Disease vocabularies in the dialog between medicine and biology

Mottaz, Anais

UNIVERSITÉ DE GENÈVE

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

Section de médecine fondamentale
Département de science des protéines humaines

Thèse effectuée sous la direction du Professeur Amos Bairoch, directeur de thèse, du Professeur Antoine Geissbühler, co-directeur de thèse, et du Docteur Anne-Lise Veuthey, superviseur de thèse

---

Disease vocabularies in the dialog between medicine and biology

THESE
présentée à la Faculté de Médecine
de l'Université de Genève

pour obtenir le grade de Docteur en sciences médicales MD - PhD

par

Anaïs MOTTAZ

de Genève (GE)

Thèse n°18

Genève
2015

UNIVERSITÉ
DE GENÈVE
FACULTÉ DE MÉDECINE

DOCTORAT EN SCIENCES MEDICALES
« MD-PhD »

Thèse de :

Anaïs MOTTAZ
originaire de Genève (GE)

Intitulée :

Disease vacabularies in the dialog between medicine and biology

La Faculté de médecine, sur le préavis du Comité directeur du MD-PhD, autorise l'impression de la présente thèse, sans prétendre par là émettre d'opinion sur les propositions qui y sont énoncées.

Genève, 5 janvier 2015

Thèse n° **18**

Henri Bounameaux

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

Remerciements

Le travail présenté dans cette thèse a été effectué principalement de septembre 2006 à décembre 2010.

Je tiens à remercier mes directeurs de thèse, le Professeur Amos Bairoch (département de science des protéines humaines, Université de Genève) qui m'a accueillie au sein de ses deux groupes de bioinformatique Swiss-Prot et CALIPHO, et le Professeur Antoine Geissbühler (département de radiologie et informatique médicale, Université de Genève) qui m'a accueillie aux réunions de son groupe de recherche en cybersanté et permis de me former dans le domaine de l'informatique médicale.

Je tiens à remercier également les Professeurs Yves Moreau (département de génie électrique, KU Leuven), Dominique Müller (département de neurosciences fondamentales, Université de Genève) et Patrick Ruch (département d'information documentaire, Haute école de gestion de Genève) qui m'ont fait l'honneur d'avoir accepté de faire partie du jury de thèse.

Je remercie chaleureusement Anne-Lise Veuthey pour sa supervision, sa disponibilité et son soutien dans toutes les circonstances.

Je souhaite aussi remercier Yum Lina Yip pour ses conseils et sa supervision relatifs au développement de SwissVar, ainsi que Fabrice David, Gregory Loichot, Harris Procopiou et Nathalie Lachenal pour leur travail sur les variants.

Mes remerciements s'adressent également à Livia Famiglietti et Arnaud Gos pour nos discussions sur l'annotation médicale de Swis-Prot, Paula Duek Roggli pour notre collaboration sur les ontologies d'anatomie, Julien Gobeill pour son travail préliminaire sur le mapping, ainsi qu'à Séverine Duvaud, Monica Pozzato, Delphine Baratin, Thomas Kappler, Elisabeth Gasteiger, Nicole Redaschi, Salvo Paesano, Karin Sonesson, Mikael Doche et Edouard de Castro pour leur disponibilité, leurs conseils et leur compétence dans le développement et le maintien des resources informatiques utilisées dans ce travail et le maintien des outils développés.

Ma gratitude va au Fonds national suisse de le recherche scientifique qui a subventionné ce travail.

Je souhaite encore remercier tous mes collègues de Swiss-Prot et CALIPHO pour les bons moments passés ensemble et également mes collègues de l'hôpital Beau-Séjour pour leur excellent accueil et particulièrement Adrian Guggisberg de m'avoir fait confiance.

Pour finir je tiens à remercier ma famille, particulièrement mes parents, mes frère et sœur, Edouard de Castro et notre fille, ainsi que mes amis de leur soutien sans faille dans toutes les décisions que j'ai prises pendant ce travail.

Abstract

Mendelian disorders, which account for most so-called “rare-diseases”, have an impact individually on a relatively small number of people but have a huge impact altogether and can greatly contribute to our understanding of disease molecular basis and cell biology. Nearly 4,000 of them have a known causative mutation, most of which have an effect on protein function through a single amino-acid change. They provide thus a direct link between a change in the DNA sequence and observable consequences on the development and functioning of the human organism through their impact on the molecular function of proteins. This unique perspective on the relationship between genotype and phenotype is highly valuable for the dialog between clinical practice and fundamental research.

In the last couple of years next generation sequencing technologies have begun to produce a huge flood of data. To cope with this “data deluge” efficient software tools are necessary to access and integrate these data and require a high degree of interoperability between the various molecular and medical knowledge resources. One of the first steps toward semantic interoperability and the representation of data into machine-processable formats consists in linking existing information to defined concepts represented in controlled vocabularies and ontologies.

The first purpose of the presented work was to find an automatic way to map the human proteins and variants that are causative of diseases annotated in the UniProtKB/Swiss-Prot knowledge base to a disease controlled vocabulary. The aim was to enhance the interoperability of Swiss-Prot with other sources of information relevant to these disorders.

The result of this mapping, updated every month, was made available to the community through the development of a web interface, SwissVar, improving the access from diseases to this major molecular biological resource. Requests could also be combined to sequence and three-dimensional characteristics of missense variants.

Besides, reviewing translational efforts in the domain of genomics revealed that much is done to predict new variants for implication in diseases using protein functional information, based on the correlation between protein function and phenotype. Less is done using disease information to prioritize protein functional information such as implication in biological process or protein/protein interactions (PPIs). In addition, distinct clinical traits including pathologies found in different Mendelian disorders are separately important. Indeed, recent evidences indicate that pleiotropy, the effect of single genes on multiple phenotypic traits, is mainly the consequence of proteins implicated in different biological processes depending on the context, in relation to modularity of cell biology. Yet these traits are only roughly represented through the disease categories of controlled vocabularies. Based on these observations, a prototype tool was developed to filter PPIs including those obtained through high-throughput technologies with Mendelian disorder phenotypes from the Human Phenotype Ontology (HPO), to

isolate biological process context. The aim of this approach was to help formulate hypotheses on the function of proteins and interactions and had never been proposed in such comprehensive manner.

Finding automatic ways to link fundamental research information to disease concepts is an essential step toward a better dialog with clinical medicine. Mendelian diseases are highly valuable as they provide a direct link between molecular data and phenotypes. Given the modular nature of cell biology, considering clinical traits separately is necessary to make the most of this relationship.

Résumé

Les maladies mendéliennes, représentant la plupart des 'maladies rares', touchent individuellement un nombre relativement faible de personnes mais ont un impact global important et se révèlent précieuses pour la compréhension de la physiopathologie des maladies ainsi que de la biologie cellulaire. Environ 4'000 d'entre elles ont une mutation causale connue dont la majorité est un changement simple d'acide aminé. Elles représentent donc un lien direct entre un changement unique dans la séquence d'ADN et ses conséquences visibles sur le développement et le fonctionnement de l'organisme humain à travers son impact sur la fonction moléculaire des protéines. Cette perspective sur le lien entre le génotype et le phénotype est importante pour le dialogue entre la pratique clinique et la recherche fondamentale.
Présentes depuis quelques années, les nouvelles technologies de séquençage ont et vont produire une quantité gigantesque de données. Pour accéder à et intégrer ces données de manière efficace, des outils logiciels sont indespensables et nécessitent un degré élevé d'interopérabilité entre les différentes resources médicales et biologiques. Une des premières étapes vers l'interopérabilité sémantique et la représentation des données dans un format lisible en machine consiste à lier les informations existantes à des concepts prédéfinis tels qu'on trouve dans les vocabulaires contrôlés et les ontologies.

Le premier objectif de ce travail a été de développer une méthode automatique pour lier à un vocabulaire médical contrôlé les protéines et variants humains causant des maladies annotées dans la base de connaissance UniProtKB/Swiss-Prot. Le but était d'augmenter l'interopérabilité de Swiss-Prot avec d'autres sources d'information concernant ces maladies. Le résultat de ce mapping, mis à jour chaque mois, a été rendu public à travers une interface web, SwissVar, améliorant ainsi l'accès à partir des maladies à cette ressource majeure de biologie moléculaire. Les requêtes peuvent également être combinées à des caractéristiques séquentielles et tridimensionnelles des variants.

Ensuite, la considération des efforts translationnels dans le domaine de la génomique a révélé que beaucoup de travaux utilisent la relation prédictive qu'il existe entre la fonction des protéines et les phénotypes pour détecter de nouveaux variants potentiellement impliqués dans des maladies. Beaucoup moins se concentrent sur l'utilisation des maladies pour déceler des implications de protéines dans des processus biologiques ou révéler des interactions protéine/protéine (IPPs). De plus, les traits cliniques, incluant les pathologies, observés dans les maladies mendéliennes sont individuellement importants puisque des études récentes indiquent que la pléiotropie, autrement dit l'effet d'un gène sur plusieurs phénotypes, est principalement la conséquence de protéines impliquées dans différents processus biologiques suivant le contexte spatio-temporel dû à la nature modulaire de la biologie cellulaire. Ces différents traits cliniques ne sont que grossièrement représentés dans les catégories des vocabulaires médicaux. Pour explorer néanmoins le potentiel de ce concept, un outil prototype a été développé pour filtrer, ou contextualiser, les IPPs avec des phénotypes de maladies mendéliennes trouvés

dans l'ontologie HPO (Human Phenotype Ontology). Cette approche a pour but d'aider à formuler des hypothèses sur la fonction des protéines et des interactions et n'a jamais été proposée de façon généralisable comme ici.

Développer des moyens automatiques de lier des informations de recherche fondamentale à des concepts de maladie est une étape essentielle pour l'amélioration du dialogue avec la médecine clinique. Les maladies mendéliennes ont une grande valeur puisqu'elles représentent un lien direct entre les données moléculaires et les phénotypes. Etant donné la nature modulaire de la biologie cellulaire, considérer individuellement leurs différentes caractéristiques cliniques est indispensable pour exploiter au mieux cette relation.

1. Introduction

1.1 Molecular and clinical data integration

Biomedical data growth

The last decades have seen a change in the scale of data production and storage, including biomedical data, enabled by technological progresses. The corpus of scientific publications reporting the results of clinical and fundamental research has grown exponentially, from hundreds of thousands to tens of millions in 50 years (Figure 1).

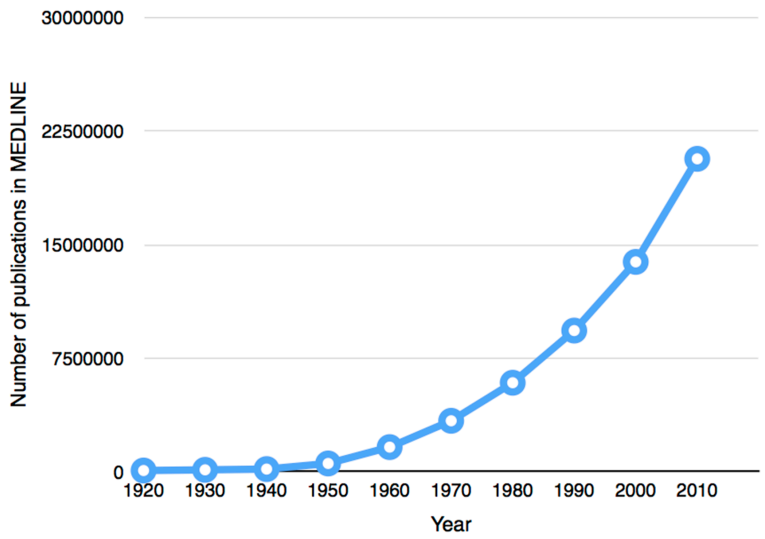


Figure 1. Exponential increase of publications as captured in PubMed.

In the genomic domain in particular, the development of personalized whole-genome sequencing, enabled by next generation sequencing technologies, will contribute to the creation of incredible amounts of data. The sequences will come along with other clinical information such as diseases and phenotypes, critical to fully exploit these data (Cordero & Ashley, 2012). Finding ways to store, organize, share, retrieve, integrate and analyze them is extremely challenging and requires the combined efforts of many research fields.

More generally, the development of semantic web technologies is illustrative of such kind of efforts. Their aim is the representation of data with formally defined languages allowing machines to treat more easily the semantic

content of web pages (Berners-Lee *et al.*, 2001). Indeed, while humans can easily understand which concepts are treated in unstructured data such as text using contextual information combined with previous knowledge, this is much more difficult for automatic approaches. These technologies are based on standards ensuring technical and semantic interoperability, controlled vocabularies and ontologies being the most important resources for semantic interoperability.

Semantic interoperability

In the biological and medical domains, standardization efforts have begun well before the emergence of the semantic web concept. In the medical domain for example, the International Classification of Diseases (ICD) was created in the 19th century to classify death causes in different countries. Things took longer in the biological domain. The Gene Ontology was created only in 1998 to help researchers standardize the representation of genes and gene products attributes across species and databases (Consortium, 2006).

A crucial requirement to achieve a seamless integration of biomedical data is the interoperability between clinical resources and fundamental research, especially around pathologies (Machado *et al.*, 2013). Indeed disease concepts are essential for clinical practice but are also important in fundamental life sciences research. Clinical practice use disease concepts to rationalize medical care and treatment. In the life sciences cellular and animal models are used to understand the molecular basis of diseases. Also, physiological function of genes and proteins are investigated through pathological phenotypic effects of molecular product deficit, using for instance gene knock-out or knock-down approaches. Therefore on one hand, molecular information relative to a disease is a key resource for the development of diagnostic tools and treatments. On the other hand, the availability of clinical findings can give ideas and directions for studying mechanisms of pathology and better understand the physiological functions of molecular products (Figure 2).

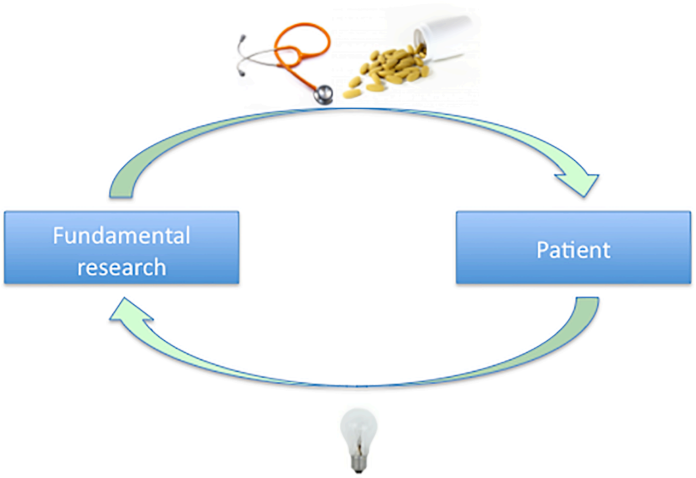


Figure 2. Illustration of the global dialog between biology and medicine: fundamental research discoveries translating into treatment, prevention and diagnostic tools; patient's information representing valuable information to better understand pathologies.

1.2 Characteristics and purposes of controlled vocabularies

Controlled vocabularies are sets of predefined terms used to identify concepts in a domain. It can go from a simple list of terms to more elaborate representation of the vocabulary with concept definition and synonyms as well as relations between concepts. As most controlled vocabularies are now organized into taxonomic hierarchies, they are often assimilated to ontologies (Bodenreider, 2008). Indeed, ontologies aim at representing knowledge by categorizing and relating things in a formal way. Ontologies gathering several domains together through different kinds of relations enable yet more elaborate automatic reasoning.

Concepts

Definition

The presence of definitions enables to clear ambiguities. This is particularly the case with homonyms. For example when dealing with the Charcot disease, it is essential to know if it refers to amyotrophic lateral sclerosis (ALS), commonly referred as Charcot disease by French people, to Charcot–Marie–Tooth disease or to a neuropathic arthropathy, known as Charcot joint. These diseases are indeed different, the ALS being a complex disease implicating the degeneration of upper and lower motor neurons, the Charcot-Marie-Tooth a Mendelian

disease responsible for a peripheral neuropathy and the neuropathic arthropathy a degeneration of joints following peripheral neuropathy.
The use of concepts helps to integrate data from different resources. One way is through the use of a similar ontology. The second way is through the use of different ontologies that first have to be aligned.
Medical decision support systems for example beneficiate from integration of information through controlled vocabularies, for example between electronic records and knowledge resource to warn for drug adverse effects or interactions (Greenes, 2011).

Synonyms

By providing synonyms, through lexical relations, ontologies enable to consider and treat ideas rather than terms. Gathering synonyms around concepts facilitates the retrieval and integration of documents relative to the same concepts even if they use different denominations. For example when looking for information on the Rubinstein-Taybi syndrome, it is useful to also retrieve documents mentioning the broad thumb-hallux syndrome because they refer to the same disease.

Taxonomic relations

Controlled vocabularies are often composed of concepts with different levels of specificity, organized in a hierarchy through taxonomic relations, or subsumption links (is-a), and sometimes partonomic relations, or composition links (part-of). As the links have a direction, and a parent cannot be its own child, they are structured as directed acyclic graphs with most of the time a single root.
Concepts can then be retrieved through flexible entry points, with slightly different levels of specificity. For example when searching for information about dwarfism, information about achondroplasia can be retrieved.
Taxonomic relations enable also to work with categories of concepts, for example to study data about cardiovascular diseases.
Moreover, they provide a mechanism to estimate the semantic similarity between concepts.

Data retrieval and aggregation

Clinical vocabularies, such as the International Classification of Diseases, ICD (www.who.int/classifications/icd/en/), enable the aggregation of diseases into categories to study variables such as survival rate (Bergeron *et al.*, 2007). They are also used for world-wide comparison of morbidity and mortality rate or to help estimate for example costs of health care for hospital billing.
Literature indexing vocabularies improve the retrieval efficiency of relevant documents. For example the major database of biomedical literature, MEDLINE, accessed through PubMed (www.ncbi.nlm.nih.gov/pubmed) indexes its articles with the Medical Subject Headings (MeSH) (www.nlm.nih.gov/mesh), to improve the retrieval of documents among 23 millions of citations.

Semantic similarity measures

Taxonomic hierarchies are used to estimate semantic similarity between concepts. Intuitively, the closer two concepts are in the hierarchy, the closer their meaning are.

Two main approaches are used: the first approach consists in using path length between concepts and the other in using information content of concepts (Blanchard *et al.*, 2005). The first approach is based on the idea that the more concepts separate two concepts, the less similar they are. The second approach is based on the idea that the more two concepts share information, the more similar they are. Evaluation of semantic similarity measures is difficult because it depends on what level similarity is interpreted. Human judgment can be used as well as other parameters known to correlate with similarity (Pesquita *et al.*, 2009).

The similarity between protein function estimated using the Gene Ontology (Consortium, 2006) has been used to find functional modules, to predict protein/protein interactions (PPIs) and implication in diseases (Wang *et al.*, 2010) and to transfer information between different species (Blake & Bult, 2006).

1.3 Medical controlled vocabularies

SNOMED-CT

The most important clinical vocabulary is the SNOMED-CT. SNOMED-CT was born from the union of an American systematized nomenclature (SNOMED) developed by pathologists, dealing with precise diagnostics, and British clinical terms (CT) more oriented toward primary care practice. Its scope encompasses diseases but also clinical findings, procedures, anatomy, pathogenic biological agents, substances, social context, etc. There are over 300,000 concepts organized in a taxonomic directed acyclic graph hierarchy, enabling several parents, with additional relations including causative and locative. Concepts are formally defined. It is mainly intended for use in electronic health records and is maintained by the International Health Terminology Standards Development Organization (www.ihtsdo.org/snomed-ct).

ICD10

ICD is the International Classification of Diseases produced by the World Health Organization. This classification was created in the 19th century to classify death causes. It spread rapidly in several countries. It expanded later, in 1949, to morbidity and was then primarily used as an epidemiological tool to register and compare international statistics of mortality and morbidity causes. As hospitals began to index medical records with it, more detailed disease information was needed with precise manifestation beside etiology. ICD was then extended to contain signs, symptoms, social circumstances and the possibility to add a manifestation site to an etiology that was implemented through a system of principal and accessory code (called dagger and asterix).

Indeed, codes are organized in a mono-hierarchical classification, they can have only one parent and a unique code represents the hierarchical position of the concept. For each code are provided a text definition, synonyms named inclusions as well as exclusion terms to indicate what it is not.

