTPase, ANK repeat and PH domain-containing protein 1	intron	CDS
2-105	AAUUAGCCAGGCAUGGUGGUGCAUGCCU CUAAUCCCAGCUACUCAGGAGGCUGAGG CACAAGAAUCUCUUG	2 +	32,541,377-32,541-447	BIRC6	Baculoviral IAP repeat-containing 6 (apollon)	intron	CDS
1-29	UACUUAAAUAAGGUUCCUCAUCAG	2 +	66,546,861-66,546,886	MEIS1	Meis homeobox 1	intron	CDS
1-15	UCCUAUAUAUCCUUGGGAAGUUACCUCU AGCAAGAGCUU	20 +	37,056,985-37,057,023	DHX35	DEAH (Asp-Glu-Ala-His) box polypeptide 35	intron	CDS
2-254	AAUUAAUAUCUAUUAAUGGCAUCAACCAU UCC	20 +	51,309,883-51,309,914	TSHZ2	Teashirt zinc finger homeobox 2	intron	3' UTR
1-64a	CCUAUGCUGCUCAGGUUUGUAUCACUUG GGCCUGGGUGUAUGACUCAG	20 +	58,246,272-58,246,319	AX747739		intron	
2-168	CAUGGAAAGAGAUCUUUGUUAGUUGGAG GAAGGAGAACUUGCCUGAGCUUCGGGCC UUUG	20 +	62,358,141-62,358,200	PCMTD2	Protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 2	intron	5' UTR
2-167	AUAAAUUCAAAACUGAUCCUUUUAGUUCC UCAUUAUAUGAUAUGAAGGGAUUAACUG UAG	21 -	39,639,669-39,639,728	HMGN1	High-mobility group nucleosome binding domain 1	**exon**	**5' UTR**
2-184	CUUUUAUCUUUUCUUUCUCAACAGCUUU CCAUUCAGUCUGGAUCCUUCCAUGACUA CAGCUAUUUAAG	21 +	16,172,954-16,173,021	USP25	Ubiquitin specific peptidase 25	3' exon	3' UTR
2-64	AAUUACAAUACUUCCCAUUAUAACAAUGA CCUCUGGCAGAUUUUUGAAAAUCCU	3 -	147,277,262-147,277,315	PLOD2	Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2	**exon**	**CDS**