It is worthwhile to note that countries produced national modified versions of the published ICD to respond to their need, particularly the United States with the ICD-9 CM (Clinical Modification), developed by Centers for Disease Control and Prevention, for more detailed morbidity. This delayed their adoption of ICD-10 that is now under the way with the ICD-10 CM, while ICD-11 is due to be released in 2015. ICD-11 is intended to have a more meaningful structure, with disease entity associated to properties such as definition, manifestation site or duration enabling more semantic operations and facilitating ontology mapping for example with SNOMED-CT.

MeSH

Medical and biological publications represent a very important source of biomedical information. The major database of life sciences and biomedical literature is MEDLINE, accessible through the PubMed search engine maintained by the United States National Library of Medicine (NLM). The NLM has developed a terminology to index articles and improve retrieval efficiency, the Medical Subject Headings (MeSH) (www.nlm.nih.gov/mesh). It is composed of descriptors organized in a directed acyclic graph, enabling several parents, with taxonomic relations. Descriptors can contain several concepts, and each concept is itself composed of several synonyms, or terms. The descriptor name corresponds to a preferred concept, while the concept name corresponds to a preferred term. Some concepts are slightly narrower concepts than their descriptor, but not enough to form a separated descriptor. In the 2014 version, MeSH contained 27,149 descriptors and more than 218,000 terms. Also 83 qualifiers can add a context to the descriptors, such as ‘congenital’ or ‘prevention’. The essential benefit of this vocabulary is that these terms are directly linked to the most basic source of information - the scientific literature.

A summary of the different characteristics of these three main vocabularies existing at the time of this work is presented on Table 2.

Table 2. Properties of main medical vocabularies.

	SNOMED-CT	ICD-10	MeSH
Number of concepts	~300,000	~14,400	~27,000
Relationships	'Is a', multiple parents allowed 'Attribute relationship' e.g. finding site, causative agent	'Is a', one parent allowed Dagger/Asterix system to add an anatomical site to an etiology	'Is a', multiple parents allowed, possible combination of site and etiology
Coverage	Clinical findings/disorders Procedures Observable entities Anatomy, morphology Chemicals names, generic drug products Generic physical devices Other etiologies of disease, including external forces, harmful events, accidents, genetic abnormalities Functions and activities Social contexts care provision Types of clinical records Staging, scales, classifications	Diseases	Anatomy Organisms Diseases Chemicals and Drugs Analytical, Diagnostic and Therapeutic Techniques and Equipment Psychiatry and Psychology Phenomena and Processes Disciplines and Occupations Anthropology, Education, Sociology and Social Phenomena Technology, Industry, Agriculture Humanities Information Science Named Groups Health Care Publication Characteristics Geographicals
Access	License	License, free for non-commercial use	No license (Terms and conditions)
Language	English (US + UK), Spanish, Danish and Swedish	42	20

Disease Ontology

The Disease Ontology (disease-ontology.org/) has been developed by the Northwestern University, Center for Genetic Medicine and the University of Maryland School of Medicine, Institute for Genome Sciences as a human disease ontology, containing 8,043 diseases classified anatomically and etiologically (Schriml *et al.*, 2012). Terms have been mapped to MeSH, ICD, NCI thesaurus, SNOMED and OMIM. It aims at providing consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts.

UMLS

The major effort to map all biomedical ontologies together is the Unified Medical Language System (UMLS) (Bodenreider, 2004). It contains nearly 3,000,000 concepts (2013 official statistics) with several kinds of relations coming from more than hundred different source vocabularies. Such effort enables the integration of data from sources using different vocabularies. Moreover, a semantic network has been created to navigate across concept categories through semantic relations, such as functionally related or spatially related. The major drawback of this kind of resources is its heaviness of use.

OntoOrpha

Orphanet is the main Mendelian diseases resource for healthcare professionals and patients led by a consortium of around 40 countries, coordinated by the French INSERM team. It provides information about many aspects of the diseases including an inventory of rare diseases with clinical descriptions. The disease descriptions have been mapped to OMIM, MeSH and ICD10. OntoOrpha has been created recently and should be worth exploring since it is an ontological representation of Orphanet knowledge that implements relations between diseases, clinical signs and genes (Olry *et al.*, 2011)

NCI thesaurus

The NCI thesaurus is a terminology and biomedical ontology around cancer containing 10,000 diseases but also substances, therapies and genes (Sioutos *et al.*, 2007). A metathesaurus has also been created with 2,000,000 concepts from terminologies mainly found in UMLS.

Human Phenotype Ontology

The Human Phenotype Ontology (HPO) was originally constructed using the Clinical Synopses from OMIM (Köhler *et al.*, 2014), a main Mendelian disease resource with a more molecular orientation than Orphanet (for OMIM description see *Mapping UniProtKB/Swiss-Prot to a disease controlled vocabulary* chapter, section *Resources description and data extraction*). The Clinical Synopses of OMIM are manually annotated clinical features found in Mendelian diseases (Hamosh *et al.*, 2005). However the vocabulary used in the clinical synopses is not normalized, meaning that the same clinical feature can be expressed in different ways. For example, ‘generalized amyotrophy’ can also be expressed as ‘generalized muscular atrophy’ or as ‘muscular atrophy’ depending on the entry. Also the granularity is not controlled, with ‘congenital heart disease’ being used in some entries while more precise terms like ‘ventricular septal defect’ are used in others.

To create HPO, synonyms were thus merged and semantic links were created between concepts to create the ontological structure, which has been manually refined, corrected, and expanded with definitions and new concepts (Robinson & Mundlos, 2010). The hierarchy is implemented as a directed acyclic graph with taxonomic links ('is-a').

HPO at this time contains over 9,500 terms organized in three ontologies, 'Organ abnormality', 'Inheritance' and 'Onset and Clinical course'. 'Organ abnormality' is the main ontology. It contains concepts as varied as 'Hypopigmented skin patches', 'Neurological speech impairment' or 'Basal cell carcinoma'. The 'Inheritance' ontology contains concepts related to the mode of inheritance of Mendelian diseases such as dominance and recessivity, as well as some concepts like somatic mutation or predisposition. Finally the 'Onset and Clinical course' contains concepts relative to the severity of the phenotype, like the age of onset or death and the pace of progression.

1.4 DNA variation and diseases

The molecular and clinical data integration efforts presented in this work concern protein variations related to diseases, in particular single amino-acid variants which result from missense variants. Understanding how genes and proteins, their main functional product, variations affect our health is currently one of the major challenges in the biomedical domain. Indeed, most of the diseases that can be cured or prevented today have an external agent as main cause. Antibiotics along with increased hygiene, sanitary rules and vaccines have decreased mortality due to infectious agents during the 20th century (Omran, 1971). Degenerative and chronic diseases are now more visible and a greater cause of morbidity and mortality (Figure 3) (Doll, 1995). The etiology of these diseases involves subtle interactions between genes and environment that remain to be elucidated.

Table 1. Main mechanisms of mutation.

Base substitutions	Structural variations
- Spontaneous loss or modification of a base. - UV rays creating cross-links between adjacent bases. - Chemicals such as tobacco smoke agents adding alkyl groups to DNA bases. - Replication errors on undamaged DNA. - Intercalating agents.	- Gamma and X-rays breaking the DNA backbone. - Transposable elements such as Alu sequences (Batzer & Deininger, 2002) - Replication error on undamaged DNA (Hastings *et al.*, 2009)

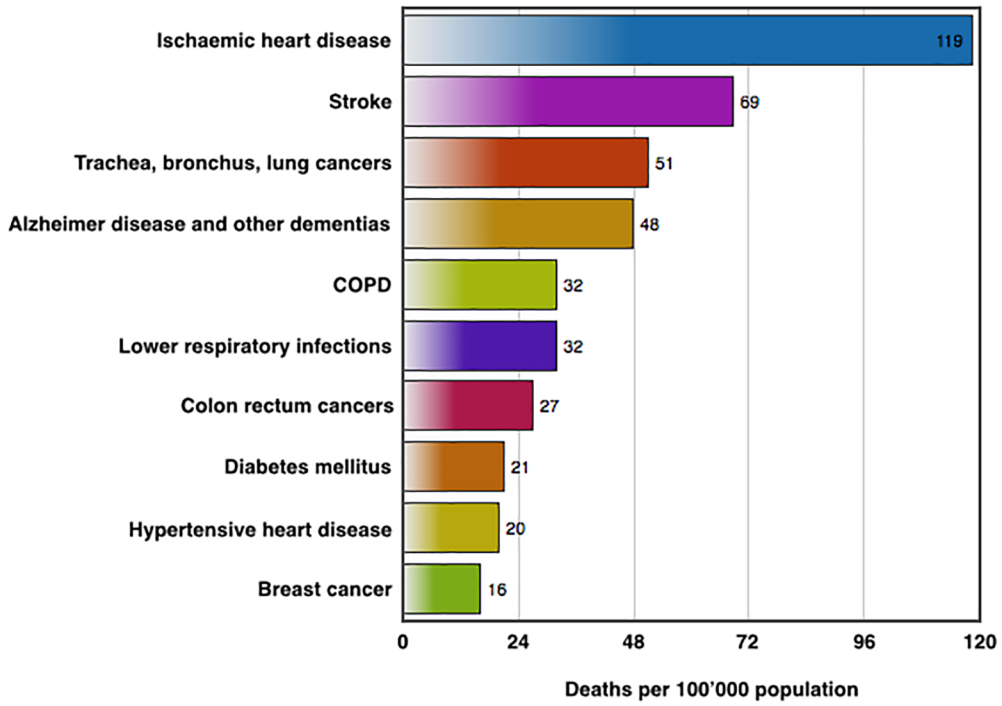


Figure 3. Main causes of mortality in high income countries (WHO).

Origin

DNA sequence variability is inherent to life adaptation in changing environments and evolution (Friedberg, 2003). Different types of variation exist: they can be subdivided in two main broad categories, base substitutions and structural variations. Base substitutions do not change the number of nucleotides but replace a base with another one while structural variations imply a change in the number of nucleotides or their order, such as insertions, deletions, duplications, inversions or translocations.

Variation can arise during replication on intact DNA or following DNA damage (Table 1). DNA damages happen thousands times a day in a given nucleated cell (Strachan & Read, 2011) and if unrepaired before replication a damaged base can lead to a variation.

Variations do not arise uniformly on the genome but some regions are more sensitive than others to different types of variations. For example sequences rich in methylated cytosine, CpG, are base substitution hotspots because deamination of a 5-methyl cytosine give directly rise to a thymine. Many other mechanisms behind region specificity to variations are under investigation.

While variations in somatic cells can lead to cancer if they are enabled to accumulate, they give rise in germ cells to a constitutive change in offspring creating a new variant in the population. It is estimated that around 60 new variations appear each generation (Conrad *et al.*, 2012). Base substitution variants are known as single nucleotide polymorphisms or SNPs. They are the most frequent variations, for example 38 millions of SNPs have been recently identified, including rare ones (Abecasis *et al.*, 2012) and when comparing two random genomes, around one of 1'000 base pairs are different. Investigation of copy number variants (CNVs), a structural variation implicating the duplication of large regions of the genome, have also revealed their importance in the human genome diversity (Redon *et al*., 2006; Conrad *et al.*, 2010), however SNPs are still the most frequent variations in term of numbers.

Discovery

The discovery that some entity is transmitted to offspring and results in a phenotypic characteristic has been described well before the identification of DNA. In the 19th century, Gregor Mendel described inheritance patterns of visible phenotypic traits in peas. Whereas at that time it was thought that traits from both parents blended together, he described the concepts of dominance and recessivity. These concepts implied an interaction between inherited factors, later called alleles. Some of the alleles need to be inherited from both parents to show an effect while others need to be given by one parent only. The former inheritance imply a "recessive" interaction between the alleles and the second a "dominant" one.

Diseases having a pattern of inheritance that can be described as dominant or recessive are called Mendelian diseases. They correspond to diseases mainly determined by a single variation, usually rare. Affected members can be easily recognized in families because the penetrance is high.
Disease-associated genes are identified by studying families in which the disease run. The identification of the causative genes begun long before the whole human genome was sequenced. The first one to be identified was a mutation in the hemoglobin responsible for sickle cell anemia (Ingram, 1956). Linkage analysis detecting chromosomal regions where the disease gene is susceptible to lie is performed with markers detection, the closer a marker is from the disease causing mutation the more chances it has to be transmitted with the disease. Up to 5 years ago identification of the causal gene was done through positional cloning, facilitated by available chromosome gene maps. This kind of approaches enabled the discovery of most of the monogenic diseases. Now the next-generation sequencing is revolutionizing the field (Koboldt *et al.*, 2013) by offering whole exome and genome sequencing at very low cost. Using these techniques, the identification of Mendelian disease genes is successful in the majority of cases (Gilissen *et al.*, 2012).

When the development of disease depends on several genes and environmental factors, pattern of inheritance cannot be easily described and diseases are referred as complex. Identification of complex diseases causal genes is much more difficult than for Mendelian diseases (Figure 4). Indeed they often have a low penetrance and several genes involved. Linkage analysis has been used for complex diseases. However when pattern of

inheritance is too far from Mendelian, non-parametric thus less powerful methods have to be used. Variants have nevertheless been identified with such methods, for example the apolipoprotein E*4 allele increase the risk for late onset Alzheimer (Pericak-Vance *et al.*, 1991).

Two hypotheses exist about the relationship between sequence variants and complex diseases. The first one is the common disease - common variant hypothesis and is based on the idea that several gene variations already present in the population slightly modify the risk of a disease. The second one suggests that complex diseases are caused by recent and rare variations with more effect and is called the mutation - selection hypothesis. These two hypotheses lead to different approaches for discovering associated variants.

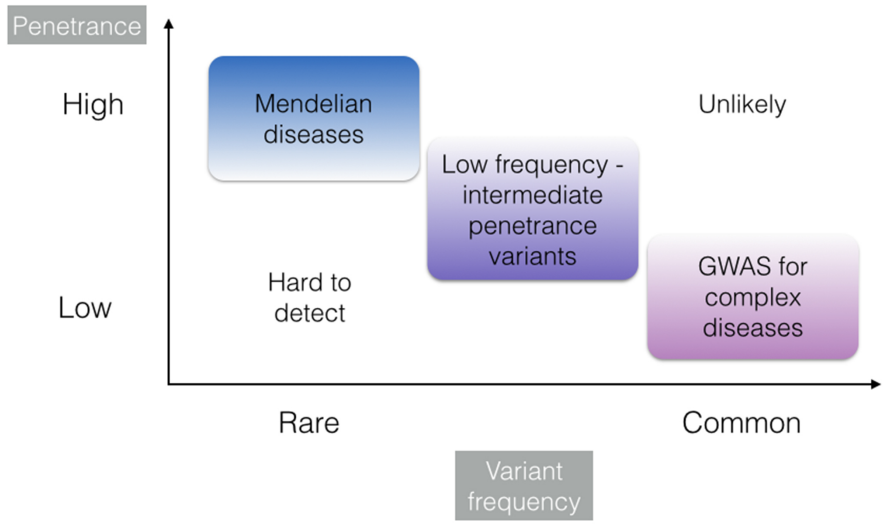


Figure 4. Detection methods of disease causing variants according to variant frequency and penetrance (inspired from Lobo, 2008).

Genome wide association studies (GWAS) investigate populations and are used to identify common variants associated to complex diseases (Stranger *et al.*, 2011). It is based on the hypothesis that susceptibility alleles are more often present in people with the disease than people without. They were enabled by the Human Genome project, a NIH project formally begun in 1990, that resulted in 2003 in the completion of the human genome sequence based on a small number of individuals (www.genome.gov/10001772) as well as the HapMap project that sequenced several individuals from different parts of the world in order to identify all SNPs present in more than 1% of the populations (hapmap.ncbi.nlm.nih.gov). Segments have been determined as regions of linked variants, long of a few thousands bases. These segments can be identified by tag SNPs, with the consequence that only half a million of tag SNPs are enough to determine all ancestral variations, 20 times less than the total number of common SNPs, and are used in GWAS.

To study association of rare variants with complex diseases, rare variants must be described at a population level. Deep sequencing of many individuals is thus necessary. Initiative such as the 1000 Genomes Project (Consortium, 2012) demonstrate that it is now feasible to sequence the complete genomes of representative members of a given population and that many rare variants can be found in such studies (Panoutsopoulou *et al.*, 2013).

Limitations

Association with a tag SNP does not necessarily imply that a variant close to the marker is responsible for the increased susceptibility to the disease. The association can arise because of population stratification, the disease population having a common ancestor not related to the disease susceptibility, or the associated variant may increase the survival of people with the disease.
Even if the association is indeed marker of a disease susceptibility, the causative variant need to be identified and they can be located in relatively extensive chromosomal regions containing many genes. Moreover variants may be found in regions that are not associated with any functional or regulatory role, in which case a functional assessment is extremely difficult.
Therefore, one of the main uses of translational genomics is the prioritization of genes implicated in complex diseases.

In this context Mendelian diseases are an important source of knowledge (Antonarakis & Beckmann, 2006; Brinkman *et al.*, 2006) by offering a direct model for studying the link between the genotype and the phenotype and are not only important in the hope of helping people suffering from these rare diseases but also in the perspective of a better apprehension of complex diseases for prevention and treatments (Craig *et al.*, 2008).

Effect at the protein level

Proteins are the principal mediators of the phenotypic expression of genes. Studying the effect of variants at their level is essential to understand the relationship between DNA sequence variations and diseases. It can vary depending on the type and location of variants. SNPs that affect the protein-coding sequence of a gene can turn one amino-acid into another, creating a missense variant, or into a stop codon, creating a nonsense variant. It can also create a synonymous variant because the genetic code is redundant, most amino acids being coded by several triplet codons. SNPs outside protein-coding regions can affect splicing or change expression level of the mRNA transcript and thus that of the protein and structural variations can lead to premature stop through frameshift or affect the expression level of proteins, such as CNVs (Strachan & Read, 2011).

Missense mutation

Missense mutation is the type of variation most frequently related to human diseases (Antonarakis & Cooper, 2001). Its effect depends on the physicochemical properties of the new amino acid compared to the original. A residue change in size, charge, polarity, or even shape can disrupt the function of a protein depending on its location. Indeed, the function and localization of a protein is based on its precise conformation, flexibility, interaction capacity with other proteins, membrane, nucleic acids or small ligands (Zhang *et al.*, 2012). These behaviors are ruled by favorable energy states through dipole and charge interactions, hydrogen bonding, van der Waals forces and hydrophobic effects (Kahraman *et al.*, 2007) provided by amino-acid side chains. Different mechanisms are presented here by which missense mutations affect protein function, with examples in relation to Mendelian disease.

Active site

A negatively charged amino acid replaced by a positively charged one in the active site of the phenylalanine hydroxylase (PAH) makes the enzyme nearly completely loses its activity and is responsible for phenylketonuria (PKU) (Erlandsen *et al.*, 2003).

Protein - protein interface

Another substitution associated to PKU is an arginine to cysteine change that disrupts a hydrogen bond at the interface between two PAH monomers, destabilizing the dimer.

Protein - DNA interaction

A missense variation in the PAX3 transcription factor leads to deafness and pigmentation abnormalities by preventing DNA binding (Fortin *et al.*, 1997).

Localization signal

A missense mutation in the nuclear localization signal of the short stature homeobox transcription factor (SHOX) abolishes its nuclear localization and leads to dwarfism (Sabherwal *et al.*, 2004; Hung & Link, 2011).

Post-translational sites (PTM)

Disruption of PTM sites might be a rather common mechanism of protein function disruption leading to diseases (Li *et al.*, 2010). For example the mutation of a phosphorylation site in the period circadian protein homolog 2 protein (PER2) is responsible for familial advanced sleep phase syndrome.

Protein stability

A correct folding and stable conformation are also essential and 70% of disease-causing missense mutations are estimated to affect the stability of the protein (Wang & Moult, 2001). For example several PAH mutations away from the active site and leading to PKU have been demonstrated to conserve their enzymatic specificity and kinetics but present an altered activity in vivo explained by misfolding leading to accelerated proteolytic degradation (Waters *et al.*, 2000). Misfolded proteins can also be pathogenic through a gain of function as seen in

Alzheimer, Parkinson or Creutzfeldt-Jakob diseases including familial forms, although the exact pathogenic mechanism is not yet clearly defined such as the role of protein aggregates (Dobson, 2003).

Nonsense mutations

Some mutations introduce a premature stop codon that results in a truncated protein or no protein at all due to nonsense mediated mRNA decay (NMD) that happens when a stop codon is upstream an exon-exon junctions and leads to the degradation of the mRNA before it is translated into a protein. Such mutation can cause for example cystic fibrosis.

Synonymous mutations

Synonymous SNPs can affect the folding of the protein because of different availability rate between tRNA (Buske *et al.*, 2013).

Loss and gain of function

Variations inducing a loss of function usually lead to recessive diseases while some variations can induce a gain of function by increasing or conferring a new activity or changing the spatiotemporal expression of a protein (Lodish *et al.*, 2000). For example, the mutation of a GTPase Ras protein can lead to an overactive form, by making it resistant to GTPase-activating proteins, predisposing to juvenile cancers (Cirstea *et al.*, 2010). A loss of function variation can nevertheless be responsible for dominant diseases through haploinsufficiency or dominant negativity phenomenon.

Genotype to phenotype relationship

Locus heterogeneity

Variations in different genes can lead to the same disease, a phenomenon called locus heterogeneity. The Bardet-Biedl syndrome (BBS) for example can be caused by a mutation in any of at least 18 genes (Katsanis, 2004; www.omim.org/entry/209900). It can arise from the disruption of a function that is performed by a combination of different proteins. Indeed, the disruption of any of these proteins would have the same consequence on the function performed by the group of proteins. In the case of the BBS, the disruption of a protein complex, the BBSome, is responsible for the same BBS phenotype. This complex is necessary for the formation of the primary cilium, an organelle present in nearly all eukaryotic cells that mediates mechanical, thermal and chemical signals (Badano *et al.*, 2006).

Clinical heterogeneity

Two identical mutations can lead to variable disease expressions, going from different degrees of severity to incomplete penetrance. Individuals affected by the BBS in the same family can display for example different ages of onset of retinopathy (Badano *et al.*, 2003). Besides other reasons such as environmental factors, one explanation lies in the presence of modifier genes, whose variation modify the expression of a disease (Genin *et al.*, 2008). For example the gene CCDC28B is a modifier of the BBS penetrance (Badano *et al.*, 2006). Extreme cases of such interaction gives rise to digenic inheritance where mutations in two different genes are required for expressing the disease (Katsanis, 2004).

Of course, outside modifier genes, clinical heterogeneity can also arise when diseases are caused by different mutations in the same gene. In this case, additional mechanisms can explain the difference such as gain versus loss of function and partially-functional versus non-functional mutations. For example the loss of function of the RET gene results in defective intestinal nerve cell migration giving rise to Hirschsprung disease while its overactivation leads to cancer syndrome and the difference between the Becker and Duchenne muscular dystrophy, both caused by mutation in the dystrophin gene, can be explained by the residual function of the protein in the less severe Becker muscular dystrophy.

Pleiotropy

Mendelian diseases often affect different systems, a phenomenon related to pleiotropy. Indeed, pleiotropy refers to the fact that one locus, extensively one gene, can affect two or more apparently unrelated phenotypic traits (Stearns, 2010). It has historically been described as resulting from different mechanisms (Hodgkin, 1998), globally arising either from different functions of a gene product, or from one function of a gene product that has several consequences. The former is often referred as authentic/horizontal/mosaic/independent pleiotropy and the latter as spurious/vertical/relational/reactive (Paaby & Rockman, 2012).

1.5 Objectives of the project

Interoperability between molecular and clinical resources is important especially to investigate the relation between DNA variation and diseases, or genotype to phenotype relationship. However, even if semantic standards exist in clinical medicine and in molecular biology, they exist independently from each other. In particular, controlled medical vocabulary use is scarce in molecular biology and the development of phenotype vocabulary standards is relatively new. Moreover, as animal models are often used, until recently these vocabularies concerned exclusively non-human species, such as the Mammalian Phenotype Ontology (Smith & Eppig, 2009) which mainly describes mouse phenotypes. Accessing biological information related to diseases and integrating them from different resources is then hindered by the different synonyms that can be used to refer

to diseases, by the different degrees of precision used to report diseases amd by the lack of categorization possibilities.

The aim of this work was therefore to enhance the accessibility and medical interoperability of UniProtKB/Swiss-Prot, a central molecular resource through the development of a mapping between its internal controlled vocabulary of Mendelian diseases and a disease controlled vocabulary.

The first task consisted in developing an automatic procedure to extract disease name from textual description and map it to the most appropriate term, if existing, in a standard vocabulary, MeSH. This work led to a publication (Mottaz *et al.*, 2008).

A second objective was to implement a web interface, SwissVar, enabling to query proteins and missense variants from the vocabulary terms and categories, combined with sequence and structural features of variants available from a previous work, leading to another publication (Mottaz *et al.*, 2010).

The aim of the final part was to see how the added knowledge offered by taxonomic relations in controlled vocabularies could contribute to translational efforts. The conclusion was that current hierarchies in disease vocabularies are far from representing the phenotypic complexity of Mendelian diseases, due to pleiotropy. Indeed, reviewing current literature revealed that most genes show some degree of pleiotropy and that it can be related to modularity of cell biology. To suggest directions for further use of Mendelian disorders in translational genomics and because few approaches use clinical data to prioritize molecular information, a prototype tool was developed to filter protein/protein interactions using single phenotypes of Mendelian disorders from HPO hopefully isolating higher level biological processes than when using global Mendelian disease similarity.

2. Mapping UniProtKB/Swiss-Prot to a disease controlled vocabulary

Disease information in UniProtKB/Swiss-Prot was, at the time of the work, presented in a textual description containing a disease term and the corresponding OMIM number.
Some keywords had been created to enhance the retrieval possibilities. However these keywords concerned only the most frequent disease categories and mixed different kinds of relations between proteins and diseases. For example many proteins were indexed with the keyword AIDS, a disease not directly caused by a defect in a protein. Moreover, no synonyms were provided and nearly no hierarchy.
Besides, the majority of proteins were cross-referenced to OMIM. While OMIM is the most important molecular resource for Mendelian diseases, it has been designed to be read by humans and not computers. It does not provide any hierarchy to enhance its access through different levels of specificity or higher level categories. Also it does not provide direct mapping possibilities to clinical resources. Moreover, not all disease annotations in UniProtKB/Swiss-Prot were referenced to OMIM because some protein-disease associations were directly reported from literature.
The automatic mapping of protein entries to a controlled vocabulary includes a step of term matching, after extraction and preprocessing of the disease name, to find which term in the controlled vocabulary corresponds to the disease. A benchmark is then used to evaluate the procedure. But before presenting our procedure, the existing approaches for term matching are overviewed .

2.1 Terminology matching and information retrieval techniques

String matching functions

Two main kinds of approach exist, comparing either letters or words (Cohen *et al.*, 2003).

‘Edit distance like’ functions

In these methodologies, single characters are compared to calculate string similarity or distance. They take into account the similar letters, the different letters or both between two terms. The Levenshtein distance, or edit distance, calculates the number of single character edits necessary to change one word into another. The Jaro-Winkler similarity takes into account the number of common and different letters as well as the transpositions (Winkler, 1999).

Token based functions

'Token based' approaches calculate similarities between two concepts by comparing tokens such as words. There are three main approaches: Jaccard distance, TFIDF similarity score and n-gram.

Jaccard distance

The Jaccard distance takes into account the different and common words, normalized by the total number of words (Jaccard, 1901).

TFIDF similarity score

The TFIDF similarity score is based on the TF-IDF statistic used in information retrieval domain. Information retrieval aims at automatically finding relevant information in resources such as text documents and many techniques are based on the TF-IDF index. TF-IDF stands for term frequency - inverse document frequency. Indeed, documents are ranked according to the frequency of the terms of interest inside the document weighted by their frequency in the complete collection. The consequence of the IDF ponderation is that common words will have a lowered impact on the score calculation. The IDF is logarithmically scaled, giving the following formula for TF-IDF in its simplest form, for a given term *t*, a given document *d* part of a given set of *N* documents *D*:

$$TFIDF(t,d,D) = freq(t,d) \times IDF(t,D)$$
$$IDF = -\log(d{:}t/N)$$

'freq' being the frequency of a term in a document and *d:t* the number of documents among *D* that contain the term *t*.

The TF-IDF score can be used to calculate a similarity between documents. Documents are represented as vectors of terms, each term being weighted by its TF-IDF score, and the cosine of the angle between documents is used.

Inspired by this technique, the TFIDF similarity score is used to calculate the similarity between terms by representing them as a vector of words weighted by their TF-IDF. Since words are not repeated in terms, the weight corresponds in fact to the IDF score alone weight.

N-gram

N-gram approaches take into account sequences of *n* characters. Strings can be compared for example using cosine distance between vectors of *n*-characters tokens weighted by their frequency, as described in TFIDF similarity score.

String-token hybrid methods have also been developed that calculate and sum the string similarity of words (Monge & Elkan, 1996).

TF-IDF or Jaccard methods can also be modified, or softened, by considering common words if their string similarity is above a given threshold.

Comparison of these different techniques have been made, showing that approaches using the information content such as TFIDF approaches work best and that a combination with string matching techniques can improve the performance even more (Cohen *et al.*, 2003).

Preprocessing

Preprocessing steps for terminology matching include syntactic and semantic approaches (Cheatham & Hitzler, 2013).

Syntactic approaches include:

- tokenization or splitting strings into their component words based on delimiters,
- splitting compound words,
- stemming or lemmatization to eliminate grammatical or derivational differences,
- stop word removal, or elimination of very common words,

Semantic approaches include the use of:

- synonyms,
- antonyms,
- translation,
- expand abbreviations and acronyms, by either looking them up in external knowledge sources or using language production rules

Evaluation of terminology matching

Entity matching is evaluated with benchmarks, or standard sets of validated matching entities. Performance is then calculated with recall and precision. Recall corresponds to the number of correct retrieved entities compared to the size of the set. Precision corresponds to the number of correct retrieved entities compared to the total number of retrieved entities.

Recall: Correctly retrieved entities / all relevant entities
Precision: Correctly retrieved entities / retrieved entities

The aim is to maximize both recall and precision. By changing the threshold of the similarity score, recall can be enhanced while lowering precision. Maximizing the mean of both measures is thus a way to obtain the best threshold. A convenient way to average rates is to use the harmonic mean as it lowers the impact of high outliers while raising the impact of small outliers (en.wikipedia.org/wiki/Harmonic_mean). Moreover, depending on which of both measures we want to favor, a ponderation can be used. A value of 2 is given to Beta (F2) to favor recall and 0.5 (F0.5) to favor the precision (Figure 5).

$$F_\beta = \frac{(1+\beta^2)\cdot(precision\cdot recall)}{(\beta^2\cdot precision + recall)}$$

Figure 5. Harmonic mean of precision and recall

2.2 Mapping procedure

Our mapping approach consisted in extracting the disease name from the UniProtKB/Swiss-Prot 'involvement in disease' annotation lines and find for each the most similar term in a given disease vocabulary, using a TFIDF weighted token based similarity score that we developed. It was preceded by term normalization. To increase the number of synonyms, and since most of Swiss-Prot annotation lines contained a reference to OMIM, names and synonyms were retrieved from OMIM to improve the mapping. To determine the score threshold and the procedure to combine SP and OMIM mapping, a benchmark was produced. The final procedure consisted in taking the best match among SP disease and OMIM synonyms above a given threshold, which had been determined by maximizing the harmonic mean of precision and recall. After the publication of the work (Mottaz *et al.*, 2008), to deal with the fact that OMIM contains included titles that are not real synonyms but slightly different concepts, matches of included titles was considered only if the disease extracted from UniProtKB/Swiss-Prot mapped the same entity. Moreover, because our mapping effort was mainly focused on the MeSH vocabulary that is used to index the MEDLINE literature, we took advantage of the MeSH descriptors indexing articles about corresponding proteins to improve the precision. Assuming that the reported association between the disease and the protein came from a publication, only retrieved MeSH descriptors were allowed for the mapping. The publications that we used were the literature references of UniProtKB/Swiss-Prot entries because they are the source of the disease annotations. Publications containing information about gene function, the GeneRIF, which annotate the NCBI Gene entries referenced in UniProtKB/Swiss-Prot, were also retrieved because they are an important source of gene-disease association (Osborne *et al.*, 2007). .

Besides, in order for the missense variants to be mapped as well to the medical vocabulary, they were linked to the disease annotation lines through acronyms present in both disease and variant annotations. The variants that

could not be linked to a disease annotation were directly mapped to the disease vocabulary with our similarity score.

Data storage

All the critical data that were extracted from the different resources and used for the mapping were recorded in a relational PostgreSQL database (www.postgresql.org) for subsequent query. Relational databases enable to store large amounts of interrelated data and query them in a very efficient way on their content. They are based on relations, or tables, containing a set of labeled data, or attributes of the same type. Query can then be made with simple operations such as selection on attribute value or joining different tables. The schema of the tables created for the mapping is presented in the Figure S1, in *Supplementary material*.

Resources description and data extraction

Medical vocabularies

Description

Different disease vocabularies are presented in the *Medical controlled vocabularies* section of the *Introduction* chapter. The main vocabularies available at the time of the choice have been taken into account for the mapping: SNOMED-CT, MeSH and ICD-10. We did not consider the Disease Ontology because it was at that time just starting to be developed, as well as OntoOrpha. The NCI thesaurus, while clearly of value, is specific for cancer and thus could not be used as a primary target vocabulary for the majority of diseases. It could nevertheless have been used for the mapping of cancer information in a second phase.
While SNOMED-CT was the most comprehensive medical vocabulary, the license restrictions made it too complicated to use. UMLS was considered too heavy for our purpose while using a vocabulary contained in this metathesaurus let the possibility of taking advantage of this resource. Therefore our efforts focused on MeSH and ICD-10.

Data extraction

MeSH

The MeSH 'Disease' and 'Psychiatry and Psychology' hierarchies were extracted from the XML file. Terms, concepts and descriptors were retrieved. Treenumbers, associated to descriptors, provided the taxonomic hierarchy. Several treenumbers could be associated to one descriptor since multiple parents are allowed. UMLS concept identifiers were extracted, as well semantic types even if unused since the tree categories 'Disease'

already corresponded to the semantic type of interest. However it could have been useful for the 'Psychiatry and Psychology' since it contains other concepts than psychiatric diseases, such as behavior or emotion concepts. The relation type between concepts and descriptors was recorded, indeed some concepts are slightly narrower concepts compared to descriptors. However since the mapping was done at the level of descriptors while all the terms were used for term matching, this information was not used. Also for this reason and to facilitate database queries since each MeSH term belongs to a concept that itself belongs to a descriptor, a direct link between descriptors and terms was added, providing a shortcut for the queries. The XML file was downloaded from the MeSH FTP server (ftp://nlmpubs.nlm.nih.gov/online/mesh/.xmlmesh/).

ICD-10

The 'master' table was extracted, which contains all the valid codes of the classification, as well as the 'libelle' table, which contains all the texts used in the classification, including terms, synonyms, exclusions, notes and explanations appearing in certain chapters. To be able to retrieve the terms and synonyms of the classification, the tables providing the correspondance between codes and libelles for these entities were extracted: the 'system' table for the systematic classification, the 'descr' table for implicit synonyms and the 'include' table for explicit synonyms. They were extracted from the XML format file retrieved from the WHO website.

UniProtKB/Swiss-Prot

Description

UniProtKB/Swiss-Prot (www.uniprot.org) is a key protein information resource worldwide for life scientists. It is part of the manually annotated section of the UniProt Knowledgebase that is the most comprehensive protein database maintained by the UniProt consortium, a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Georgetown University Medical Center's Protein Information Resource (PIR). It contains, among other species, information on all human proteins in a non-redundant manner (UniProt Consortium, 2009). The annotations are of high quality thanks to the manual curation process that consists of analyzing, comparing and merging all available sequences for a given protein as well as a critical review of associated data from the literature (Boutet *et al.*, 2007). The information concerns both sequence and functional attributes. It contains a wealth of cross-references to other protein and gene resources such as gene expression databases, protein interaction or pathway databases, thus acting as a main hub for data integration in the biomedical domain.

Sequence annotations

The sequence annotations are described on a protein canonical sequence, chosen based on isoform prevalence, similarity with orthologous proteins and what enables the best annotation possibilities such as sequence length.

Their description is at the level of amino-acid sequence. They include regions, such as domains, and sites, such as metal binding, that mediate numerous functional mechanisms. They include also post-translational modifications (PTMs), essential as well for the function. Finally different types of variations are listed, experimental and natural. Most natural variations are single amino-acid variants while small insertions and deletions are sometimes added.

Information on variants include their position on the canonical sequence, the original and the substituting amino acid, the implication in diseases for non-polymorphic variants, the origin of the tissue for somatic mutations, the effect of the mutation on the protein function, links to publications and reference to dbSNP (Sherry *et al.*, 2001) when they exist (Figure 6).

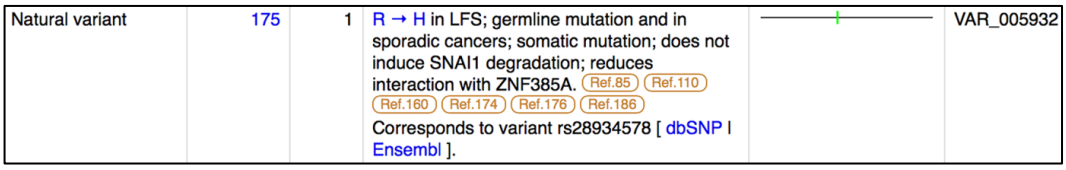


Figure 6. Variant annotation as found in UniProtKB/Swiss-Prot.

The disease information is either given with an acronym whose significance is found in the involvement in disease annotation lines, or directly with a disease name when it corresponds to a somatic mutation. Currently around 70,000 variants, related to disease or not, can be found.

Functional annotations

Information about the function of proteins is presented in a full-text form but also in the form of keywords and Gene Ontology concepts with source references. This resource is used in the *Phenotype-based PPI contextualization* chapter.

Involvement in disease

Information about the implication of proteins in disease is presented in disease annotation lines which format has slightly changed after the beginning of the project. It contains the disease name with a link to the OMIM database and a definition of the disease. The disease annotation on Figure 7 corresponds to the current annotation format where disease name is presented in a standardized way. An acronym of the disease is also given that is used in the variant sequence annotation.

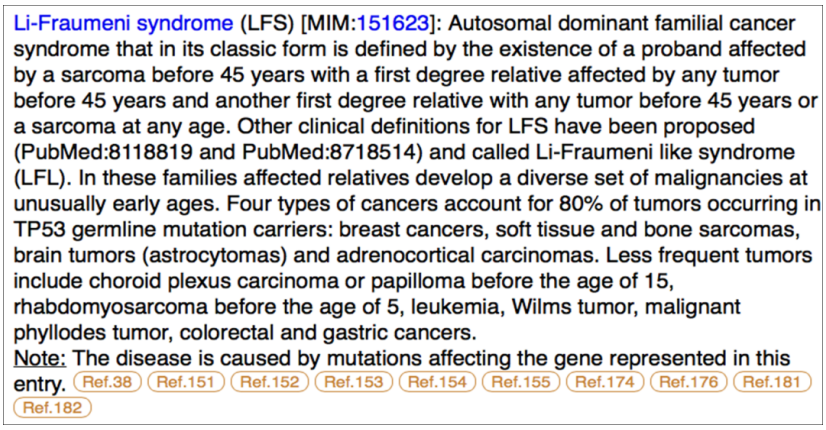


Figure 7. Disease annotation as found in UniProtKB/Swiss-Prot.

Currently, more than 20,200 human proteins are annotated in Swiss-Prot. 3,000 have at least one involvement in disease annotation and about 24,000 single amino acid variants are related to them. Among the 5,000 disease annotations, a majority (86%) are referenced to OMIM.

References

References section contains citations of the literature used to annotate the entry.

Cross-references

Cross-references point to information related to the entry found in other data resources, including the NCBI Gene identifiers.

Data extraction

Data were extracted from the UniProtKB/Swiss-Prot flat file parsed with the Swissknife Perl module (Hermjakob *et al.*, 1999).

Diseases

First the use of 'involvement in disease' annotation required an automatic extraction of the disease name from the full-text annotation. This was done using regular expressions recognizing the context in which diseases were cited (see Additional figure 3, Mottaz *et al.*, 2008 in *Supplementary material*). As already mentionned, the structure of the disease lines have changed after the publication of this work.