1-86	AGUGCUGAUGCCUGAAUUUAACAGUGAG	3 -	178,391,952-178,391,979	TBL1XR1	Transducin (beta)-like 1X-linked receptor 1	**intron**	**5' UTR**
1-119	UGGGUGCUCCAGAUGUUUCCCCACAGAAC	3 +	138,180,728-138,180,756	IL20RB	Interleukin 20 receptor beta	**intron**	**CDS**
1-21	AAAGCUUCACUUCAAAACCCCUGCAAUAGCUGGGUUUACA	3 +	14,502,887-14,502,927	SLC6A6	Solute carrier family 6 (neurotransmitter transporter, taurine), member 6	3' exon	3' UTR
1-67b	CUUAAAAUUAUGUAUUUUGUCUUGGGCUGCAAUUU	3 +	153,666,130-153,666,164	MBNL1	Muscleblind-like (Drosophila)	3' exon	3' UTR
1-11	AAGUAGAAAACCAAAUAUUAUAGAAACUUAGAUUUCCAAUCCAACUUAGCGAAAUCAAUGUA	3 +	32,775,257-32,775,318	CNOT10	CCR4-NOT transcription complex, subunit 10	**intron**	**CDS**
1-39	AUUAAAUCUUAGUGACUCUCAUUUCAGCAUCCUU	3 +	51,549,372-51,549,405			intergenic	
1-60	ACAGAGAGUGGGUGACCCUCCCCUCUGCUCCCAGCUAACACACACCUCAG	3 +	70,032,047-70,032,096	MITF	Microphthalmia-associated transcription factor	intron	CDS
2-116b	AAUGUCAUUCUCAUUUCUGUAGUGGAAAGAGCAUAGACUUUGGAGUCAGACAGACUUUGUUUCAAAUCUUAGCU	3 +	9,449,587-9,449,660	SETD5	SET domain containing 5	intron	5' UTR
2-5	UGUGCUGUGCUUCUCCGGUUGCCUGCUGCCUCCCCAGCUUGGGCAUGCACAGGUGUCCUGUGGCAAG	4 -	4,582,221-4,582,289	STX18	Syntaxin 18	intron	CDS
2-99	UUCUACCCUUUGUCUUCCUCCCCCAGUGCCACAGACCCCUCAGCCCCAGUCCCCAUGGCCUGUGUGUUG	4 +	1,148,203-1,148,271			intergenic	
1-99	AAAUUAACUCUAAAACCUUGAUCCUAAAAUUUAAU	4 +	41,643,375-41,643,409	TMEM33	Transmembrane protein 33	intron	CDS
1-20a	AGCAACUAAUUUCUUAAUUGUCCUUUUCUUCCUCACACAGUAUUG	5 -	111,643,915-111,643,960	EPB41L4A	Erythrocyte membrane protein band 4.1 like 4A	intron/3' SS	CDS
2-192	UUAGCUCAAGCGGUUACCUCCUCAUGCCGGACUUUCUAUCUGUCCAUCUCUGUGCUGGGGUUCGAGACCCGCGGGUGCUUACUGACCCUUUU	5 -	135,444,085-135,444,176	hsa-mir-886	hsa-mir-886	ncRNA	
2-159	AUUUCUCCCCUCUUUUCCCCCUCAUCAAGAAUGAAACCCNUGACUAUGCCCCAG	5 -	136,576,688-136,576,741	SPOCK1	Sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 1	intron	CDS
2-42	AAACAAAUUUCUGAUGUGGAACGUGAUGUUCAGAAUUUCCCAGCUGCCACAG	5 -	140,887,356-140,887,407	DIAPH1	Diaphanous homolog 1 (Drosophila)	exon	CDS
1-25	AAAAAUUGGCUCCUGUGUACACCCCCCAG	5 -	149,806,401-149,806,429	RPS14	Ribosomal protein S14	**intron**	**CDS**
1-85 1-115	CAGGAUGGACUUGCUCCCUGUCCUCAUGCAGCCUAAU	5 -	150,427,405-150,427,443	TNIP1	TNFAIP3 interacting protein 1	**intron**	**5' UTR**
1-6b	UAUCUCUUUAGCCAUAUUGUCUGUGUACCUAAGUCU	5 -	158,553,647-158,553,682	RNF145	Ring finger protein 145	intron	CDS
2-94	AACACAUUUCCCAAGCCCCUGGAGACCCACAGGGUUUGGCACUGUCCGUGAGGCUGUGCUCCUGAGGACGGA	5 -	178,969,130-178,969,201		on opposite strand of RUFY1	intergenic	
2-238	UCUCUUUAAUCUGUAGGUUCUCACUCCAUUUCUCCUUUUUUUCCCUUGCAAAC	5 -	31,494,829-31,494,881	RNASEN	Ribonuclease III, nuclear	intron	CDS
2-109	AAUUAUAAUUCUCCACUGUGUUUAGAAUAACAUCUGAGGUUCUUAACCUGGCCUUG	5 -	40,867,352-40,867,407	RPL37	Ribosomal protein L37	**3' exon**	**3' UTR**
1-120a	AAAAUUCUUAGACACUCUACACAAUUCUCUUAUCU	5 -	41,816,421-41,816,455	OXCT1	3-oxoacid CoA transferase 1	**intron**	**CDS**
2-155	ACUGUAGUAAUGAAAGGGUCAGUUCUCUACUUUUGAAUUUCCC	5 -	78,694,953-78,694,995			intergenic	
1-117	GUAAAAACUAAUUAAGCUUACUAUUUUUCGCUGCUGUUCAGGAUG	5 -	80,932,625-80,932,669	SSBP2	Single-stranded DNA binding protein 2	**intron**	**CDS**
1-106a	AAAAUAAUCUUAUAAACUGCUUUACCUUAUCUUUCCCCUGGCUCCAGGCAUUUAAUGG	5 +	102,370,912-102,370,969	PAM	Peptidylglycine alpha-amidating monooxygenase	intron	CDS
2-97	AAGUACCUUCCCUCCAGCAUGCUGAUCCCCUGCCUGUGUGCUGGACUUUUGAGUCCUCAGC	5 +	135,427,670-135,427,730			intergenic	
1-34	AUUGCUUAAUCAACUGAGACUUGUUGGCU	5 +	138,149,965-138,149,993	CTNNA1	Catenin (cadherin-associated protein), alpha 1, 102kDa	**intron**	**CDS**
2-216 2-262	AAAGUUGGUUAAUGAUGGUGCUGUUCCUGACCUCAUUCUUCUGACUUUAGAGUCCA	5 +	147,621,829-147,621,884			intergenic	
2-247	UAGUCUUGCUGGAAUUCUGAUUAAUCAUGGUCCUUCCCAAUCUCUCAGAAUGACCCAUAUCUUUCAG	5 +	17,306,722-17,306,788	BASP1	Brain abundant, membrane attached signal protein 1	intron	5' UTR
2-33	AAUGAUUGCUAUCCUCACAGAAAACUAUGGGGGCAAAUGGUAAUUUUUGUCACUGUCUUUUUUUUCUG	5 +	33,498,027-33,498,094	TARS	Threonyl-tRNA synthetase	intron/5' SS	CDS
2-40	UUUUACCAUGUUGGCCAGGCUGGUCUCAAACUCCUGACAUG	5 +	37,428,401-37,428,441	WDR70	WD repeat domain 70	intron	CDS