A unique identifier was attributed for single disease annotations. Indeed one protein can be associated with several diseases, for example the GTPase KRas (P01116), is implicated in five diseases including 'Noonan syndrome 3' and 'Cardiofaciocutaneous syndrome 2', but no unique identifier is given to refer to them.

OMIM cross-references

OMIM identifiers were retrieved also from the 'involvement in disease' annotation and not from the cross-reference section, enabling them to be linked to corresponding disease annotations and thereby to variants.

Variants

Missense variant identifiers were extracted from the sequence annotation, along with either the acronym that relates them to the disease line and OMIM entry, or a disease name. Since there is no way to easily know if the sequence annotation refers to an acronym, disease, or other types of information, regular expressions were used. First only what followed 'found in', 'detected in', or simply 'in', was considered. Then the extracted text was split around ',' and 'and' in case several diseases were mentioned, and cleaned if necessary, removing words like 'patient' 'family affected by' etc. Cleaned text was then mapped to the acronym extracted from disease lines and in case no corresponding disease line was found, it was mapped directly to the disease vocabulary. Many variants that do not correspond to any disease line in fact correspond to somatic mutation. This information about somatic mutation, contained in the variant annotation, was also extracted.

PubMed identifiers

PubMed identifiers were extracted from the reference section to retrieve associated MeSH terms.

NCBI Gene identifiers

Gene identifiers were extracted from the cross-reference section, to retrieve MeSH terms associated to GeneRIF annotations.

GeneRIF

Description

Gene References Into Function, GeneRIF (www.ncbi.nlm.nih.gov/gene/about-generif), is a resource of the NCBI database of gene specific information (Maglott *et al.*, 2007). It enables to annotate a gene with a concise phrase describing a function from a referenced publication.

Data extraction

NCBI Gene entries in XML format were retrieved through the NCBI API from Gene identifier (http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=gene). PubMed identifiers were extracted from the GeneRIF section to retrieve associated MeSH terms.

PubMed

Description

MEDLINE is the major database of biomedical literature and is accessed through PubMed. It indexes articles with MeSH.

Data extraction

PubMed entries in XML format were retrieved through the NCBI API from PubMed identifier (http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed) and MeSH descriptors extracted to be used as filter for relevant disease terms.

OMIM

Description

The Online Mendelian Inheritance in Man (OMIM) is the most important resource on Mendelian diseases and contains information on all known Mendelian disorders (www.omim.org/). It is the online version of the database initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of Mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). The database is hosted at University of California Santa Cruz (UCSC) Genome Bioinformatics. The web access to the database is provided by the National Center for Biotechnology Information (NCBI), a service by the National Library of Medicine (NLM), while the content is edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine (Hamosh *et al.*, 2005).

OMIM contains around 8,000 different diseases and non-pathologic phenotypes. Among them, 4,000 have a known molecular basis, '#' entries corresponding to a phenotype with several associated locus and '+' entries to a phenotype associated to one locus only. Other types of entries are '*' genes, '%' mendelian phenotype or phenotypic locus for which the underlying molecular basis is not known, and phenotypes for which the mendelian basis, although suspected, has not been clearly established or which separateness from that in another entry is unclear.

Each phenotype entry contains an extensive full-text summary of knowledge on the disease and references to publications. Information on genes is available as well as tools such as search by genomic regions (www.omim.org/search/advanced/geneMap). Also a summary of clinical findings is provided containing the clinical traits found in the disease. Indeed, genetic diseases are often composed of several different traits. They can go from simple non-pathogenic phenotypes like café-au-lait spots to more pathological phenotypes such as increased risk of leukemia or mental retardation (see *Phenotype-based PPI contextualization* chapter).

Data extraction

OMIM titles and alternative titles were extracted from the flat file downloaded from the OMIM FTP server (ftp.omim.org/OMIM/omim.txt.Z). The extraction procedure had to deal with irregular formatting of titles and alternative titles. For example some titles were separated by semicolon, simple or double, others by newline, while some titles were split in half by newline. We used regular expressions to deal with these formatting variations. Unfortunately the XML format did not seem to resolve all the formatting problems, therefore we kept the flat file extraction procedure.

Programming languages

Programs were implemented with the Perl 5 programming language (www.perl.org/).
The access to the database was implemented using the DBI module (dbi.perl.org/).
XML files were parsed with the Perl XML::TWIG module, efficient to process large XML files by building only selected parts of XML tree (xmltwig.org/module/).
The programming code was organized into modules, one for each resource, grouping together the functions necessary to download the resources, extract the data from them, create, fill and query the database tables containing the retrieved information (Table 3).

Table 3. Perl modules and related database tables

Perl Modules	Database Tables
Sp.pm	swissprot acsec spdisease spdisease_variant spdisease_omim variant_mesh
Omim.pm	omim omim_title
EntrezGene.pm	-
Pubmed.pm	pubmed_mesh
Mesh.pm	term concept

	descriptor treenumber semantictype concept_semantictype conceptulms
Mapping.pm	-
Normalize.pm	-
Result.pm	sp_mapping_mesh omim_mapping_mesh final_mapping_mesh
DbConnection.pm	-
UnimedConfig.pm	-

Similarity score

A similarity score was calculated between the extracted disease name and the terms of the medical vocabulary to find the most similar one, preceded by term preprocessing.

Exact match

A match was considered exact when both terms were composed of exactly the same words.

Partial match

The similarity score calculation between the diseases terms and the MeSH terms was inspired from the TF-IDF measure and the Jaccard index (Figure 8).

It consisted of decomposing the terms into words, or tokenization and summing the common words and subtracting the different ones. Each word was weighted according to the logarithm of the IDF evaluated with its frequency in the whole set of OMIM titles, alternative titles and Swiss-Prot diseases annotations. The score was then divided by the number of words composing the disease to match. An example is presented on Figure 9.

$$\frac{\sum \log(1/freq(cw)) - \sum \log(1/freq(ncw))}{size(disease)}$$

Figure 8. Similarity score formula

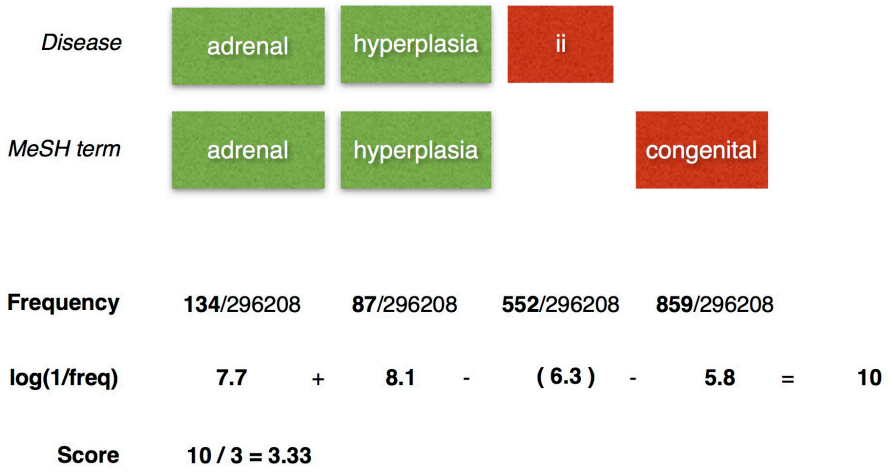


Figure 9. Example of similarity score calculation

Preprocessing

Syntactic

To deal with syntactic issues, we used a normalization program, Norm, distributed by the UMLS, as part of the Specialist Lexical tools (/www.nlm.nih.gov/pubs/factsheets/umlslex.html). It enabled to deal with:

Inflection: 'cancer, esophageal' and 'cancers, esophageal'

Stop words: 'NOS' , 'and', 'to', …

Tokenization enabled to deal with:

Syntaxy: 'cancer, esophageal' and 'cancer of the esophagus'

Tokenization helped also to deal with hyphenated terms but they were treated in a special way to avoid false positive matches without penalizing the sensitivity. Each of their components was considered as distinct word. If all components had a matched equivalent, their respective weights were summed up in the score calculation. Otherwise, their weights were subtracted.

To deal with small words such as numbers, any word composed of three letters or less was not taken into account except if all other words matched.

Semantic

Synonyms

All the synonyms provided by OMIM were used, as well as the synonyms from MeSH. The best match was considered. Combining the scores of the different synonyms into a global score could have been useful but has not been done during this project.

Evaluation and results

The benchmark set was composed of 200 disease annotations from randomly selected human UniProtKB/Swiss-Prot entries manually mapped in the framework of this work to the MeSH terminology and to the ICD-10 classification. The mapping was done and validated at the level of descriptors for MeSH and codes for ICD-10. When the manual mapping had to be done on several codes or descriptors, the automatic mapping was considered correct if any of them was mapped. Recall and precision were calculated according to the formula presented in the *Evaluation of terminology matching* section of this chapter and, for the first evaluation, separately for UniProtKB/Swiss-Prot disease names (SP) and OMIM. The harmonic mean favoring precision (F0.5) was calculated for different score threshold values and the threshold was chosen to maximize this measure with SP mapping (see Mottaz *et al.*, 2008). The combination of SP and OMIM mapping was finally decided to be the union of both mapping, that is the best mapping among SP disease names and OMIM titles and alternative titles. In the mapping examples below, true positive mappings correspond to correct mappings above the threshold, false negative to correct mappings below the threshold, true negative to wrong mappings below the threshold and false positive to wrong mappings above the threshold.

Mapping to ICD-10

The mapping to ICD-10 was rapidly abandoned due to very low recall, around 35%, and precision, around 66%. As seen in the false positive mappings (Table 4), ICD-10 is meant to be used with the knowledge of the whole classification. The approach developed here consisting in simply selecting the most similar term in the vocabulary could not work. The mapping of 'Childhood ataxia with central nervous system hypomyelinization' for example to the term 'Nervous system (central) NOS' is not appropriate because the ICD-10 term refers to the code D33.9 which is in fact a child of 'benign neoplasm of brain and other parts of central nervous system'. Finding ways to deal with this kind of problems did not seem worth the effort given the coarse granularity of Mendelian disease codes in ICD-10. For example in the false negative mappings (Table 4), the otopalatodigital syndrome was wrongly mapped to the orofaciodigital syndrome but the corresponding code, Q87.0, referring to 'Congenital malformation syndromes predominantly affecting facial appearance' was correct. Interoperability with such loss of information was not considered of interest in this work.

Table 4. ICD-10 examples of mapping		
OMIM or Swiss-Prot disease	**Automatic mapping**	**Manual mapping**
TRUE POSITIVE		
Idiopathic generalized epilepsy	Generalized idiopathic epilepsy and epileptic syndromes	Generalized idiopathic epilepsy and epileptic syndromes
Limb-girdle muscular dystrophy type 2B	Muscular dystrophy limb-girdle	Muscular dystrophy limb-girdle
Autosomal dominant rhegmatogenous retinal detachment	Rhegmatogenous retinal detachment	Rhegmatogenous retinal detachment
Epidermolysis bullosa simplex Dowling-Meara type	Epidermolysis bullosa simplex	Epidermolysis bullosa simplex
Tangier disease	Tangier disease	Tangier disease
Hypophosphatasia infantile	Hypophosphatasia	Hypophosphatasia
Nemaline myopathy type 1	Myopathy nemaline	Myopathy nemaline
FALSE NEGATIVE		
Squamous cell carcinoma of the head and neck	Head, face and neck	Head, face and neck
Otopalatodigital syndrome type 1	Syndrome oro-facial-digital	Congenital malformation syndromes predominantly affecting facial appearance
Posterior polymorphous corneal dystrophy 2	Hereditary corneal dystrophies	Hereditary corneal dystrophies
Hyperinsulinemic hypoglycemia, familial, 3	Other hypoglycaemia	Hyperinsulinism NOS
Pachyonychia congenita type 2	Pachyonychia	Pachyonychia
Malignant hyperthermia susceptibility 5	Malignant hyperthermia due to anaesthesia	Malignant hyperthermia due to anaesthesia
Microphthalmia, isolated, with coloboma 5	Coloboma NOS	Coloboma of iris / Coloboma of the fundus / Congenital malformation of choroid
TRUE NEGATIVE		
Short qt syndrome 2	Short rib syndrome	Arrhythmia (cardiac) NOS
Bleeding disorder	Puberty bleeding	Other specified haemorrhagic conditions
Alternating hemiplegia of childhood	Hemiplegia	Other specified paralytic syndromes
Trifunctional protein deficiency	Protein deficiency anaemia	Disorders of fatty-acid metabolism
Iridogoniodysgenesis anomaly	Congenital anomaly NOS	Other congenital malformations of anterior segment of eye
Endometrial stromal tumors	Tumour NOS	Uterus
Solitary median maxillary central incisor	Median nerve NOS	Hypodontia
FALSE POSITIVE		
Myopathy, distal, with anterior tibial	Anterior tibial syndrome	Muscular dystrophy distal

onset		
Femoral head, avascular necrosis of	Avascular necrosis of bone	Idiopathic aseptic necrosis of bone
Childhood ataxia with central nervous system hypomyelinization	Nervous system (central) NOS	Other specified demyelinating diseases of central nervous system
Cataract, embryonic nuclear	Nuclear sclerosis cataract	Congenital cataract
Senile cataract	Senile cataract	Senile cataract, unspecified
Cardiomyopathy, familial hypertrophic, 8	Other hypertrophic cardiomyopathy	Obstructive hypertrophic cardiomyopathy
Pfeiffer syndrome	Pfeiffer's disease	Congenital malformation syndromes predominantly affecting facial appearance

Mapping to MeSH

The mapping to MeSH yielded better results than ICD-10. A recall of 64% with a precision of 86% was obtained. The whole benchmark mapping is presented in the Additional figure 1, Mottaz *et al.*, 2008, *Supplementary material* section. The analysis of the results led to the conclusion that the lack of coverage of the automatic mapping was due to an incomplete coverage of Mendelian diseases by MeSH (Table 5). Indeed, nearly half the diseases, 86 of 200, had been manually mapped to more than one descriptor. It means that these diseases do not have any descriptor directly corresponding to them. Improving the procedure by trying to map to more general categories have been considered. However, categories are based, besides transmission type and etiology, on affected systems or anatomy. Mapping to pathologies would require parsing description of diseases to extract the pathological traits. For example the otopalatodigital syndrome should map to 'X-linked genetic disease', 'Multiple abnormalities', 'Osteochondrodysplasia' and 'Craniofacial abnormalities'. Such mapping is of great interest as seen in the last section of this work. However it would have required consequent efforts to map to a relatively coarse granularity hierarchy. Such efforts would have been better employed mapping all clinical synopses into phenotype ontology, effort that have been done meanwhile by other groups and used in the last part of the work.

We compared our similarity score with a promising cosine similarity TFIDF score taking advantage of synonyms and partial string matching, kindly provided by its author (Ha-Thuc & Srinivasan, 2007). Comparing recall and precision with different thresholds, the results appeared not better and even slightly lower than with our approach on the benchmark (see Mottaz *et al*., 2008).

Table 5. MeSH examples of mapping		
OMIM or Swiss-Prot disease	**Automatic mapping**	**Manual mapping**
TRUE POSITIVE		
Epidermolysis bullosa herpetiformis, Dowling-Meara type	Epidermolysis bullosa herpetiformis Dowling Meara	Epidermolysis bullosa herpetiformis Dowling-Meara
Autosomal dominant nocturnal frontal lobe epilepsy type 3	Frontal lobe epilepsies	Genetic disease, inborn | Epilepsy, frontal lobe
Corneal dystrophy, Fuchs endothelial, 1	Fuchs endothelial dystrophy	Fuchs endothelial dystrophy
Hypokalemic periodic paralysis	Hypokalemic periodic paralysis	Hypokalemic periodic paralysis
Isolated ectopia lentis	Ectopia lentis	Genetic disease, inborn | Ectopia lentis
Reading disability, specific, 2	Developmental reading disabilities	Dyslexia | Genetic predisposition to disease
Familial hemiplegic migraine 2	Familial hemiplegic migraines	Hemiplegic migraine, familial
FALSE NEGATIVE		
Metatropic dwarfism, type II	Dwarfism	Genetic disease, inborn | Abnormalities, multiple | Osteochondrodysplasia | Dwarfism | Craniofacial abnormalities
Osteoarthritis with mild chondrodysplasia	Osteoarthritides	Genetic disease, inborn | Osteochondrodysplasia | Osteoarthritis
Polydactyly, preaxial II	Polydactylies	Limb deformities, congenital | Genetic disease, inborn | Polydactyly | Syndactyly
Autosomal recessive osteopetrosis	Osteopetrosis	Osteopetrosis | Genetic disease, inborn
Osteopetrosis, autosomal recessive 5	Osteopetrosis	Osteopetrosis | Genetic disease, inborn
Alport syndrome, mental retardation, midface hypoplasia, and elliptocytosis	Alport's syndrome	Genetic disease, X-linked | Abnormalities, multiple | Nephritis, hereditary | Elliptocytosis, hereditary | Craniofacial abnormalities | Mental retardation, X-linked
Arthrogryposis, distal, type 7	Arthrogryposis	Abnormalities, multiple | Arthrogryposis
TRUE NEGATIVE		
Peeling skin syndrome, acral type	Skin diseases	Skin disease, genetic | Skin abnormalities | Skin disease, vesiculobullous
Costello syndrome	Syndromes	Genetic disease, inborn | Abnormalities, multiple | Craniofacial abnormalities | Skin abnormalities | Heart defects, congenital
ICOS deficiency	Deficiency diseases	Common variable immunodeficiency
Glaucoma iridogoniodysplasia, familial	Glaucoma	Abnormalities, multiple | Eye disease, hereditary | Glaucoma, angle-closure | Eye abnormalities

Peters anomaly	Anomalies, pupillary	Eye disease, hereditary | Eye abnormalities
Episkopi blindness	Blindness	Genetic disease, X-linked | Eye disease, hereditary | Retinal dysplasia
Polyposis syndrome, hereditary mixed, 2	Familial polyposis syndrome	Intestinal polyposis | Neoplastic syndrome, hereditary | Colonic neoplasms
FALSE POSITIVE		
Distal myopathy with anterior tibial onset	Tibial syndrome, anterior	Distal muscular dystrophy
Amyloidosis, corneal	Amyloidoses	Corneal dystrophy, hereditary
Stem cell leukemia lymphoma syndrome	T-cell leukemia-lymphoma, adult	Precursor cell lymphoblastic leukemia-lymphoma
Pro-lymphocytic T-cell leukemia	Leukemia, t-cell	Leukemia, prolymphocytic, T-cell
Microphthalmia and esophageal atresia syndrome	Esophageal atresias	Anophthalmia | microphthalmos
Chromosome 22q13.3 deletion syndrome	Deletions, chromosome	Genetic disease, inborn | abnormalities, multiple | autosomal chromosome disorder
Inclusion body myopathy type 2	Inclusion body myopathy, sporadic	Myopathy

Final procedure

After our results were published (Mottaz *et al.*, 2008), we improved the precision to 93% and kept the recall to 63% by mapping to a selection of MeSH descriptors indexing the publications referenced in Swiss-Prot as well as those indexing the GeneRIF publications. As already mentioned, the idea behind was that the association between the disease and the protein reported in Swiss-Prot came from a published result referenced in Swiss-Prot. To enhance the coverage of pertinent publications, we added GeneRIF because they are an important source of gene-disease association (Osborne *et al.*, 2007). Also we considered OMIM included title matches only if it matched the same descriptor than SP disease. This avoided a wrong mapping when the included title and the main OMIM title corresponded to different MeSH descriptors. For example the OMIM entry 'Maturity onset diabetes of the young type 2' (MIM number 125851) has an included title 'diabetes gestional' both corresponding to different MeSH descriptors.

The final global automatic mapping procedure of Swiss-Prot entries to the MeSH vocabulary is presented on Figure 8.

Currently 68% of disease annotations are mapped to MeSH, along with associated missense variants (Table 6). Nearly 5,000 variants are directly mapped to MeSH, often corresponding to somatic mutations.

Table 6. Mapping statistics, UniProtKB/Swiss-Prot 2014_07		
	Extracted	**Mapped to MeSH**
Number of disease annotations (with OMIM)	5,116 (4,391)	3,468 (3,057)
Number of disease related variants	31,086	Through disease annotation: 22,462 Directly to MeSH: 4,575

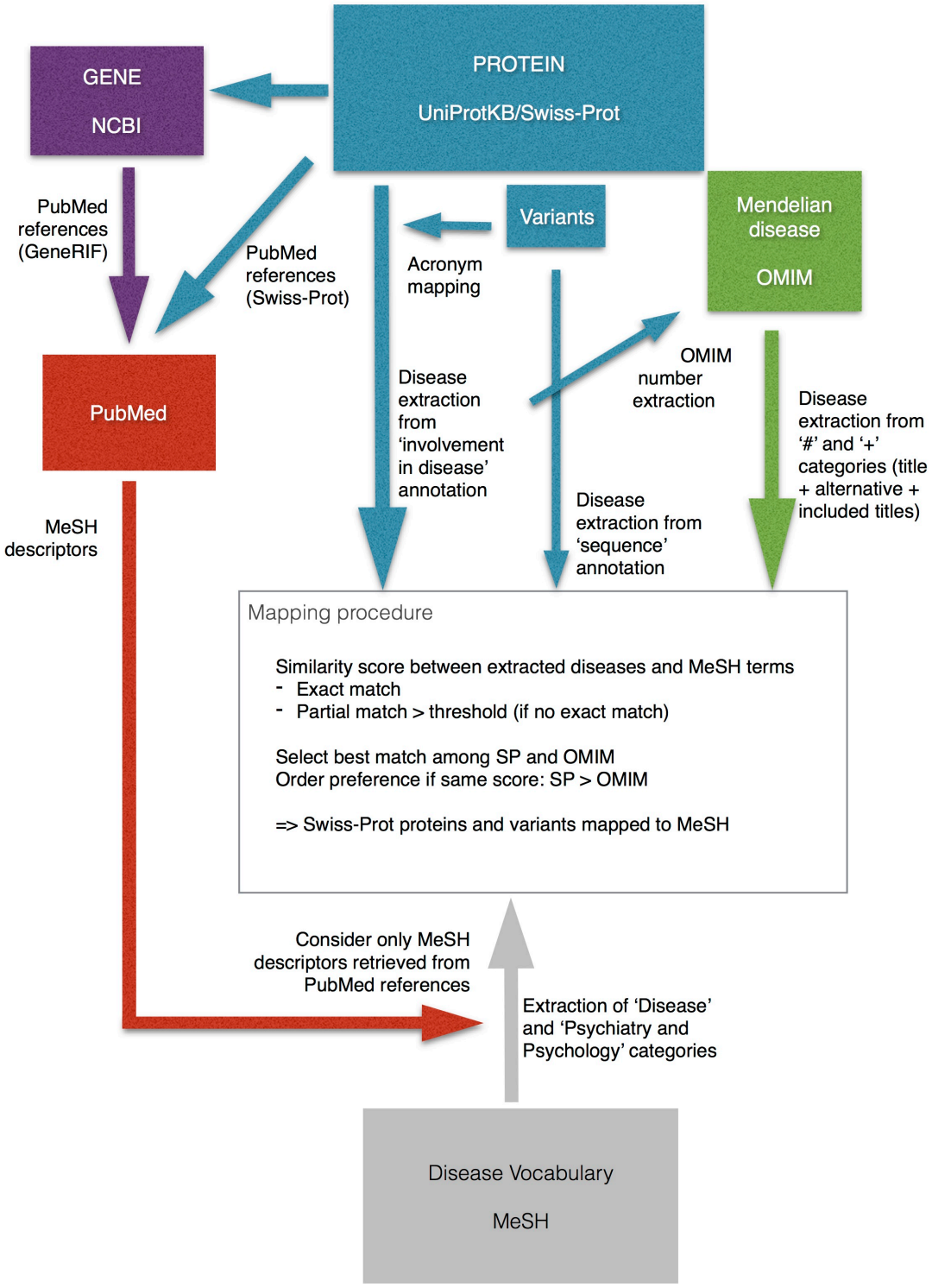


Figure 10. Final procedure of UniProtKB/Swiss-Prot mapping to MeSH.

BMC Bioinformatics

BioMed Central

Proceedings

Open Access

Mapping proteins to disease terminologies: from UniProt to MeSH

Anaïs Mottaz[1], Yum L Yip[1,2], Patrick Ruch[3] and Anne-Lise Veuthey*[1]

Address: [1]Swiss-Prot Group, Swiss Institute of Bioinformatics, 1211 Genève 4, Switzerland, [2]Department of Structural Biology and Bioinformatics, University of Geneva, 1211 Genève 4, Switzerland and [3]Medical Informatics Service, Hôpitaux Universitaire de Genève, 1211 Genève 4, Switzerland

Email: Anaïs Mottaz - anais.mottaz@isb-sib.ch; Yum L Yip - lina.yip@isb-sib.ch; Patrick Ruch - patrick.ruch@sim.hcuge.ch; Anne-Lise Veuthey* - anne-lise.veuthey@isb-sib.ch

* Corresponding author

from 10th Bio-Ontologies Special Interest Group Workshop 2007. Ten years past and looking to the future
Vienna, Austria. 20 July 2007

Published: 29 April 2008

BMC Bioinformatics 2008, **9**(Suppl 5):S3 doi:10.1186/1471-2105-9-S5-S3

This article is available from: http://www.biomedcentral.com/1471-2105/9/S5/S3

Abstract

Background: Although the UniProt KnowledgeBase is not a medical-oriented database, it contains information on more than 2,000 human proteins involved in pathologies. However, these annotations are not standardized, which impairs the interoperability between biological and clinical resources. In order to make these data easily accessible to clinical researchers, we have developed a procedure to link diseases described in the UniProtKB/Swiss-Prot entries to the MeSH disease terminology.

Results: We mapped disease names extracted either from the UniProtKB/Swiss-Prot entry comment lines or from the corresponding OMIM entry to the MeSH. Different methods were assessed on a benchmark set of 200 disease names manually mapped to MeSH terms. The performance of the retained procedure in term of precision and recall was 86% and 64% respectively. Using the same procedure, more than 3,000 disease names in Swiss-Prot were mapped to MeSH with comparable efficiency.

Conclusions: This study is a first attempt to link proteins in UniProtKB to the medical resources. The indexing we provided will help clinicians and researchers navigate from diseases to genes and from genes to diseases in an efficient way. The mapping is available at: http://research.isb-sib.ch/unimed.

Background

Biomedical data available to researchers and clinicians have increased drastically over the last decade because of the exponential growth of knowledge in molecular biology. While this has led to the creation of numerous databases and information resources, the interoperability between the resources remains poor to date. One of the main problems lies in the fact that medical terminologies are scarcely used in molecular biology. For instance, while the UniProt Knowledgebase (UniProtKB) - the most comprehensive protein warehouse with extensive cross-references to other database resources [1] – contains more than 2,000 human proteins with manually curated information related to their involvement in pathologies, this

information is not easily accessible for clinical researchers. This is due to the fact that UniProtKB does not use standard medical vocabularies to describe diseases associated to proteins and their variants.

In order to increase the interoperability between the biomolecular and clinical resources, one of the key solutions lies in the development or unification of common terminologies capable of acting as a metadata layer to provide the missing links between the various resources. In the medical/clinical domain, there have already been numerous and successful efforts to implement controlled vocabularies for pathologies. Terminologies such as MeSH - the controlled vocabulary thesaurus used for biomedical and health-related documents indexing [2], ICD-10 - the official disease classification provided by the World Health Organisation (WHO) for diagnostic information [3], and SNOMED-CT – the clinical terminology used for clinical information [4], have all served well in their respective domain of application. Most of these terminologies are collected and organised into concepts in the UMLS, a major repository of biomedical standard terminologies [5].

The recent integration of the Gene Ontology (GO) [6] into the UMLS, as well as the development of numerous biological ontologies under the Open Biological Ontologies initiative (OBO) [7], have opened new ways of linking biological and medical resources via terminologies. Therefore, terminology and ontology mapping has become an active field of research, the objective being identifying correspondence between concepts of different resources. The National Library of Medicine (NLM) made an important pioneer effort through the integration of more than 60 medical vocabularies in the UMLS Metathesaurus and the development of lexical tools for this purpose [8]. In parallel, many approaches have been developed which integrate lexically-based, as well as knowledge- and semantics-based methods to map, for instance, GO terms to UMLS concepts [9,10], representations of anatomy [11], genotypic and phenotypic data [12,13]. In the biological field, identical initiatives are emerging for linking OBO ontologies [14]. It was shown that the mapping could be improved by a combination of lexical alignments and hybrid mapping techniques which integrate structural properties of the ontologies. The most advanced tools for aligning and merging ontologies indeed take advantage of both the similarity between terms and the structural features of the resources.

In this study, we tested different automatic approaches to map the disease terms in UniProtKB to MeSH. The MeSH thesaurus is the NLM's controlled vocabulary for subject indexing in MEDLINE [2]. It is structured in a hierarchy of descriptors, with each descriptor including a set of concepts, and each concept itself containing a set of terms, which are synonyms and lexical variants. This rich vocabulary is included in the UMLS and, therefore, is linked to many other biomedical terminologies. The mapping procedures described below took advantage of the manual annotation in UniProtKB as well as the curated links of UniProtKB entries to OMIM, a comprehensive knowledge base of human genes and genetic diseases [15]. A benchmark set was created for the evaluation and refinement of term matching algorithms.

Results

Overview of the mapping procedure

We mapped the disease names extracted from Swiss-Prot annotations to terms from the disease category of the MeSH terminology. The complete procedure is summarised in Fig. 1. It consisted of three successive steps:

(1) we extracted the disease names from the Swiss-Prot and OMIM entries;

(2) for each disease name, we looked for an exact match with a MeSH term where all words composing the name had an identical correspondent in a MeSH term and vice versa;

(3) when the previous step failed, we looked for partial matches by decomposing the name into its word components and calculate a similarity score with MeSH terms.

To define the whole procedure, a benchmark set was created for the evaluation and refinement of term matching algorithms. Different methods adapted from textual information retrieval techniques were tested. Namely, we evaluated the effect of linguistic pre-processing of the terms to get rid of word lexical variations (with/without normalisation). A method developed by Ha-Thuc and Srinivasan for gene name recognition was also tested [18].

The methods were assessed in term of *retrieval*, *recall* and *precision*, which measure the proportion of terms mapped among all terms, the proportion of terms correctly mapped among all terms, and the proportion of terms correctly mapped among mapped terms, respectively. A detailed description of the methodology is provided in the Methods section.

The benchmark set

We constructed a benchmark set consisting of 200 randomly selected diseases manually mapped to one or several MeSH terms. The principal problem encountered in this manual mapping process was the lack of specificity of MeSH in the field of genetic diseases. This means that only a quarter of the disease names (52) were mapped to a term of similar meaning. For the other 148 ones, we mapped to

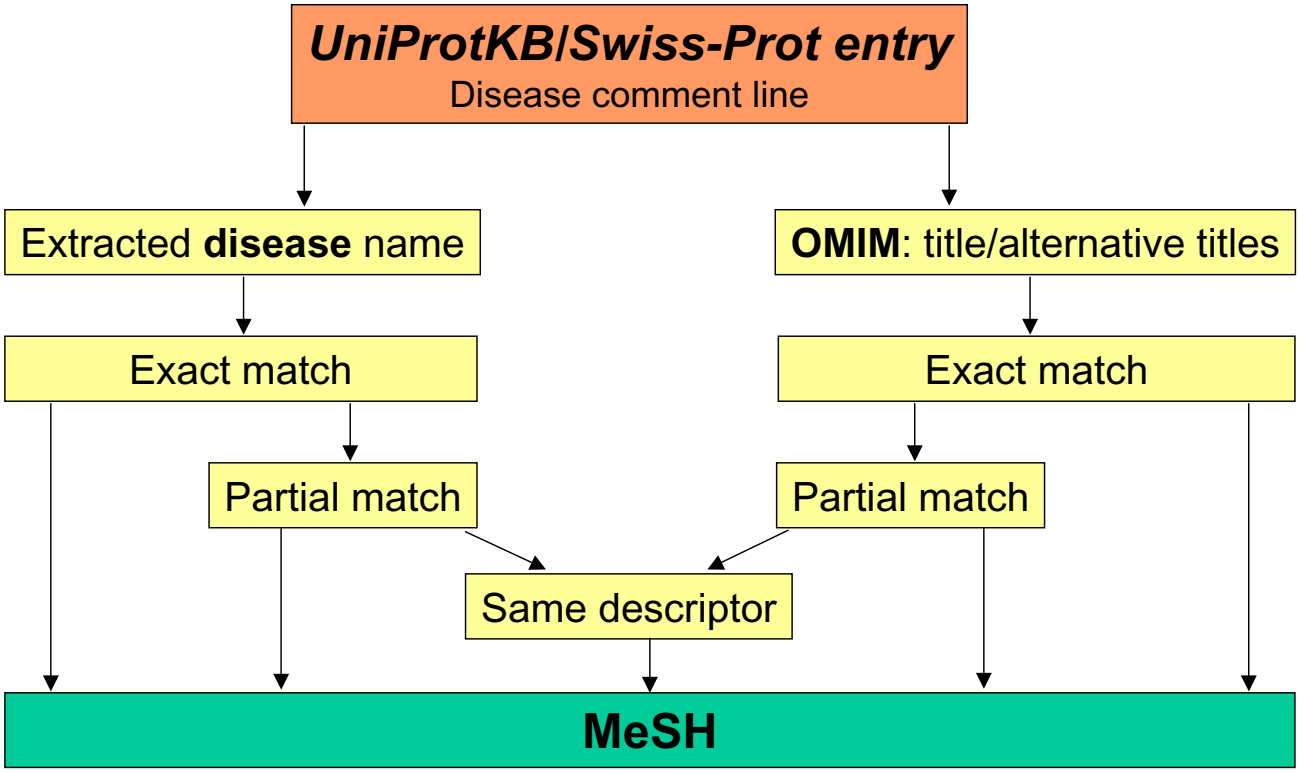


Figure 1
Procedure of the mapping of UniProtKB/Swiss-Prot disease comment lines to MeSH terms.

a term with coarser granularity and, for 90 of them, we had to choose more than one parent term since the same term could belong to several branches in the MeSH hierarchy. For instance, the disease name *X-linked congenital idiopathic intestinal pseudoobstruction* (P21333) was associated to the MeSH term *Intestinal Pseudo-Obstruction*. However, this term is in no way linked to a branch indicating the genetic origin of the disease. Therefore, we mapped the disease to two other coarser terms belonging to other hierarchies: *Genetic Disease*, *X-Linked* and *Digestive System Abnormalities*.

The manually mapped terms were used to evaluate the performance of automatic procedures described below.

Disease name extraction

In Swiss-Prot, the manually annotated section of UniProtKB (release 54.1), 2,252 human protein entries contained information on the involvement of these proteins in a total of 3,408 diseases, mainly of genetic causes (Fig. 2). We extracted almost all disease names from the UniProtKB/Swiss-Prot free text comment lines with a set of regular expressions. The extraction failed in only 7 comment lines where a clear reference to a disease was not expressed, for instance:

"(CBL) can be converted to an oncogenic protein by deletions or mutations that disturb its ability to down-regulate RTKs." (P22681)

By manually assessing the extraction results, we noticed that as the system was constructed to extract only a single disease name per line, it was unable to treat lines such as:

"KRT16 and KRT17 are coexpressed only in pathological situations such as metaplasias and carcinomas of the uterine cervix and in psoriasis vulgaris." (P08779)

We did not investigate further these cases, as the structure of disease lines is scheduled for revision as part of Swiss-Prot annotation standardization efforts.

In parallel, we extracted disease names and synonyms from the 2,087 OMIM phenotypes (#) and genes with phenotypes (+) entries cited in the 2,601 Swiss-Prot disease lines. This corresponded to 82% of the total OMIM

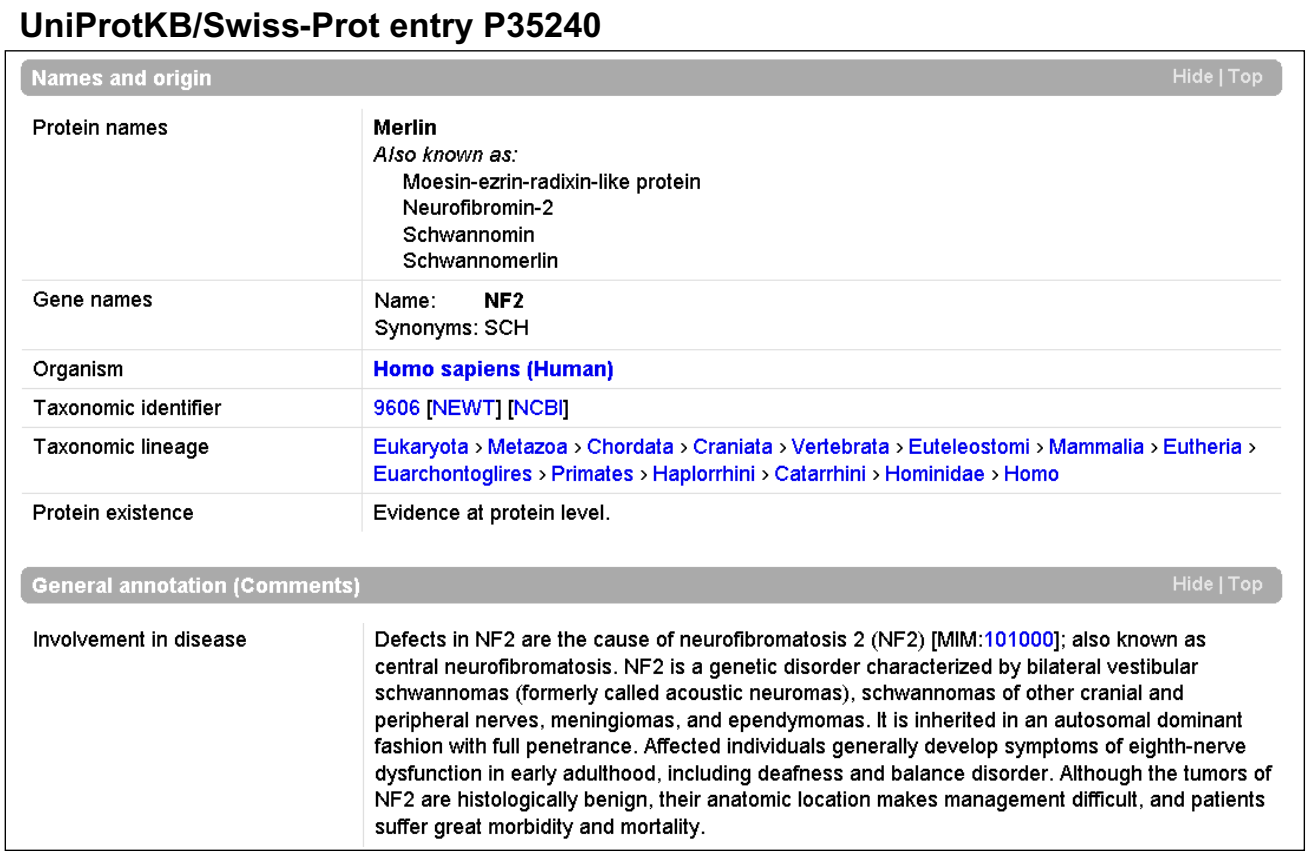


Figure 2
Disease comment lines in a UniProtKB/Swiss-Prot entry.

entries on phenotypes with a known molecular basis (v. August 2007).

Establishing the mapping procedure using the benchmark set

The 200 disease names of the benchmark set and their associated OMIM terms were automatically mapped to the "Diseases" and "Psychiatry and Psychology" categories of the MeSH (v. August 2007). This subset of MeSH consists of 43,220 different terms. The automatic mapping procedure was done independently on disease names from Swiss-Prot and from OMIM. Different techniques were evaluated to maximize the number of exact and partial term matches.

Exact matches

Briefly, the step consisted of transforming all terms into *bag of words* either with or without word normalisation. The word normalisation step was performed using the *Norm* program of the NLM [16]. The effect of term preprocessing was found to be not significant on this dataset, the two procedures giving exactly the same results (Table 1, columns 1-3). All exact matches provided by Swiss-Prot disease names were correct. It was found that the coverage obtained using OMIM terms was better. This could be explained by the presence of synonyms for each disease, which increased matching opportunities. The presence of synonyms however also augmented the risk of possible incorrect mappings. Indeed, the only three false positive matches were caused by a difference of classification between MeSH and OMIM. For instance, two types of *epidermolysis bullosa*, which are distinct MeSH descriptors, are synonyms in OMIM. When we gathered the exact matches provided by Swiss-Prot and OMIM, the recall increased to 26%, with a precision of 96%. It should be noted that the overlap of disease mapping from the two resources did not necessarily mean that the matching terms were the same, but rather that they belonged to the same descriptor in the MeSH terminology.