2-17	ACUUUUUUUUUUCUUCUCUUUCAGGACU UCAAACAAGUGUG	5 +	700,895-700,936	CEP72	Centrosomal protein 72kDa	intron/3' SS	CDS
2-146	AUUACCUCCUGUACCACCUGUGCUACCU CUGUGCCUCUCUGAACCA	5 +	76,611,342-76,611,387	PDE8B	Phosphodiesterase 8B	**intron**	**CDS**
2-12	AGAAAAAAUAAAAAAAAUAAAACUGCCCU UUUCU	5 +	94,074,439-94,074,473		on opposite strand of MCTP1	intergenic	
2-178	GUUGUCUUUGACUGCUUCUCACAUGCAA UAUCUGAUCUAUUCAGAAAAUCCUACUG	6 -	166,255,550-166,255,608	AK090688		intron	
1-17	CACCAUGGUUUAGUUCUUCGUCUCUCAU CCCAGCAAGACUGCACUAG	6 -	166,262,386-166,262,432	AK090688		intron	
2-237 2-137 2-267	AAGUCACACUUACCUGCCUGCUGCCUGC CUUCCUUCUCCCAUAGCCAUGUUCUUGG CUGGGGAGAAAGGGUAGAUGUAG	6 -	64,295,883-64,295,961		on opposite strand of PTP4A1	intergenic	
2-131	ACUCUUAAUCCUUACAGAUGGGAAAGGG CUCCUUCAAGUAUGCCUGGG	6 -	74,285,930-74,285,977	EEF1A1	Eukaryotic translation elongation factor 1 alpha 1	**intron/3' SS**	**CDS**
1-63	AAUCAAUGAUGAAACCUAUCCCGAAGCUG AUAACCUGAAG	6 -	86,444,056-86,444,095	SNHG5	Homo sapiens chromosome 6 open reading frame 160 U50B CD box snoRNA (complete overlap)	**intron/ ncRNA**	**non-protein-coding multiple snoRNA host gene**
2-173 2-263	AAUUAUGCUCAAGUCAUAUCUGUCUCAU UUCCCUGUAGGUGACCAUGCUUUGGGAU UUUCAUAG	6 +	111,931,154-111,931,217	BX647114		intron/3' SS	
2-217	ACUUCCUGACUCUGCUGGCCCCGUG	6 +	117,911,064-117,911,088	DCBLD1	Discoidin, CUB and LCCL domain containing 1	intron	CDS
2-228	UCAUUCUCAUACAAUACAGCUCACCCACU UAAAAUGUACAAUUCAGUGAUUUCUAG	6 +	133,643,547-133,643,602	EYA4	Eyes absent homolog 4 (Drosophila)	intron	CDS
2-135 2-235	UAGCUUGGACCCAUGGCCACAUUACGGC CAUGUUGUCACAGGGAUCUACCAUACUC UAUUGACUUUUCUUGUGUUCUGCCAUCC UG	6 +	157,351,798-157,351,883	ARID1B	AT rich interactive domain 1B (SWI1-like)	**intron**	**CDS**
1-10	AAAGUUUUAUCUCACCAAGGAGGCCUUC UCUUUCCCUCU	6 +	167,355,838-167,355,878	FGFR1OP ---- CCR6	FGFR1 oncogene partner ---- Chemokine (C-C motif) receptor 6	**intron/3' SS** **----** **intron/3' SS**	**CDS** **----** **5' UTR**
2-39 2-67	UUUCUUCACUAAUGUUCUUACUGUUCUA GGAUCCAACGCAGCAUACCACAUUG	6 +	18,278,583-18,278,637	AOF1	Amine oxidase (flavin containing) domain 1	intron	CDS
2-130	CUGUGAUAAUUUCUCCUGGGUUUAAGUU ACAUGACUGAUUCCUCAGAG	6 +	7,536,665-7,536,712	C6orf151	Chromosome 6 open reading frame 151	intron	CDS
1-113	AAGCUGCCAUGGCAUAAUACACAAUUUAU UUUGGAAUAGCAUAGAAGCUCAUCUG	7 -	110,408,100-110,408,154	IMMP2L	IMP2 inner mitochondrial membrane peptidase-like (S. cerevisiae)	**intron**	**CDS**
2-112	AGUUUGCUGGGCCUUGCGAGACUUCACC AGGUUGUUUCGAUAGCUCACACUCCUGC ACUGUGCCUG	7 -	128,384,522-128,384,587	TNPO3	Transportin 3	**exon**	**STOP**
1-134	AUGUUCUAUUUCAUGUAAGAGUUA	7 -	138,400,650-138,400,673	ZC3HAV1	Zinc finger CCCH-type, antiviral 1	intron	CDS
2-28	UCAAAGAGCAUGUCUUGCGAGUGGGAAU GUCAUUGGUACUACCACUUGGAAAACCA UUUGGAAUUACCUUGUAAUGUAAUGCCC AACCU	7 -	32,609,867-32,609,955	AK026768		intron	CDS
1-12	AAGAGAUGACUCUGGUAGGAGCUUACUU AGCUACAUUGGAAGGUGCUCCCUUUG	7 -	9,834,906-9,834,959			intergenic	
2-123 2-133 2-191 2-201 2-223	AGAUGUUCAGAUCAUUCUGGAGAGCUGC CACAACCAGAACAUUCUGGGUACCAUCC UUCAUCCAAAUGG	7 +	127,130,550-127,130,613 and 127,132,129-127,132,133	SND1	Staphylococcal nuclease and tudor domain containing 1	**exon/exon junction**	**CDS**
2-203	AACUUUACUUUUUCCCAAUGUGUUUAAC CUUGCACACAAACUCCAG	7 +	28,587,326-28,587,374	CREB5	CAMP responsive element binding protein 5	intron	CDS
2-245 2-251	GAAUUCAUAAAAUGCAAAGAACCCAG	7 +	30,437,455-30,437,482		on opposite strand of NOD1	intergenic	
2-118	CACUUUCUGAAAACUCCUGCUGAAGUAA CUGCACAAGAAUCCUUG	7 +	55,120,780-55,120,824	EGFR	Epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	intron	CDS
2-104	UUAAUAGCUAGCUGGAUGAAUGUUUAAC UUCUAGGCCAGGCACUACUCUGUCCCAA CAAUAAGCCCUG	8 -	128,300,360-128,300,427	AK125310		exon	
1-67a	CUUNUUUUCAUUAACCUGCUAUCCUCUU UACAAAUGGGAUUCUUCACCCACUCCCU UCUUGUAG	8 -	128,300,572-128,300,635			intergenic	
2-121b 2-232b	AAUAAUGACUGCUGACUCCUAAUUCAGU GGAUCUUCCCUGGCCACCGUUUUGUAUU	8 -	128,310,424-128,310,482	BC106081		exon	

	GAG						
2-160 2-239	UUGCUUUUCUGACAAAUUCUACUUAUUC UGUG	8 -	60,175,738- 60,175,769	TOX	Thymocyte selection-associated high mobility group box	intron	CDS
1-116 1-132	CAAUUAUUUUUAAAUAAUCUUGUUUcUCC UGCAGACAG	8 -	63,311,374- 63,311,411			intergenic	
2-80	CUUUUUUCUCCUUUCCUACAGUGAAACA AAACUUGAC	8 -	97,340,753- 97,340,789	MTERFD1	MTERF domain containing 1	**intron**	**5' UTR**
1-124	AGUAGGAUCAAUUCUGGCCUCUUG	8 +	124,569,992- 124,570,015			intergenic	
1-1	UUGCCAUCUUUNGNUUAUUUGCUNUGAA UCUAAGUCUUCNA	8 +	137,984,980- 137,985,019			intergenic	
2-171	UUGGUGUUUUUCCCCGUCCUUAGGUCCU AAUGCCCCGAGCCAAGUUCCACUGGUUC UUCUUGAAG	8 +	19,728,194- 19,728,258	INTS10	Integrator complex subunit 10	**intron/3' SS**	**CDS**
2-32	AAAAUGAUCUAGAAAUUAAAGAGGAAUAU GGAGGUUCUUUCCAGUCUCAAAUUA	9 -	102,313,293- 102,313,346	TMEFF1	Transmembrane protein with EGF-like and two follistatin-like domains 1	intron	CDS
2-205 2-266	ACCUGGGCCACCAUGAGACUCGAAGGCU UCUAGCUGUCCUACGACACGAGCCACGG UGGCCCUGCCUCUCCCCGCUCAUAGAAG CCUGGUCCCUCUG	9 -	126,320,336- 126,320,432	NR6A1	Nuclear receptor subfamily 6, group A, member 1	3' exon	3' UTR
2-176	UCAUUUCUCUAAAACAGUUUCUGCCACG UCAGAUUUGUGAAGGCUUGGUUUUGCCC UCAGCCACCCGCCACUGCACCUGACCAG GUC	9 -	129,250,707- 129,250,793	RPL12	Ribosomal protein L12; U65 H/ACA box snoRNA (partial overlap)	intron/ ncRNA	CDS
1-16c	UUAUGUUUUCCAGCAACUUCUUCCUAGG AGCUUUUGUG	9 -	134,134,673- 134,134,710	SETX	Senataxin	**intron/3' SS**	**CDS**
2-265	AAAAGACAGAGACUCUUUCAGUUUGAG	9 -	134,193,584- 134,193,612	SETX	Senataxin	**exon**	**CDS or 5' UTR**
1-48	UUUAUUAUCCUCACCGCCCACUGAGUCU CUGCAGAAUCCUCCAGUCUUGAUUGCUC AGCCU	9 -	138,489,390- 138,489,450	SEC16A	SEC16 homolog A (S. cerevisiae)	**5' exon**	**CDS**
2-61	AAGUCACCUCCAUCAAUUAAAGUCCUGCA GGUACAACUGAUGCACCUGUAGCCCCGC UGUCCCUCCAG	9 -	138,821,429- 138,821,497	AK055547	on opposite strand of KIAA1984	intron/3 'SS	
2-163	AAAAGAUAUUCCUUUCUCUUGACAUUGC UUCCUUAAGAACUGANAACAGAAUGGAG GAACUGUCAAAUGG	9 -	21,468,514- 21,468,585	LOC554202		intron	
1-81b 1-111b	GGUAUGGCUGCCUUGGUUGGGAGUGGU GGGCCCUCCACCCUCCUUUCUCCCUCA	9 -	21,525,603- 21,525,656	LOC554202		intron	
1-40	UUUUACAUUAAUCUACUACACCAUUUUUU GUAUUUAGAUAUUC	9 -	30,281,947- 30,281,989			intergenic	
1-102	ACUUGCCUGUUUUGAAGUCAAAUAUUGU ACCAGUAAAUCUAG	9 -	33,122,712- 33,122,880	B4GALT1	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 1	**intron**	**CDS**
1-16d	UGUCCCUUUCUUGUUGCUUCAGUUGAGA CUGAGCUCUCCAACU	9 -	93,704,736- 93,704,778	ROR2	Receptor tyrosine kinase-like orphan receptor 2	**intron**	**CDS**
2-234a	UUUACUAAUUUCUCCUUAGUUUGGAUCA CUAACA	9 +	106,575,830- 106,575,863	NIPSNAP3B	Nipsnap homolog 3B (C. elegans)	**3' exon**	**3' UTR**
2-7b	AUUUCAGUCAUUUGAGUCCUAAGAUAUU	9 +	109,129,204- 109,129,231	RAD23B	RAD23 homolog B (S. cerevisiae)	**intron**	**CDS**
1-16b	CUCUCUUUUGGGAUUCCAUAGAGAAACU GACUGCACUC	9 +	130,359,449- 130,359,486	SPTAN1	Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)	intron	5' UTR
2-26	AAAUAACUAUCCUUUUCUCAAGACUGGCA GUCCAGGCAGGGCCCAGCCACAG	9 +	130,410,887- 130,410,938	SPTAN1	Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)	**intron/3' SS (36 nt)**	**CDS**
1-88	AGUGGAGCCUGUGGGGGCCGGGCUGGG CUGGGGCAGCUGCUGUGUUGGUG	9 +	133,274,791- 133,274,840	KIAA0515	KIAA0515	intron	
2-30a	AAAUGUAAUAUAAAACAUAGCUCAGUGGC CUACUGCACUACUACUACUG	9 +	35,219,651- 35,219,697	UNC13B	Unc-13 homolog B (C. elegans)	**intron**	**CDS**
2-206 2-259	AUACUAAUUUUUUUUUUCUCCUAGCGAC AUGUUGGCAAAAUCUGUCCUUACA	9 +	74,049,766- 74,049,818	GDA	Guanine deaminase	intron	CDS
2-69	UCUUGAAACCUGAUCUGUCAUUCCCCUU ACCCCCAAUUCUCUGUUUAGAUUCUCUU CCU	9 +	99,463,001- 99,463,059	NCBP1	Nuclear cap binding protein subunit 1, 80kDa	intron/3' SS	CDS
2-59	AUUUCAUAUGUGAUCUUGGUAUUUUGAG UCUAAAGCACAUUUUCUCUUUA	X -	119,463,249- 119,463,299	LAMP2	Lysosomal-associated membrane protein 2	**intron**	**CDS**
1-81c	UGUCUUUCUCUUUUAGAUAUUGUAUCAA UCUUUAGAAAAGGCAUAGUCUA	X -	152,944,538- 152,944,587	MECP2	Methyl CpG binding protein 2 (Rett syndrome)	3' exon	3' UTR
2-182 2-215	GGAAGAAGCUAGUGUUUCCUAUGACCUG AUUGUGCCUUUGGGGUUUUUCAUGCAUG G	X -	53,722,486- 53,722,543	HUWE1	HECT, UBA and WWE domain containing 1	intron	5' UTR
1-150	CAUUUAGUUUUACAAUUGUUAACCCCAUU CUUUAUAUACCA	X +	135,408,880- 135,408,920	HTATSF1	HIV-1 Tat specific factor 1	intron	CDS