Partial matches

The disease names not mapped by exact matches went through a partial matching procedure. For this, three separate procedures were tested in order to evaluate the effect

of term pre-processing as well as the use of different scoring functions:

Procedure 1: Term pre-processing followed by calculation of a similarity score for matching terms based on an adaptation of the weighting schema 'Term Frequency x Inverse Document Frequency' (TFIDF) [17];

Procedure 2: No term pre-processing followed by calculation of the same similarity score as in procedure 1;

Procedure 3: Use of the program developed by Ha-Thuc and Srinivasan [18].

The weighting schema TFIDF is commonly used in information retrieval techniques. This scoring method allows evaluate the informative content of a word in a collection or documents. Ha-Thuc and Srinivasan's program uses a different adaptation of TFIDF which allows partial matches at the word level [19,20]. The method also takes advantage of synonymy resources to improve the similarity scoring by increasing the weights or words common to several synonyms.

The three procedures were evaluated in terms of trade-off between recall and precision (Fig. 3). As already noticed with exact matches, the global performance was better with OMIM terms rather than with Swiss-Prot disease names. This is because of the richer terminology used to define OMIM phenotypes. Likewise, we did not observe significant differences due to term pre-processing. This lack of effect could be explained by the fact that the MeSH vocabulary already includes lexical and orthographic variants, therefore reducing the utility of term normalization.

The performance of the Ha-Thuc's synonym-based similarity scoring was slightly lower than the simpler scoring system we developed. This could be due to the fact that their program calculated a vector similarity measure using the *cosine coefficient*. Indeed, in a first attempt to set up a scoring schema, we noticed that the *cosine coefficient* was less effective on our data. It appears therefore that this similarity measure, although widely used in information retrieval from texts, is less efficient for terminology mapping.

Based on these evaluations, we decided to set up the complete mapping procedure using the scoring method we developed. The word normalisation pre-treatment was included in the procedure even though it did not result in a real gain of performance. The reason for this choice was due to our intention to map Swiss-Prot diseases to ICD-10, which does not include lexical resources. Therefore, a word normalization step could be essential.

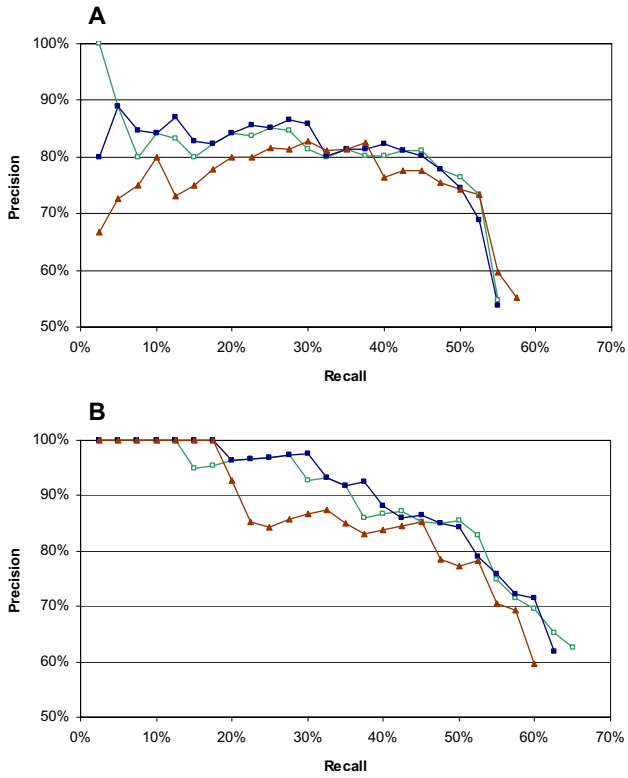


Figure 3
Recall –precision curves for partial matches of Swiss-Prot disease names (A) and OMIM *titles* and *alternative titles* (B) to the disease MeSH terms, with term normalisation (blue squares), without normalisation (green empty squares), and with the method developed by Ha-Thuc (red triangles). The data have been ordered according to the score and the precision is calculated at increasing recall intervals.

With the choice of the scoring schema, we proceeded to select a similarity score threshold above which a partial mapping could be considered as correct. The threshold was selected by determining the maximal performance of the system estimated with the *F*- measure, which is the weighted harmonic average of precision and recall (Fig. 4). As the prerequisite for a fully automatic mapping process was high precision, the *F*-measure was parameterized accordingly. We chose a score threshold of -2.5 around which maxima of *F*-measure were found for both OMIM and Swiss-Prot mappings.

The overall system performance was assessed using this threshold for partial matches of the benchmark dataset (Table 1, columns 4-6). It was found that when combining exact and partial matches of Swiss-Prot disease names and OMIM terms, a recall of 64% for a precision of 86% were obtained (Table 1, columns 7-9). While this precision is clearly sufficient to aid manual curation, we could further improve the mapping procedure in terms of preci-

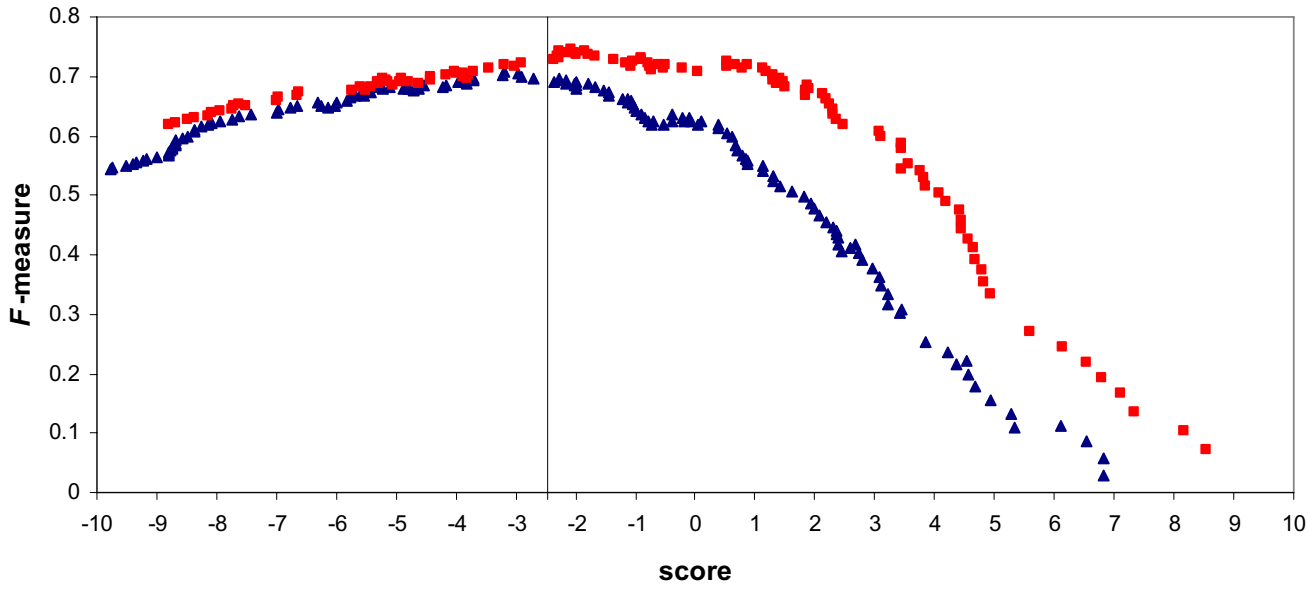


Figure 4
F-measure in function of the score of partial matching to MeSH terms with Swiss-Prot disease names (blue triangles) or OMIM terms (red squares).

sion. For this purpose, we took advantage of the independence of mappings from Swiss-Prot and OMIM, and included an additional condition: the respective mappings should point to the same MeSH descriptor in case of partial matches. Under this condition, and keeping the union of exact matches, the precision increase to 92%, with a drop in recall to 51.5%. This means that more than the half of the benchmark disease names can be mapped to MeSH with a precision above 90%. This value could be considered as sufficient to completely automate the mapping procedure.

The mappings of the benchmark, both manual and automatic, are available in additional file 1.

Automatic mapping of UniprotKB/Swiss-Prot disease comment lines

The mapping procedure was used to map the 3,408 disease comment lines present in UniProtKB/Swiss-Prot. About 76% of them had a corresponding OMIM entry. The results of the mapping are presented in Table 2 (see additional file 2 for the detailed results). Following the safe combination method described previously, we obtained a global performance of 1613 mapped terms, representing 47% of the total number of disease comment lines. The decrease in mapping coverage with OMIM terms (53% compared to 63% of the benchmark) can be explained by the higher proportion of lines having an OMIM citation in the benchmark (87%). Of course, the precision of the mapping cannot be assessed, and the results are expressed in terms of retrieval instead of recall. However, as the figures above do not differ significantly

Table 1: Evaluation of the mapping of 200 UniProtKB/Swiss-Prot disease lines (173 with a reference to OMIM)

	Exact match			Partial match			Total		
	Retrieval	*Recall*	*Precision*	*Retrieval*	*Recall*	*Precision*	*Retrieval*	*Recall*	*Precision*
SP	35(17.5%)	35(17.5%)	100.0%	91(45.5%)	73(36.5%)	80.0%	126(63%)	108(54%)	86.0%
OMIM	43(21.5%)	40(20%)	93.0%	84(42%)	68(34%)	81.0%	127(63.5%)	108(54%)	85.0%
SP ∩ OMIM	23(11.5%)	23(11.5%)	100.0%	58(29%)	51(25.5%)	88.0%	93(46.5%)	86(43%)	92.5%
SP ∪ OMIM	54(27%)	52(26%)	96.5%	95(47.5%)	76(38%)	80.0%	149(74.5%)	128(64%)	86.0%

SP: UniProtKB/Swiss-Prot
SP ∩ OMIM: both mappings correspond to the same MeSH descriptor.

Table 2: Mapping on MeSH of the 3408 UniProtKB/Swiss-Prot disease lines (2601 with a corresponding OMIM entry)

	Exact match	Partial match	Total
SP	637 (18.7%)	1332 (39%)	1969 (57.8%)
OMIM	745 (21.9%)	1063 (31.2%)	1808 (53.1%)
SP ∩ OMIM	397 (11.6%)	645 (18.9%)	1289 (37.8%)
SP ∪ OMIM	968 (28.4%)	1362 (40%)	2330 (68.4%)

SP: UniProtKB/Swiss-Prot
SP ∩ OMIM: both mappings correspond to the same MeSH descriptor.

from the benchmark, it is likely that the performance is comparable.

As a first assessment, we checked if, in case of exact matches, corresponding Swiss-Prot and OMIM terms mapped to identical MeSH descriptors. This statement was confirmed in all but 17 cases. These discrepancies in descriptor matching were mainly due to differences in classification, with OMIM synonyms corresponding to distinct descriptors in MeSH. Another minor cause was the mention of multiple diseases in the UniProtKB/Swiss-Prot comment line. In these cases, the disease name with an OMIM reference was different from the one extracted.

Discussion

In this study, we designed a mapping procedure to link the UniProtKB/Swiss-Prot human protein entries and the corresponding OMIM entries to the MeSH disease terminology. MeSH was chosen as it is interlinked with many biomedical terminologies within the UMLS. More importantly, its intimate association with literature will provide us with a valuable means for knowledge discovery using data-mining in the future.

To derive an efficient mapping procedure, alternative methods were tested in order to evaluate the effect of term pre-processing and the use of different similarity scoring systems. It was found that these methods did not differ drastically in terms of performance. Clearly, the benchmark dataset used for evaluation could be too small to draw definite conclusions. However, the fact that MeSH includes many lexical and orthographic term variations does provide an explanation for the low benefit obtained from term normalisation. On the other hand, as both MeSH and OMIM have synonym resources, the mapping procedure should have been improved with the Ha-Thuc's method which cleverly takes into account the word frequency in a set of synonyms. It is possible that the parameters used in Ha-Thuc's program, which was initially developed for gene name entity recognition in textual documents, need to be re-adjusted to better suit the purpose of terminology mapping.

The final mapping procedure we set up by combining exact and partial matches of disease names from OMIM and Swiss-Prot was able to provide a high precision mapping for more than half of the total number of disease comment lines in UniProtKB/Swiss-Prot. Although this retrieval could be considered as low for certain applications, it should be noted that stringent conditions were chosen on purpose to provide a high quality fully automated mapping procedure. If manual curation could be solicited, we could accept a reduced precision.

Recently, the same approach was used to map diagnosis-related annotations of tumor tissue microarrays to the NCI thesaurus [25] with better results (a mapping coverage of 86% and an estimated precision of 86%). These differences in performance could be simply explained by the richness of the domain-specific NCI-T vocabulary compared to the MeSH. Indeed, one of the main problems encountered in the mapping process lay in the difference of granularity between the terminologies, with MeSH being relatively coarse-grained for genetic diseases. Therefore, one strategy to increase the performance of the system would be to allow the mapping to less specific concepts. For instance, the system should be able to map the disease name, *pyruvate dehydrogenase e3-binding protein deficiency*, to its correct parent, *pyruvate dehydrogenase complex deficiency disease*, which currently had a similarity score below the threshold value. To achieve this, one can try to improve the word weighting in order to get rid of rare words without disease-related meaning, such as *e3-binding protein* . This can be done by considering either a common English word thesaurus or a greater biomedical resource, such as the whole MEDLINE database, for the word frequency calculation. More sophisticated linguistic methods could also be applied to analyse the syntactic and semantic structure of the term. Finally, it may be worth integrating information from the MeSH terminology structure in the score calculation as such a strategy has been successfully used for categorising OMIM phenotypes using MeSH terms [26].

Apart from the direct mapping strategy, preliminary work was done to evaluate several indirect mapping strategies that exploit the textual information provided by UniProtKB/Swiss-Prot and OMIM. The first method consisted in using a generic categorizer, XMap [21], to associate Swiss-Prot diseases comment lines with a ranked set of MeSH descriptors. The preliminary results on the benchmark were not convincing (data not shown). This is in agreement with other studies using MetaMap – a similar program developed by the NLM [22] - which reported that these complex methods did not outperform simpler heuristics such as ours in categorising structured database annotations [23,24]. Nevertheless, the method could be

more efficient on longer texts such as the OMIM disease *description* fields.

The second method consisted in using the textual information from the biomedical literature cited in Swiss-Prot and OMIM. Indeed MeSH is used to index MEDLINE documents and this information can be used to find the correct term. In a preliminary attempt, all disease MeSH terms in OMIM's citations were extracted and ranked according to their frequency. The precision for the first ranked terms was found to be 57%. The result was rather promising given the fact that the method was not based on term similarity. In future developments, we may consider using this complementary method in combination with the direct mapping.

Nevertheless, the problem of MeSH granularity will hardly be completely solved by these methods. We need definitely to explore the use of other medical terminology resources, such as ICD-10 or SNOMED-CT.

Conclusions

In conclusion, this work represents the first step in standardizing the medical vocabularies in the UniProt Knowledgebase. Through this effort, we provide a bridge for the medical informatics community to explore the genomic and proteomic data present in biological databases which could be of value for disease understanding.

Methods

Extraction of disease names

In UniProtKB/Swiss-Prot, disease information related to a protein entry is expressed in free text comment lines (category 'Involvement in disease'). We proceeded by first manually establishing a list of regular expressions that indicated the presence of disease names within a Swiss-Prot comment line such as 'cause(s)', 'cause of', 'involved in', 'contribute(s) to'. The expressions are listed in the additional file 3. The extraction of complete disease names was relatively easy as they are usually located at the end of a sentence or before a conjunction or a relative clause or directly followed by a corresponding OMIM identifier.

In parallel, the fields *Title* and *Alternative titles; symbols* were extracted from the cited OMIM entries. These two fields provide the disease names in OMIM as well as a set of synonyms. For names coming from "gene and phenotype (+)" entries, both gene names and diseases names were included in the disease list.

Term pre-processing

The mapping procedure was tested with and without word normalisation. The word normalisation was done using the program *Norm* from the lexical tools provided by the NLM [16]. *Norm* removes stop words and plural forms, uninflects verbs, lowercases words etc. For the mapping without word normalisation, we simply lowercased the term components, removed punctuation signs and unspecific words such as "susceptibility to", "development of" from the disease names extracted from Swiss-Prot (see additional file 3). The word "included" which qualifies a synonym of closely related meaning was also removed from OMIM *Alternative titles*. The terms were transformed into "bags of words", without taking collocations into account, except for hyphenated words.

Mapping procedures

The extracted disease names were mapped to the MeSH terms in two successive term matching steps (Fig. 1). First, we looked for exact matches, where all words composing the name had an identical correspondent in a MeSH term and vice versa. The word order and the case were not taken in consideration. When this step failed, we looked for partial matches by calculating a similarity score which is a function of the number of words in common minus the number of words which differ. The similarity score was calculated according to the following formula:

$$S = \frac{\sum_{cw} \log_2\left(\frac{1}{freq(cw)}\right) - \sum_{ncw} \log_2\left(\frac{1}{freq(ncw)}\right)}{size(disease)}$$

Where *freq=n/N*, with n the number of occurrence of the word in all OMIM (Titles, Alternative titles), MeSH terms (disease category) and Swiss-Prot disease comment lines, and N the total number of words in these documents. *cw* and *ncw* stand for words in common and not in common, respectively, between the two mapped terms, and *size(disease)* is a normalization factor consisting of the number of words composing the disease name to be mapped.

We also calculated term similarity using the program kindly provided by Ha-Thuc and Srinivasan [18]. The implemented procedure uses a 'soft' TFIDF approach which introduces a character-based similarity between words [19,20]. In addition, it takes into account the word frequencies in a set of synonym names by increasing the TF scores of words that are common to several synonyms of a disease name.

Mapping evaluation

In order to evaluate the mapping procedure, 200 disease comment lines from 95 UniProtKB/Swiss-Prot entries were manually mapped to MeSH by a medical expert. Swiss-Prot entries were selected randomly. However, care was taken so that the chosen sample of entries would be representative and lead to a proportion of exact and par-

tial matches similar to that found in a preliminary mapping attempt.

The mapping procedure was assessed in terms of precision, *p=TP/(TP+FP)* and recall, *r=TP/total number of terms*, where *TP* is the number of correct mapping (true positive) and *FP* is the number of incorrect mapping (false positives). Since the system was forced to retain only the best match, we considered, in case of diseases manually mapped to several MeSH terms, that the automatic mapping was correct if at least one of these terms was mapped.

To estimate the performance of the system, the *F*-measure was also calculated according to this formula:

$$F_\beta = (1+\beta^2)\frac{pr}{r+\beta^2 p}$$

The β value was set to 0.5 so as to favor the precision of the mapping.

Competing interests

The authors declare that there are no competing interests.

Authors' contributions

AM developed the matching procedure and did the manual mapping. YLY participated in the study's design and helped write the manuscript. PR participated in the study's design. ALV conceived, coordinated the study and wrote the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

This file contains the manual mapping of 200 Swiss-Prot disease names to Mesh terms, and corresponding automatic mapping with scores. (html format).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-S5-S3-S1.htm]

Additional file 2

This file contains the automatic mapping of all Swiss-Prot disease names with a matching score above the threshold (html format).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-S5-S3-S2.html]

Additional file 3

This file contains the regular expressions used to extract disease names from the UniProtKB/Swiss-Prot disease comment lines (pdf format).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-S5-S3-S3.pdf]

Acknowledgements

This work was funded by the Swiss National Science Foundation (grant No 3100A0-113970). We are grateful to Viet Ha-Thuc who kindly provided us with his program. The authors also wish to thank Julien Gobeill for performing the preliminary indirect mappings using Xmap and Violaine Pillet for her comments on the manuscript.

This article has been published as part of *BMC Bioinformatics* Volume 9 Supplement 5, 2008: Proceedings of the 10th Bio-Ontologies Special Interest Group Workshop 2007. Ten years past and looking to the future. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/9?issue=S5.