2-98	AGGGCUCCAGCAUCCUCUGCUUCCCCCA CCACGUUCCCAUCACCCACCUCAUUGAU CCACUGACC	X +	153,324,264-153,324,328	GDI1	GDP dissociation inhibitor 1	**3' exon**	**3' UTR**
1-9b	UACAUGCUGAUAUGAGAAGCAGAAUGCA GCAGAACGUGGCCCACUGCAUGCUGGUG	X/Y -	2,417,720-2,417,773	ZEBD1 ---- DHRSX	zinc finger, BED-type containing 1 ---- Dehydrogenase/reductase (SDR family) X-linked	**3' exon** **----** **intron**	**CDS** **----** **CDS**

[1]For CLIP tags mapping to overlapping genes both annotations are given. Genes containing multiple CLIP tags are boxed.

[2]The location of CLIP tags in introns, exons, non-coding RNAs (ncRNAs), on the opposite strand of genes or in intergenic regions is given. CLIP tags were defined to locate at splice sites (ss) when the proximal end was within 40 nt of the splice junction. In cases where CLIP tags did not overlap with a splice site the distance (in nt) is given.

[3]The location within a gene is shown. CDS, coding sequence; START or STOP, overlap of the CLIP tag with a translation initation or termination codon, respectively; UTR, untranslated region. Locations in exons or introns associated with annotated alternative splicing events are indicated in bold.

Table S2: Weight matrix of the binding specificity of SF1

Position	A	C	G	U	Consensus	Information score
-5	0.272	0.338	0.137	0.252	N	0.045
-4	0.780	0.058	0.031	0.131	A	0.654
-3	0.174	0.693	0.039	0.094	C	0.478
-2	0.051	0.029	0.030	0.890	U	0.922
-1	0.502	0.072	0.249	0.177	N	0.198
BP	0.755	0.086	0.072	0.087	A	0.562
+1	0.148	0.601	0.099	0.152	C	0.283

Table S3: Sequence motifs that are enriched in intronic and exonic CLIP tags relative to shuffled sequences

Intronic tags (149 total)

Motif	Enrichment	Count-in-tags	Total-nt-in-tags	Count-in-background	Total-nt-in-background	Probability*
CCUG	2.43	59	6900	25857	7344000	1.000
UCCU	1.92	88	6900	48871	7344000	1.000
CUUG	1.85	63	6900	36291	7344000	0.982
UUCU	1.57	102	6900	68957	7344000	0.972
UGGG	2.40	32	6900	14206	7344000	0.959
CUGA	1.97	48	6900	25891	7344000	0.952
CUGC	1.97	47	6900	25406	7344000	0.940
GCUG	2.13	38	6900	18970	7344000	0.925
AAAA	1.97	45	6900	24367	7344000	0.910

Exonic tags (44 total)

Motif	Enrichment	Count-in-tags	Total-nt-in-tags	Count-in-background	Total-nt-in-background	Probability*
UCCU	2.48	29	2191	12390	2320000	0.982
CCUG	2.80	22	2191	8329	2320000	0.964

*This column shows the posterior probability that the frequency of the motif in the tags is not the same as the frequency of the motif in randomized sequences. Only enriched motifs for which the probability is higher than 0.9 are shown.

Table S4: Comparison of the frequency of all tetrameric motifs in exonic/intronic tags relative to random fragments of the same length taken from human exons/introns. The mean and standard deviation of the counts of each of the motifs in 100 randomized sets of exonic/intronic tags were used to compute the z-value.

Exonic tags

Tetramer	CountInTags	z-value	Tetramer	CountInTags	z-value	Tetramer	CountInTags	z-value
GGAG	3	-3.23	TCGC	1	-1.18	TGGT	8	-0.55
GAGG	3	-2.69	TGAA	10	-1.18	ACCG	3	-0.54
GGAC	2	-2.62	GCCA	9	-1.17	GGGT	4	-0.54
AGGA	6	-2.40	GTAC	2	-1.16	AGTC	5	-0.53
AAGT	2	-2.25	TACC	3	-1.16	CGGT	2	-0.51
AGCA	5	-2.20	GGTC	3	-1.15	CCTA	4	-0.50
GCCC	5	-2.01	TCCG	2	-1.14	CACC	10	-0.46
GACA	4	-1.99	GGCA	7	-1.13	CGAG	5	-0.45
CAGG	6	-1.96	TGCG	2	-1.11	ATGT	7	-0.44
CAAG	6	-1.92	TTGG	6	-1.10	ATGG	10	-0.43
GGAA	7	-1.90	AAGA	13	-1.03	AGAC	8	-0.42
GCCG	1	-1.89	CTCG	2	-1.01	AGGC	9	-0.41
AGAG	6	-1.88	GCGG	4	-1.00	ATGC	7	-0.40
GAAG	10	-1.86	GAAC	6	-0.97	GGTT	4	-0.39
CGGA	1	-1.83	GGCC	10	-0.96	GTGG	12	-0.38
CGCC	1	-1.82	AGCC	9	-0.96	CGCT	4	-0.38
AGTA	1	-1.81	GTGC	6	-0.96	CTGA	12	-0.35
GAGC	6	-1.81	TGTG	9	-0.94	GCAT	6	-0.35
AGGG	3	-1.81	CAGC	14	-0.92	CTGG	18	-0.33
CCGG	2	-1.76	CCAA	8	-0.89	GACT	7	-0.30
GCAC	3	-1.74	GACC	7	-0.88	GCGA	3	-0.30
GAAA	8	-1.67	GTCA	5	-0.86	AGAT	10	-0.29
GAGT	3	-1.62	AGAA	15	-0.86	GATG	12	-0.28
CCCG	2	-1.60	CCGT	2	-0.83	TGAG	10	-0.25
CCAG	11	-1.60	ACGA	2	-0.82	GTCT	6	-0.23
TGGA	12	-1.58	CCCA	10	-0.80	CAGT	10	-0.20
GCAG	10	-1.52	ACCT	8	-0.79	GTTC	6	-0.20
CCGC	2	-1.50	GGCG	4	-0.79	CGAT	2	-0.20
GTGA	5	-1.43	ACAC	5	-0.77	TACG	2	-0.18
AAGC	6	-1.41	ACGC	2	-0.71	CCTC	12	-0.18
CACA	5	-1.41	TTCG	2	-0.67	GGTG	10	-0.18
CGGC	3	-1.38	GGGG	7	-0.67	CTAT	5	-0.15
CGGG	3	-1.35	TACT	4	-0.67	CACG	4	-0.15
CATG	6	-1.32	GACG	3	-0.65	GTCC	7	-0.15
CGAC	1	-1.32	CAAC	7	-0.65	GGCT	11	-0.14
CGTC	1	-1.31	CCCC	9	-0.64	TACA	7	-0.14
CAGA	11	-1.30	CGTG	4	-0.63	CTGC	17	-0.12
GAGA	9	-1.28	TGGC	11	-0.63	AGTG	10	-0.12
GCAA	5	-1.25	TGCC	10	-0.63	TCGA	3	-0.11
GGGA	7	-1.25	GGGC	9	-0.59	AAGG	11	-0.11
ACCC	5	-1.24	CGAA	2	-0.58	ATAC	4	-0.08
CCGA	2	-1.24	CAAA	9	-0.58	ACGT	3	-0.05
CTAC	4	-1.23	AAAG	11	-0.56	GTAG	3	-0.00
ACAT	5	-1.21	ACGG	3	-0.56	GCGT	3	0.01
ACAG	8	-1.21	AGCG	4	-0.56	TGCA	12	0.03
AACG	1	-1.20	GCTG	17	-0.55	TAGA	5	0.12

Exonic tags (continued)

Tetramer	CountInTags	z-value	Tetramer	CountInTags	z-value	Tetramer	CountInTags	z-value
ACTC	7	0.13	TTTG	13	0.62	TAAG	7	1.65
TCAA	10	0.13	GGTA	4	0.63	CTTG	13	1.68
ACTG	11	0.14	TTGC	9	0.64	TAAA	12	1.74
AACC	8	0.16	GCTT	10	0.73	TCCA	18	1.79
ACCA	12	0.19	CTTA	6	0.74	GTAT	8	1.83
AACT	9	0.20	ATTT	12	0.83	TTCA	17	1.90
AAAA	17	0.21	AAAT	14	0.83	ATAG	7	1.92
TATC	5	0.22	GCTA	7	0.88	ATTC	12	1.92
ACTT	9	0.29	GCCT	15	0.88	AATC	11	2.03
TGTC	9	0.29	GCTC	13	0.90	CACT	14	2.06
AGGT	7	0.31	CATT	11	0.93	TTGT	13	2.10
TCTA	6	0.33	TATT	9	0.94	GGAT	15	2.14
ATAT	7	0.33	CTCT	14	0.94	CAAT	12	2.17
TGAC	10	0.33	ATAA	8	0.96	TCTG	20	2.23
CCCT	13	0.36	CATA	6	1.01	ATTA	11	2.27
GTAA	4	0.37	CTCC	17	1.02	GATC	13	2.31
TCAG	13	0.38	ACTA	7	1.10	AATT	15	2.44
TATA	5	0.38	ATTG	11	1.16	TAAC	8	2.67
AAAC	10	0.39	TAGG	4	1.17	TGTA	12	2.68
CTCA	12	0.40	TCTC	13	1.22	ATCT	16	2.79
TCAC	9	0.42	TATG	10	1.23	TGTT	17	2.84
TTAC	6	0.42	GATT	10	1.23	GTTT	15	2.93
ATGA	14	0.43	CATC	16	1.24	CTTC	24	2.97
GTTG	7	0.43	GTTA	6	1.25	TAGT	8	3.03
TGCT	15	0.45	TTTT	17	1.33	TCTT	20	3.15
GTGT	9	0.46	TTGA	13	1.34	CTAG	9	3.23
ATCA	11	0.50	TGAT	14	1.34	TTCT	23	3.57
CTAA	5	0.50	AATA	10	1.34	TTAG	10	3.63
AGTT	9	0.50	AGCT	18	1.37	TTTA	17	3.76
GATA	6	0.53	CCAC	16	1.40	CTTT	22	3.81
GAAT	10	0.53	CCAT	14	1.43	TAGC	10	3.86
CCTG	22	0.53	TTAT	10	1.44	TTCC	23	3.98
ACAA	12	0.54	CTGT	17	1.52	TTTC	21	4.09
TGGG	14	0.54	TAAT	8	1.60	TTAA	16	4.19
AATG	12	0.56	TCAT	14	1.61	TCCT	29	4.31
CGTT	3	0.62	TCCC	15	1.64	CCTT	25	4.96