References

1. **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2007, **35:**D193-D197.
2. Nelson SJ, Schopen M, Savage AG, Schulman JL, Arluk N: **The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation.** *Medinfo* 2004, **11(Pt 1):**67-69.
3. **International Statistical Classification of Diseases and Health Related Problems.** In *(The) ICD-10* Second Edition edition. *WHO Press, Geneva* .
4. Donnelly K, SNOMED-CT: **The advanced terminology and coding system for eHealth.** *Stud Health Techno Inform* 2006, **121:**79-90.
5. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology.** *Nucleic Acids Res* 2004, **32:**D267-D270.
6. **The Gene Ontology (GO) project in 2006.** *Nucleic Acids Res* 2006, **34:**D322-D326.
7. Ashburner M, Mungall CJ, Lewis SE: **Ontologies for biologists: a community model for the annotation of genomic data.** *Cold Spring Harbor Symp Quant Biol* 2003:227-236.
8. **UMLS Lexical Tools.** . [http://www.nlm.nih.gov/research/umls/tools.html].
9. Sarkar IN, Cantor MN, Gelman R, Hartel F, Lussier YA: **Linking biomedical language information and knowledge resources: GO and UMLS.** *Pac Symp Biocomput* 2003:439-450.
10. Cantor MN, Sarkar IN, Gelman R, Hartel F, Bodenreider O, Lussier YA: **An evaluation of hybrid methods for matching biomedical terminologies: Mapping the Gene Ontology to the UMLS.** *Stud Health Technol Inform* 2003, **95:**62-67.
11. Zhang S, Mork P, Bodenreider O, Bernstein PA: **Comparing two approaches for aligning representations of anatomy.** *Artif Intell Med* 2007, **39:**227-236.
12. Lussier YA, Li J: **Terminological mapping for high throughput comparative biology of phenotypes.** *Pac Symp Biocomput* 2004:202-213.
13. Cantor MN, Sarkar IN, Bodenreider O, Lussier YA: **GenesTrace: Phenomic knowledge discovery via structured terminology.** *Pac Symp Biocomput* 2005:103-114.
14. Johnson HL, Cohen KB, Baumgartner WA, Lu Z, Bada M, Kester T, Kim H, Hunter L: **Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies.** *Pac Symp Biocomput* 2006:28-39.
15. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33:**D514-517.
16. **The Specialist Lexical Tools.** . [http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html]
17. Shatkay H: **Hairpins in a bookstacks: Information retrieval from biomedical text.** *Brief Bioinform* 2005, **6:**222-38.
18. Ha-Thuc V, Srinivasan P: **Exploiting synonym relationships in biomedical named entity matching.** In *BioLINK SIG 2007, ISMB/ECCB Vienna*; 2007. July
19. Bilenko M, Mooney R, Cohen W, Ravikumar P, Fienberg S: **Adaptive name matching in information integration.** *IEEE Intellig Sys.* 2003, **18:**16-23.
20. Cohen W, Ravikumar P, Fienberg S: **A comparison of string distance metrics. for name-matching tasks.** *Proc JCCAI Conf* 2003:73-78.

21. Ruch P: **Automatic assignment of biomedical categories: toward a generic approach.** *Bioinformatics* 2006, **22:**658-664.
22. Aronson AR: **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.** *AMIA Annu SympProc* 2001:17-21.
23. Butte AJ, Kohane IS: **Creation and implications of a phenome-genome network.** *Nat Biotechnol* 2006, **24:**55-62.
24. Butte AJ, Chen R: **Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics.** *AMIA Annu SympProc* 2006:106-110.
25. Shah NH, Rubin DL, Espinosa I, Montgomery K, Musen MA: **Annotation and query of tissue microarray data using the NCI Thesaurus.** *BMC Bioinformatics* 2007, **8:**296.
26. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA: **A text-mining analysis of the human phenome.** *Eur J Hum Genet* 2006, **14:**535-542.

2.3 Mapping availability through the SwissVar website

We developed SwissVar (swissvar.expasy.org) to offer a web access to protein entries in UniProtKB/SwissProt and variant pages through disease, gene or variant requests that can be combined (Mottaz *et al*., 2010).

The disease query can be made either from a MeSH term, an OMIM or a Swiss-Prot disease, proposed through an autocomplete functionality. Identifiers of these resources can also be used. Moreover, MeSH terms can be browsed through an integrated MeSH tree browser, indicating for each category the number of proteins implicated in the selected disease and children. For example users can easily visualize how many proteins are implicated in an endocrine disorder and how it is divided among children such as dwarfism, thyroid diseases etc. The query retrieves all the proteins annotated with the selected disease and children, along with the variants associated to the disease. But some proteins may be related to diseases with no associated variants. To avoid retrieving these proteins, an option can be selected named 'Only proteins having variants related to the disease'.

Queries can also be performed using gene and protein names or identifiers. It takes advantage of a database of gene and protein synonyms, GPSDB, populated from 14 different resources including non-human (Pillet *et al.*, 2005).

Variants can be filtered according to specifications such as properties of implicated amino acids. Properties include what amino-acid is substituted or the substituent, or if it is hydrophobic or polar. They can also be filtered according to the probability of substitution, using the Blosum62 matrix (Henikoff & Henikoff, 1992). Sequence proximity to a feature can also be indicated. Features correspond to Swiss-Prot sequence annotation features (Table 7), such as active site, PTM or metal binding site. Proximity in the three-dimensional (3D) space, in ångström, can also be calculated for proteins whose 3D structure has been determined or predicted by modeling approaches. Variants queries were made possible thanks to a previous work that mapped the UniProtKB protein sequences to the corresponding 3D structures at the level of the residue, thus allowing to calculate the spatial distances between the variant position and other amino acids (David & Yip, 2008). Variants can also be filtered according to their germinal or somatic origin.

A general query is also proposed, enabling in one field to query through either disease gene or variants and allowing partial matches of terms.

Importantly, the different queries can be combined. This can be especially useful for understanding the deleterious effect of variants that depends on the arrangement of amino acid in the 3D space, or on proximity to a feature such as a PTM. For example one can query all variants close to a metal binding site implicated in any brain metabolic disease (see Supplementary figure 1, Mottaz *et al.*, 2010).

Table 7. Swiss-Prot features used to query variants for sequence or 3D proximity.

Active site
Alternative sequence
Binding site
Calcium binding
Cross-link
Disulfide bound
DNA binding
Domain
Glycosylation
Lipidation
Metal binding
Modified residue
Motif
Mutagenesis
Nucleotide binding
Zinc finger

The result of the query includes protein accession number and name, disease name as extracted from Swiss-Prot disease annotation line, the three letter code variant description according to HGVS containing the wild type amino acid, the position in the Swiss-Prot canonical sequence and the substituting amino acid. HGVS is the human genome variation society that edits standards for the nomenclature of sequence variant description (Den Dunnen *et al.*, 2000). When available, the position of the variant on the 3D structure along with the references to the 3D structure in PDB. A query result can be seen in the Supplementary figure 2, Mottaz *et al.*, 2010.

The interface gives also access to the variant pages that have been created to present a summary of available information on variants present in UniProtKB/Swiss-Prot, such as residue change and physico-chemical properties of the amino acids, involvement in disease and sequence annotations around the variant residue (Yip et al., 2004). An example of variant page can be found in the Supplementary figure 3, Mottaz *et al.*, 2010.

Results can be downloaded in XML or tab delimited format. Programmatic access is also possible through URI with appropriate parameters (see *SwissVar documentation page* in the *Supplementary material*).

The web html pages are dynamically generated through a Common Gateway Interface (CGI), executing Perl scripts requesting information from the postgreSQL databases, representing three-tier architecture. Indeed the user interface, the functional unit and the data storage are separated entities. The functional unit is composed of a module that prepares the result of the request based on the data retrieved from the database through other modules specific for each resource.

The data queried from the database, including the automatic mapping to MeSH described in the 'Mapping procedure' section of this chapter, are updated every four weeks in synchronization with each UniProtKB/Swiss-Prot release.

BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 6 2010, pages 851–852
doi:10.1093/bioinformatics/btq028

Databases and ontologies

Advance Access publication January 26, 2010

Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar

Anaïs Mottaz[1,2,*], Fabrice P.A. David[1,2], Anne-Lise Veuthey[1] and Yum L. Yip[1,2,*]

[1]Swiss Institute of Bioinformatics and [2]Department of Structural Biology and Bioinformatics, Centre Médical Universitaire, 1, rue Michel-Servet, 1211 Geneva 4, Switzerland

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: The SwissVar portal provides access to a comprehensive collection of single amino acid polymorphisms and diseases in the UniProtKB/Swiss-Prot database via a unique search engine. In particular, it gives direct access to the newly improved Swiss-Prot variant pages. The key strength of this portal is that it provides a possibility to query for similar diseases, as well as the underlying protein products and the molecular details of each variant. In the context of the recently proposed molecular view on diseases, the SwissVar portal should be in a unique position to provide valuable information for researchers and to advance research in this area.

Availability: The SwissVar portal is available at www.expasy.org/swissvar

Contact: anais.mottaz@isb-sib.ch; lina.yip@isb-sib.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 24, 2009; revised on January 18, 2010; accepted on January 19, 2010

1 INTRODUCTION

Human variation data is one of the most valuable information originating from the Human Genome Project (HGP). The current challenge is how to optimally exploit this data to better understand disease association and accelerate the pace towards personalized treatments. Indeed, there are still numerous unanswered questions on the exact relationship between genetic variations, phenotypes and diseases. A plethora of databases or prediction tools exist (Thusberg *et al.*, 2009). Among the databases, only few are central databases covering mutations on all genes. They are mostly gene-centric, with little information related to the proteome. The disease and phenotype information are also currently unstructured, making specific queries difficult. This is a pity, particularly in the context of the recently proposed molecular view on diseases, which emphasizes the relationship between the disease/phenotypic networks and the underlying protein interaction or functional networks (Lage *et al.*, 2007; Oti *et al.*, 2008). Indeed, the possibility to query for similar diseases, as well as the underlying protein products and the molecular details of each variant might prove extremely useful for researchers to study a particular family of disorders or to formulate hypotheses for further research.

*To whom correspondence should be addressed.

In this article, we present the SwissVar portal (www.expasy.org/swissvar), which provides access to a comprehensive collection of single amino acid polymorphisms (SAPs) and diseases in the UniProtKB/Swiss-Prot knowledgebase via a unique search engine. This represents nearly 3300 diseases and 60 000 human protein variations (release 57.10) (Yip *et al.*, 2008). In addition, SwissVar gives direct access to the newly improved Swiss-Prot variant pages that are widely cited by the community but can not be queried, up to now.

2 IMPLEMENTATION

SwissVar accesses two relational databases that store data on variants and diseases. The database UniMed contains disease information extracted from UniProtKB/SwissProt and their mapping to MeSH terms (Mottaz *et al.*, 2008). The variant data is found in the ModSNP database (Yip *et al.*, 2004). Structural information is calculated through SSMAP, a residue–residue mapping of Protein Data Bank (PDB) structures (David *et al.*, 2008). The databases are implemented in PostgreSQL 8.1.9 and are updated at each UniProt release.

The system implementation is based on a three-tier architecture. CGI programs written in Perl query the databases and dynamically generate the web pages. The interface is accessible with the main web browsers.

3 FEATURES

3.1 Query options

Three main search categories are provided: (i) by diseases, (ii) by gene/protein names and (iii) by variant types or functional/structural features.

Query by disease terms enable search using disease names, OMIM identifiers or MeSH terms of the disease category. This query is powerful in that it exploits the mapping between Swiss-Prot disease names and MeSH terms (Mottaz *et al.*, 2008), as well as the hierarchy in MeSH to assemble groups of diseases to a granularity defined by users. For example, the users can query for all proteins related to metabolism diseases, and gather in one click proteins and variants related to refsum disease, gout etc. The representation of the MeSH hierarchy further enables the visualization and navigation inside the categories of diseases in which the queried proteins are implicated.

The second axis of query is protein centric. Users can search with a protein or gene name, as well as Swiss-Prot identifiers (AC or ID). Queries with gene names are automatically normalized using a list of

synonyms. This option could be particularly useful when analyzing gene or protein expression data.

Finally, variants recorded in Swiss-Prot/UniProtKB can be searched by their molecular characteristics. Several attributes of the amino acid concerned by the mutation can be specified, e.g. the conservation score of the residue, its surrounding environment (both sequential and structural), its surface accessibility as well as its involvement in interfaces are all adjustable parameters. The variants can also be queried using Swiss-Prot feature identifier (FTID), dbSNP rsID, the position of the mutation or the type of amino acid change.

The combination of all search parameters is possible. This combination strongly enhances the query power and the information content of the tool. For example, it is possible to retrieve all variants implicated in metabolic brain diseases, which are within 4 Å of a metal binding site (Supplementary Fig. 1).

3.2 Result pages

The result of the search is presented in a table (Supplementary Fig. 2), from which the users can have direct access to the original UniProtKB/Swiss-Prot entry, the MeSH descriptor data, the Swiss-Prot variant pages and the mapped PDB structure when available. The Swiss-Prot variant pages concisely present a complete outline of known information on each variant (Supplementary Figs 3 and 4). They were recently improved by newly added features which include the display of conservation score of the mutated residue at sequence and structural level; the display of protein features in the local structural environment of the variant (e.g. residues involved in ligand binding or post-translational modifications) as well as residues involved in protein–protein interaction when experimentally resolved 3D information is available. It is hoped that these information will further aid the users in understanding or evaluating the potential functional effect of SAPs. New articles on variants automatically retrieved through text-mining methods are also proposed on the pages (Yip *et al*., 2007).

Results can be downloaded as lists (e.g. a list of the protein accession numbers, a list of variant FTIDs or rsID) or in a tab-delimited or XML format containing all the information.

4 DISCUSSION

With the completion of the Human proteome, the UniProtKB/Swiss-Prot database has a complete collection of 20 330 human proteins with increasingly detailed functional annotation (The UniProt Consortium, 2009). The SwissVar portal gives access to this wealth of data by further providing the possibility to gather proteins/variants related to similar diseases, and allowing queries on variants using a range of sequence and structural parameters.

Further improvement of the portal and the information content is planned. First, data coverage: the current SAPs coverage is clearly not exhaustive. However, as a partner of the GEN2PHEN consortium (www.gen2phen.org), it is anticipated that data related to SAPs from consortium members will be made visible via UniProtKB and the Swiss-Prot variants pages. As such, the SwissVar portal will continue to gain its value as the amount of data grows. Second, disease terminology/phenotype information: the portal currently relies on MeSH classification that offers a reasonably broad coverage of diseases including genetic diseases. The classification is nevertheless not entirely based on phenotypic similarities. Incorporating comprehensive structured phenotype information could enhance the disease query. New resources, such as Human Phenotype Ontology (Robinson *et al*., 2009), are currently being studied for this purpose. Finally, it is planned that pathway information will be incorporated in the near future to allow seamless integration and search between diseases, phenotypes, pathways and detailed sequence and structural information of the variants.

5 CONCLUSION

In summary, the SwissVar portal provides a unique environment and search facility to investigate the relationship between human variants and phenotypes, with a particular focus on human proteome. To the knowledge of the authors, no online servers offer this kind of search possibilities that directly link molecular details of SAPs to disease classification. The current application also illustrates our ongoing effort in bridging biological and medical information. The SwissVar portal can be accessed via www.expasy.org/swissvar.

ACKNOWLEDGEMENTS

We would like to acknowledge Harris Procopiou, Gregory Loichot and Nathalie Lachenal who have contributed to the development of the Swiss-Prot variant pages.

Funding: Swiss National Science Foundation (3100A0-113970); European Community's Seventh Framework Programme under grant agreement 200754 (the GEN2PHEN project).

Conflict of Interest: none declared.

REFERENCES

David,F.P. and Yip,Y.L. (2008) SSMap: a new UniProt-PDB mapping resource for the curation of structural-related information in the UniProt/Swiss-Prot Knowledgebase. *BMC Bioinformatics*, **9**, 391.

Lage,K. *et al*. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Mottaz,A. *et al.* (2008) Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics,* **9**(Suppl. 5), S3.

Oti,M. *et al.* (2008) Phenome connections. *Trends Genet.,* **24**, 103–106.

Robinson,P.N. *et al.* (2009) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.,* **83** 610–615.

The UniProt Consortium (2009) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **37**, D169–D174.

Thusberg,J. and Vihinen,M. (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum. Mutat.,* **30**, 703–714.

Yip,Y.L. *et al.*(2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat.,* **23**, 464–470.

Yip,Y.L. *et al.* (2007) Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase. *J. Bioinform. Comput. Biol.,* **5**, 1215–1231.

Yip,Y.L. *et al.* (2008) Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum. Mutat.*, **29**, 361–366.

3. Phenotype-based PPI contextualization

Besides interoperability, the aim of mapping molecular information to a disease controlled vocabulary was to use the knowledge contained in the taxonomic relationship. Disease hierarchies are mostly based on affected anatomical sites and systems. Using the disease categories in MeSH was therefore promising to study proteins according to the clinical presentation of their defect.

To best use this information, relating characteristics of Mendelian diseases such as pleiotropy to characteristics of protein function such as modularity gives perspectives. Besides, an overview of current efforts in translational genomics is also interesting to find in which directions efforts are worth.

The conclusion of these analyzes is that distinct clinical manifestations are valuable separately as they should correspond to different molecular spatio-temporal contexts and processes, due to an overlooked consequence of modularity in proteins function. Since few efforts use clinical information to interpret molecular data, a prototype tool has been developed to prioritize protein/protein interactions with single clinical manifestations of Mendelian disorders. A case study is proposed to apprehend the potential of such approaches for further exploration.

3.1 The modular nature of protein function

One important aspect that has to be considered in the understanding of protein function and how it relates to diseases is the dynamic of cellular functioning. Indeed, space and time consideration is essential because of the modular nature of cell biology (Hartwell *et al.*, 1999). A modular system is defined by entities with specific functions that, depending on which other entity it is combined with, can serve different, more global, functionalities. This can be observed at many different levels in biology and in particular when proteins are combined into different complexes and functional units, themselves combined, or integrated, into more general processes. Therefore the global functionality of a protein depends on the cellular state, or spatio-temporal context, affecting the presence of interacting partners, post-translational modifications (Lin *et al.*, 2007), or state of other pathways (Natarajan *et al.*, 2006).

This organization enables the control of sophisticated behaviors of many different cell types with only a few thousand genes, much less than all biological roles (Pawson & Nash, 2000). It may also facilitate evolution by allowing changes in the function of a cell through alteration in the connections between the modules instead of modifying all components of a process.

Also, it can help interpret the genotype to phenotype relationship (see *Introduction* chapter, *DNA variation and diseases* section for an introduction to the genotype to phenotype relationship) through two main consequences. First, a protein is implicated together with other proteins in a given process. Second, a protein with one given

molecular function can be implicated in different processes depending on the context such as which other proteins it interacts with.

Locus heterogeneity for example can reflect the fact that mutations in any of several proteins implicated together in a functional module are responsible for the same phenotype. Indeed, as described later, proteins interacting together tend to be implicated in similar diseases.

Regarding pleiotropy, if one protein affects different processes at different times and places, as predicted by the second consequence of modularity mentioned above, the result of a mutation should be a multisystemic involvement as observed in Mendelian diseases. Confirming this interpretation, the degree of pleiotropy of a gene has been correlated to the number of interactions its coding protein has and to the number of biological processes it is implicated in, but not to the number of different domains it has or molecular functions (He & Zhang, 2006; Su *et al.*, 2010). Therefore pleiotropy starts to be considered as a consequence of modularity (Wagner & Zhang, 2011). This could explain also why nearly all genes display a certain and limited degree of pleiotropy. Indeed most genes are estimated to affect around seven traits (Stearns, 2010).

Also, modifier proteins responsible for clinical heterogeneity regarding isolated traits should be related to a common global process rather than to precisely the same functional unit or pathway, that seems to be the case (Genin *et al.*, 2008).

To study the impact of this understanding on the dialog between genes and phenotypes through Mendelian diseases, current efforts in the translational genomic domain should be considered first.

3.2 Current efforts in translational genomics

Disease gene prediction

Various approaches have already been investigated to predict new disease genes (Moreau & Tranchevent, 2012; Bromberg, 2013). They are used to prioritize both Mendelian and complex disease genes. Many are based on the observation that similar Mendelian diseases are caused by mutations in proteins with similar function.

One of the first demonstrations of this correlation used the description and clinical synopsis from the OMIM database. Each OMIM entry was represented by a vector of MeSH disease and anatomy concepts weighted by their information content and the similarity was estimated with their cosine coefficient. A positive correlation was found between the disease similarity measure and functional similarity indices of associated proteins, such as sequence similarity, number of common annotated GO terms and probability to interact (van Driel *et al.*, 2006; Gandhi *et al.*, 2006).