Intronic tags

Tetramer	CountInTags	z-value	Tetramer	CountInTags	z-value	Tetramer	CountInTags	z-value
AGGG	11	-4.12	GGAG	23	-2.94	CACG	1	-2.40
GAGG	17	-3.51	GAGA	17	-2.90	GGCC	19	-2.36
TGAG	19	-3.50	GGGC	17	-2.86	GCGG	1	-2.35
GGGA	18	-3.49	GGTG	18	-2.83	AAGG	17	-2.31
GGGG	11	-3.40	AGCC	18	-2.70	GGCA	20	-2.22
CTGG	28	-3.30	AGGA	21	-2.65	GCCC	19	-2.21
GGGT	12	-3.26	GCAG	20	-2.65	GGCG	3	-2.20
CAGG	27	-3.04	GAGC	13	-2.56	GTAG	10	-2.19
AGGC	18	-2.97	GTGG	22	-2.56	GAAA	25	-2.16
TGGG	32	-2.97	AGAG	21	-2.55	TAGG	11	-2.12

Intronic tags (continued)

Tetramer	CountInTags	z-value	Tetramer	CountInTags	z-value	Tetramer	CountInTags	z-value
AAAG	24	-2.12	ACAC	15	-1.04	TTAA	38	-0.26
GGCT	25	-2.06	GCAC	16	-1.03	TGTG	42	-0.26
CGGG	5	-2.05	ACAA	19	-1.01	ACAT	27	-0.23
GGAC	10	-2.02	AGTT	25	-0.98	ACGT	5	-0.22
GGAT	14	-1.98	CAAA	28	-0.95	GAAT	26	-0.17
AGAA	28	-1.92	TGGC	29	-0.92	GTTG	23	-0.16
CAGC	25	-1.87	TTAT	33	-0.88	GATC	16	-0.15
AGAT	19	-1.87	GAGT	22	-0.87	GATA	17	-0.12
AAGA	23	-1.79	ATAG	17	-0.86	TTTA	51	-0.06
GTTA	11	-1.78	CAGT	27	-0.84	GCAA	20	-0.05
CGTG	4	-1.76	TATA	24	-0.83	CAAC	16	0.02
GCCG	3	-1.74	CGCA	3	-0.83	ATGG	30	0.08
CCAG	35	-1.72	GCCT	35	-0.83	CGAC	3	0.09
AAAA	45	-1.72	AACA	24	-0.82	CCGT	7	0.13
TGAA	26	-1.70	AGCG	4	-0.81	CTAG	19	0.14
GTGA	21	-1.70	GATG	21	-0.80	CCGA	6	0.15
TCGG	2	-1.67	ACGC	3	-0.80	CAAG	27	0.16
CGGC	3	-1.64	GTAT	17	-0.79	CGAG	8	0.19
ACCG	1	-1.63	TGCG	4	-0.78	ATTA	34	0.21
GACC	12	-1.63	AATA	31	-0.77	ATTT	63	0.23
GCGT	2	-1.61	ACGA	2	-0.77	CATC	24	0.25
GCGC	2	-1.58	GCTC	22	-0.76	AGCT	34	0.26
GGTC	13	-1.58	AGAC	20	-0.73	TACA	26	0.26
AGTG	26	-1.54	GCTG	38	-0.69	GGTT	26	0.28
CCGG	5	-1.53	GTGC	21	-0.69	GTCA	23	0.34
AAGC	15	-1.53	CCAC	27	-0.68	TTAG	28	0.35
ACGG	2	-1.52	CACA	27	-0.65	CCCC	41	0.36
GTGT	20	-1.50	CGTA	2	-0.64	ATAT	38	0.39
AGCA	22	-1.50	GATT	23	-0.64	AATG	34	0.40
ATCA	16	-1.48	GCTA	14	-0.63	TAAT	37	0.42
GAAG	22	-1.46	TAAG	19	-0.63	GTTT	42	0.43
CGCC	5	-1.45	TGGA	31	-0.63	TCAC	30	0.46
TTGA	25	-1.45	TAGT	19	-0.62	TGTT	46	0.50
TAGA	18	-1.41	ACAG	30	-0.57	CCTG	59	0.53
AGTA	17	-1.38	CCAA	22	-0.57	GTAC	13	0.54
GGTA	10	-1.38	TCAG	33	-0.53	AAGT	33	0.56
GACA	16	-1.37	CCCA	41	-0.53	TGCC	39	0.60
CCGC	4	-1.37	TAAA	39	-0.52	AACC	22	0.63
GTCG	1	-1.36	CGCT	5	-0.49	GCGA	6	0.64
GAAC	12	-1.35	ATGA	26	-0.47	CCCG	15	0.65
ATCG	1	-1.33	ACCC	23	-0.44	AACG	5	0.66
CTCG	4	-1.33	CGTT	4	-0.43	TTTT	121	0.67
CGGT	2	-1.28	GACG	4	-0.40	TGCA	36	0.67
AGGT	22	-1.28	AAAT	48	-0.40	TACG	4	0.71
GCCA	24	-1.26	TCCG	5	-0.33	ATGC	24	0.72
TATT	34	-1.21	TGGT	30	-0.32	GCAT	26	0.76
GGAA	26	-1.19	TCGC	4	-0.31	ATGT	37	0.76
CAGA	29	-1.19	CGAA	3	-0.30	TTCG	6	0.82
ATAA	25	-1.15	TCGT	4	-0.30	TTCA	42	0.88
CACC	24	-1.05	CCCT	39	-0.27	ACCA	30	0.98

Intronic tags (continued)

Tetramer	CountInTags	z-value	Tetramer	CountInTags	z-value	Tetramer	CountInTags	z-value
CTTA	29	1.00	CTGA	48	1.76	TCAA	37	2.45
GTAA	26	1.04	TTGG	45	1.78	CTAT	30	2.48
TTTG	58	1.16	CTGT	54	1.80	TTGC	39	2.49
ATTG	32	1.17	CATT	46	1.84	ATAC	28	2.49
GTTC	26	1.19	TCTA	32	1.86	ATTC	40	2.64
TGTA	37	1.23	ACCT	39	1.87	TGAT	44	2.67
CGTC	8	1.26	ACTG	40	1.90	ACTA	29	2.78
ACTC	29	1.29	CTCT	58	1.91	CTAC	27	2.85
CTGC	47	1.31	GTCT	39	1.95	CTCA	56	2.85
TTAC	28	1.37	CTCC	58	1.99	AGTC	34	2.93
TCTG	53	1.39	TAAC	25	2.00	TATC	30	3.11
CCTC	55	1.41	TCCC	54	2.03	CCTT	56	3.13
CTAA	29	1.42	CATA	31	2.04	CCAT	46	3.24
CACT	39	1.44	TTGT	53	2.17	ATCT	50	3.82
AAAC	35	1.44	TACT	34	2.17	TCTC	70	4.05
TCCA	39	1.46	CAAT	28	2.25	CTTT	81	4.24
TAGC	24	1.51	ACTT	44	2.28	CTTC	62	4.65
TATG	31	1.53	ATCC	32	2.31	TTCC	65	4.71
TGAC	30	1.56	TGTC	41	2.34	TTTC	86	5.12
TCAT	40	1.56	AATT	58	2.40	CTTG	63	5.13
GCTT	36	1.58	CCTA	29	2.40	TACC	37	5.69
CATG	40	1.59	AACT	38	2.42	TCCT	88	6.26
GACT	29	1.71	GTCC	32	2.42	TTCT	102	6.53
AATC	29	1.72	TGCT	51	2.45	TCTT	95	6.82

Table S5: Sequences of RNA linkers, PCR primers and siRNAs

Name	Sequence[1]
RNA linkers	
RL5	OH-AGGGAGGACGAUGCGG-OH
RL3	P-GUGUCAGUCACUUCCAGCGG-puromycin
DP5	AGGGAGGACGATGCGG
DP3	CCGCTGGAAGTGACTGACAC
DP5EcoRI	GAGCCAACAGGCACCGAATTCAGGGAGGACGATGCGG
DP3EcoRI	GACTAGCTTGGTGCCGAATTCCCGCTGGAAGTGACTGACAC
PCR primers	
ALG8-E2F	CCGAAACTGGCTTGCTAT
ALG8-E6R	TCGCAGCAGATATACACC
FGFR1OP-E1F	GGTCCTGAACCGCATCAA
FGFR1OP-E9R	TGCATCCGAGAGCGAGGC
PDE3A-E1F	GCTTGGCCGCCGCGACAT
PDE3A-E4R	TGCTGGTGCTGCGGTCTC
PLOD2-E11F	GGACTTTTGCCGTCAGGA
PLOD2-E17R	CCAGACCATTTGCCGTAA
TNIP1-E1F	TCGTATTCTCCTGCCCTTCC
TNIP1-E6R	GCTGCTCTCTGGTGAATTCT
UPF3A-E1F	AGAAGCTGTCGGCCCTAGAA
UPF3A-E5R	CAGAGTCTCAGGGTTGGCAC
siRNAs[2]	
SF1 #1	GACCUGACUCGUAAACUGCdTdT (nts 178–196 of the coding sequence)
SF1 #2	UGGACUUACUCGAGAACAAUUdTdT (nts 126-144)
SF3a120	GGAGGAUUCUGCACCUUCUdTdT (nts 89–107)
LUC	CGUACGCGGAAUACUUCGAdTdT (=GL2 of Elbashir et al., Genes Dev. 2002)

[1]All sequences are shown in the 5' to 3' direction.
[2]Only the sense strand is shown.