Functional approaches

Functional approaches use functional similarity between known disease genes and often use sequence similarity, GO terms similarity, common domains and protein interactions (Turner *et al.*, 2003; Oti *et al.*, 2006; Perez-

Iratxeta *et al.*, 2007; Schlicker *et al.*, 2010; Franke *et al.*, 2006). Extending these approaches to more indirect indices of protein functional similarity, the use of gene co-expression (Adie *et al.*, 2006) tissue expression specificity (Tiffin *et al.*, 2005) and implication in similar pathways (Aerts *et al.*, 2006; Franke *et al.*, 2006; George *et al.*, 2006) were also proposed.

Sequence approaches

Global sequence properties of proteins implicated in Mendelian diseases have also been used, such as their tendency to be longer, the presence of more homologs in distant species and fewer highly conserved paralogs in the human genome (López-Bigas & Ouzounis, 2004).

Global network approaches

Attempts to prioritize new disease proteins have been made using the properties of the protein-protein interaction (PPI) networks (Gonzalez & Kann, 2012). PPI networks have been extensively studied in yeasts, revealing clusters of interconnected proteins as well as hub proteins with a high number of connections (Ideker & Sharan, 2008). This configuration enables the network to have small-world property, making each protein close to any other one in term of interactions even within a large network, while a random node deletion has few chances to affect this property, making it quite robust (Barabási & Oltvai, 2004).

Degree property

Therefore, an important global property of proteins in a PPI network is the degree of connectivity, or number of interactions. This property has been studied in proteins implicated in diseases compared to other proteins. Depending on the studies, different observations were made. Cancer related genes, either differentially expressed (Wachi *et al.*, 2005) or mutated (Jonsson & Bates, 2006) in cancer tissues were found to have more connections than other proteins. However, when studied with Mendelian diseases, an intermediate connectivity degree was associated to proteins while a high connectivity degree was more correlated to essential genes (Goh *et al.*, 2007; Feldman *et al.*, 2008), essentiality being defined as the existence of a mouse orthologous gene whose disruption results in embryonic or postnatal lethality. Disease gene prioritizers have been developed based on connectivity degree and other measures of centrality (Ortutay & Vihinen, 2009).

Cluster property

Another important property of proteins in the PPI network is the tendency to form cluster, or to interact with proteins that also interact with each other. These clusters of proteins are implicated in similar cellular function and correspond to either protein complexes or dynamic functional units (Spirin & Mirny, 2003). Considering the demonstrated correlation between protein functional similarity, including high probability to interact, and

associated disorder similarity (van Driel *et al.*, 2006), genes implicated in diseases should also form clusters when linked through disease similarities. Indeed, when creating a network based on link between genes associated to identical diseases, the genes tends to form clusters implicated in identical and similar diseases (Goh *et al.*, 2007), as determined by a manual classification of disorders based on the affected system.
Thus, the tendency of proteins in PPI networks to form clusters with other proteins implicated in phenotypically similar diseases have been used for disease gene prioritization. In one study a ranking of candidate genes for a given disease was proposed according to their protein product interaction topology with other proteins implicated in similar phenotype (Lage *et al.*, 2007). In another study, the general property of disease proteins to interact in cluster with other disease proteins was used to prioritize genes to be associated to any disease (Xu & Li, 2006).

Finally general network properties evaluated with elaborate techniques, such as random walk techniques or web ranking pages techniques, have also been proposed to prioritize any disease related gene (Chen *et al.*, 2009; Erten *et al.*, 2011).

Mutation approaches

First, genes can be prioritized based on their variation, with the hypothesis that the more a variant is deleterious to the function of a protein, the more it has chances to be associated to a disease. Predictors are based on sequence features such as local sequence environment (Capriotti *et al.*, 2006) or conservation in orthologs and paralogs (Sim *et al.*, 2012) on the assumption that mutations in conserved regions have more chances to affect the protein function. Others are based on the physico-chemical properties of the amino-acids in the context of the protein three dimensional structure (Bromberg & Rost, 2007; Adzhubei *et al.*, 2010) including their predicted effect on the structure stability (Yue *et al.*, 2006).

Cross-species approaches

Associations between diseases and genes can also be transferred across species. By calculating the phenotypic similarity between different species, a known phenotype - gene association in one species can prioritize the orthologous gene for implication in similar phenotypes (Washington *et al.*, 2009). Such approach requires considerable efforts to compare cross-species information and benefits from applications mapping phenotype and anatomical ontologies between different species, such as UBERON (Mungall *et al.*, 2012) or PhenomicDB (Kahraman *et al.*, 2005). Cross-species information transfer is also useful for validation of co-expressed cluster of proteins (Ala *et al.*, 2008).

Protein function prediction

While so many approaches take advantage of the correlation between functional similarity and disease similarity to prioritize disease genes, few approaches use implication in disease to prioritize functional data. Yet clustering

genes according to the similarity of their related phenotypes in animals, found through phenomicDB, proved to be efficient to infer gene function (Groth *et al.*, 2008). Moreover, as presented in the beginning of this chapter, protein function is highly dependent on the spatio-temporal context. Therefore, finding ways to add such context to protein data is of interest. For example, GO terms corresponding to biological processes have been used to score experimental PPIs and protein-DNA interactions, to reveal context-dependant pathways in a framework that could be generalized to different contexts represented by processes (Lan *et al.*, 2013). Another approach used tissue expression in addition to biological processes to contextualize experimental PPIs around disease related protein pairs. It helped reveal phosphorylation pathways relevant for Alzheimer's disease using as context brain tissue and cell death (Schaefer *et al.*, 2013). While context can be given by clinical traits found in Mendelian diseases, no framework has been proposed, to my knowledge, using human phenotypes to add a context to protein data and help predict their function.

3.3 Prototype tool

The approach developed here is based on the consequence of two observations: the fact that phenotypes can help predict protein function and the fact that pleiotropy in Mendelian diseases is explained by the implication of proteins in different processes depending on the context.

Starting from any protein, experimentally observed PPIs are retrieved, including those obtained through high-throughput methodologies, two levels deep around the selected protein. The resulting network may then contain thousands of interactions, potentially representing all interactions that may happen in different contexts. The network is then filtered according to the implication of proteins in a given phenotype that may be encountered in different syndromes.

The aim is to extract meaningful interactions and proteins in relation to the process behind the phenotype, that may represent high level processes, since the syndromes are not selected based on their global similarity but only on the presence of one common phenotype.

Intermediate proteins not known to be implicated in the phenotype are kept in the network. This enables to consider proteins that lack such annotation, either due to incomplete annotation coverage or a yet unknown implication in phenotype, or the presence of proteins too essential to cause a viable syndrome.

The mapping to the MeSH vocabulary could have been useful for this task thanks to the taxonomical hierarchical based on affected anatomical site and systems. Unfortunately, a lot of phenotypic information about Mendelian diseases is missing in controlled vocabularies. This can be illustrated with an example, comparing the Clinical Synopsis section of OMIM with the MeSH hierarchy of the Rubinstein-Taybi syndrome (Table 8). This syndrome is described in OMIM with nearly 90 different clinical traits while it has as few as four parents in MeSH. Only the most striking pathological traits are represented, in a quite unspecific manner, such as craniofacial abnormalities summarizing microcephaly, cataract, strabismus, coloboma, heavy eyebrows, beaked nose, etc.

Table 8. OMIM Clinical synopsis compared to MeSH hierarchy in the Rubinstein-Taybi syndrome.

OMIM Clinical synopsis	MeSH parents
Short Stature Average adult male height 153 cm Average adult female height 147cm Obesity after puberty Postnatal growth retardation Microcephaly Large anterior fontanelle Late closure of fontanelle Frontal bossing Low anterior hairline Hypoplastic maxilla Micrognathia Retrognathia Grimacing or unusual smile with almost closing of the eyes Low set ears Hearing loss Recurrent otitis Heavy eyebrows Highly arched eyebrows Long eyelashes Ptosis Epicanthal folds Strabismus Nasolacrimal duct obstruction Cataracts Glaucoma Coloboma Downward slanting palpebral fissures Beaked nose Deviated nasal septum Broad nasal bridge Small opening of the mouth Narrow palate High-arched palate Dental crowding Talon cusps Crossbite Screwdriver permanent incisors Enamel hypoplasia Enamel discoloration Atrial septal defects Ventricular septal defects Patent ductus arteriosus Capillary hemangiomas Recurrent respiratory infections Sternal anomalies Constipation Hypospadias Shawl scrotum Cryporchidism Delayed skeletal maturation Joint hypermobility Large foramen magnum Parietal foramina Scoliosis Spina bifida occulta Small flared iliac winds	Dysostosis Craniofacial abnormalities Intellectual disability Multiple abnormalities

Patellar dislocation Broad thumbs with radial angulation Fifth finger clinodactyly Persistant fetal fingertip pads Syndactyly Polydactyly Single transverse palmar creases Broad great toes Plantar crease between first and second toes Pes planus Keloid formation in surgical scars Capillary hemangiomas Café-au-lait spots Hirsutism Mental retardation (average IQ 51) Agenesis of corpus callosum Severe expressive speech delay Poor coordination EEG abnormalities Seizures Hypotonia Hyperreflexia Good social contacts Short attentions span Labile mood Recurrent infections Polysaccharide antibody response defect Increased risk of tumor formation, especially of the head Increased risk of leukemia	

Many efforts have already been done to extract phenotype information from full text and clinical synopsis from OMIM entries. Our method developed in the first part of the work could be useful for such approaches using term-matching techniques. However we used the Human Phenotype Ontology, HPO, which has been developed from the clinical synopsis of OMIM; see *Medical controlled vocabularies* section in *Introduction* chapter.

The use of HPO is straightforward since it is directly mapped to OMIM (Köhler *et al.*, 2014). Therefore, taking advantage of the UniProtKB/Swiss-Prot references to OMIM and the Human Phenotype Ontology mapping to OMIM, HPO terms can be retrieved for any protein having a disease annotation with a reference to an OMIM entry mapped to HPO. It is the case for more than half of the protein with a disease annotation (Table 9).

Table 9. Proteins-OMIM-HPO statistics, UniProtKB/Swiss-Prot 2014_07

	With disease annotation	**With disease annotation referenced to OMIM**	**With disease annotation referenced to OMIM linked to HPO**
Number of proteins	3,266	2,898	1,826

Figure 11 gives a more general idea of the difference between the number of MeSH parents compared to HPO concepts, calculated for all diseases mapped to MeSH with our approach and mapped to HPO through OMIM. It is easily visible that the majority of diseases have four parents or less in MeSH while a majority is linked to more than six HPO phenotypes.

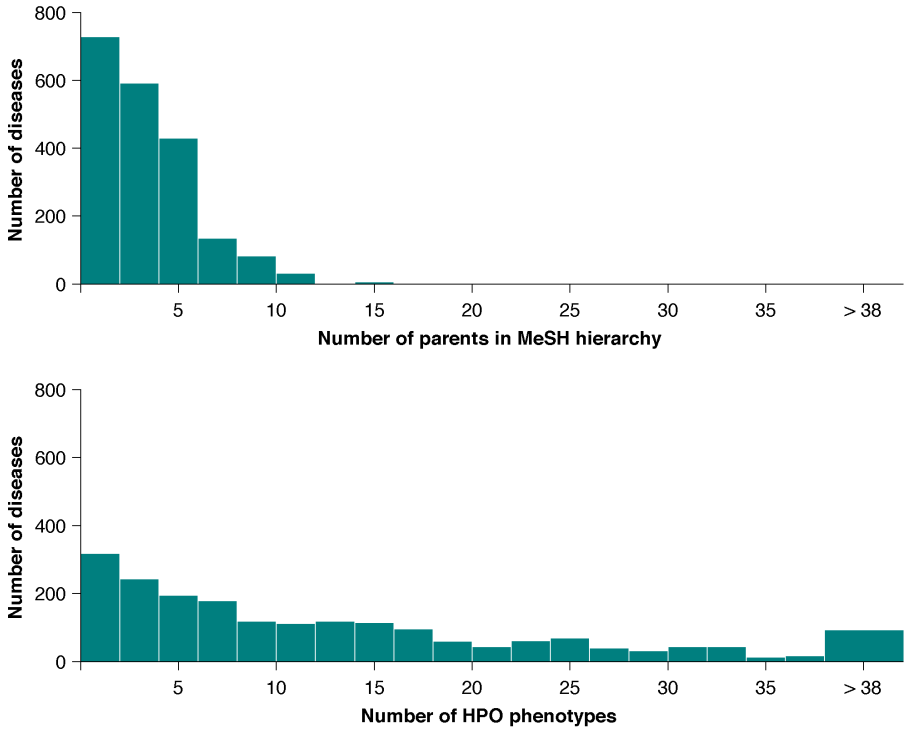


Figure 11. HPO phenotypes compared to MeSH hierarchy.

Combining the mapping between UniProtKB/Swiss-Prot and HPO, through OMIM, and the cross-references to STRING, a database of protein interactions, any phenotype present in the Human Phenotype Ontology, as varied as 'Mental retardation', 'Hypopigmentation' or 'Leukemia', can be used to filter protein interactions found in STRING.

Resources description and data extraction

UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot accession numbers and OMIM references, found in *swissprot* and *spdisease_omim* tables regularly updated for the SwissVar website, were used (see *Supplementary material*, Figure S1, and Chapter 2 *Mapping UniProtKB/Swiss-Prot to a disease controlled vocabulary*).

Human Phenotype Ontology

Description

See *Medical controlled vocabularies* in *Introduction* chapter.

Data extraction

The Human Phenotype Ontology was downloaded from the HPO website as an obo file (*human-phenotype-ontology.obo* version 1.2) and the mapping between OMIM and HPO as a tab-delimited file (*phenotype_annotation.omim*). Only the 'Organ abnormality' ontology was used.

STRING

Description

STRING (string-db.org/) is a database whose acronym stands for Search Tool for the Retrieval of Interacting Genes/Proteins and that aims to collect all reported PPIs, either known or predicted, and either direct or functional (Jensen *et al.*, 2009). It integrates data from more than 1,000 different species and transfer information across them when possible. Physical interactions are retrieved from experimental interaction databases such as BIND, DIP, GRID, HPRD, IntAct, MINT, and PID. Functional interactions are extracted from curated pathways databases such as Biocarta, BioCyc, GO, KEGG, and Reactome, but also co-expression data, genomic context such as neighborhood fusion or co-occurrence, and automatic extraction from publications using text mining techniques. Scores are attributed to evaluate the confidence of predicted interactions by benchmarking the performance of the predictions against a common reference set of trusted, true associations (von Mering *et al.*, 2005).

Data extraction

Experimental interactions were retrieved from the STRING flat file that contains protein network data and subscores for the different types of links and that is distributed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License, available from the STRING website (*protein.actions.detailed.v9.1.txt.gz*). Protein identifier were extracted as well as the score for each type of link (neighborhood, fusion, coexpression, co-occurrence, experimental, database, text-mining, combined_score). The

mapping between the protein identifiers used by STRING and the UniProtKB/Swiss-Prot AC was downloaded from the STRING website (*release.2012_1.vs.human.string.v9.1.via_blast.v1.02172012.txt*).

Gene Ontology

Description

The Gene Ontology (GO) structured controlled vocabularies is maintained by the GO consortium and is used to describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner (Consortium, 2006).

Data extraction

The Gene Ontology gene_ontology.1_2.obo was downloaded from the Gene Ontology website (www.geneontology.org).

Data storage

UniProtKB/Swiss-Prot accession numbers and OMIM references, found in the tables *swissprot* and *spdisease_omim,* were retrieved from the database regularly updated for the SwissVar website (see *Supplementary material*, Figure S1, and Chapter 2 *Mapping UniProtKB/Swiss-Prot to a disease controlled vocabulary*). Interactions were stored in the same PostgreSQL database, one table containing the interactions between proteins and associated scores, and the other the mapping between the protein identifiers used by STRING (ensp) and the UniProtKB/Swiss-Prot accession number (see *Supplementary material*, Figure S2). Concerning HPO and GO, data were loaded in the working memory through object oriented modules, see *Programming languages* below.

Programming languages

Programs were implemented with the Perl 5 programming language (www.perl.org/).
The access to the database was implemented using the DBI module (dbi.perl.org/) and for HPO and GO, the data were loaded in the working memory with the Bio::OntologyIO and accessed with the Bio::Ontology::OBOEngine (Antezana *et al.*, 2008).

Network construction

Starting from a given protein, all experimentally interacting proteins, as found in STRING and of any confidence score, were retrieved two layers deep. Proteins kept were those implicated in a given HPO concept (or any of its descendants) as well as 'intermediate layer' proteins, that is the proteins that were not implicated in the HPO concept but connecting the starting protein with the other 'HPO proteins'. The network construction was achieved with a recursive algorithm that avoided loops, calculated in a time order of seconds.

Case study

The presented network was constructed around a protein implicated in DNA repair, the Bloom (BLM) syndrome protein. It is an helicase with a role in double-strand break (DSB) repair and whose mutation predisposes to various developmental defects as well as malignancies (MIM number 210900; Orphanet number ORPHA125), in particular leukemia. The leukemia concept (HP_0001909) was chosen to study interactions and proteins potentially implicated in the process behind the predisposition to leukemia.

Starting from 1,658 proteins connected by 14,891 interactions (Table 10), we obtained after applying the phenotype filter 53 proteins in the subnetwork connected by 290 edges (*Supplementary material*, Table S1). 36% of found proteins were known to be implicated in leukemia, according to HPO phenotypes, and 34% were known to be directly implicated in double-strand break (DSB) repair, such as ATM, H2AX or BRCA1, according to UniProtKB/Swiss-Prot GO annotations (GO:0006302 and children 'is_a' and 'part_of') (Table 11). The proportion is much higher than when no filter was applied and even when any phenotype was considered. Interestingly, as seen in Table 11, the process enrichment found in the leukemia network was even more important in proteins not known to be directly implicated in leukemia (14 proteins implicated in DSB repair among 34 "non-leukemia" proteins) than in proteins known through HPO to be implicated in leukemia (4 proteins implicated in DSB repair among 19 "leukemia" proteins).

Table 10. Network features around Bloom syndrome protein according to filter criteria.

	No filter	Any HPO phenotype	Leukemia
Number of interactions	14,891	1,130	290
Number of proteins	1,658	278	53
Proportion of proteins implicated in double-strand break repair	3%	8%	34%

Table 11. Proteins found in the 'Leukemia' network around the 'Bloom syndrome protein'.
5' exonuclease Apollo Adenomatous polyposis coli protein Bloom syndrome protein* Breast cancer type 1 susceptibility protein* Caspase-3 Cellular tumor antigen p53* Chromatin assembly factor 1 subunit A CREB-binding protein Cyclin-dependent kinase inhibitor 2A, isoforms 1/2/3 DNA mismatch repair protein Mlh1* DNA mismatch repair protein Msh2* DNA mismatch repair protein Msh6 DNA repair endonuclease XPF* DNA repair protein complementing XP-G cells DNA repair protein RAD50* DNA repair protein RAD51 homolog 1* DNA repair protein RAD52 homolog* DNA topoisomerase 1 DNA topoisomerase 2-alpha DNA topoisomerase 2-beta DNA topoisomerase 3-alpha Double-strand break repair protein MRE11A* Exonuclease 1 Fanconi anemia group A protein Fanconi anemia group C protein Fanconi anemia group D2 protein Fanconi anemia group E protein Fanconi anemia group M protein Flap endonuclease 1* H/ACA ribonucleoprotein complex subunit 4 Histone H2AX* Interferon-induced GTP-binding protein Mx1 Mast/stem cell growth factor receptor Kit Meiotic recombination protein DMC1/LIM15 homolog Mismatch repair endonuclease PMS2 Mitotic checkpoint serine/threonine-protein kinase BUB1 beta Mitotic spindle assembly checkpoint protein MAD2A Nibrin* RecQ-mediated genome instability protein 1 Replication factor C subunit 1 Replication protein A 32 kDa subunit* Replication protein A 70 kDa DNA-binding subunit* Retinoblastoma-associated protein Serine-protein kinase ATM* Serine/threonine-protein kinase Chk1 Structural maintenance of chromosomes protein 1A Telomeric repeat-binding factor 1 Telomeric repeat-binding factor 2 TFIIH basal transcription factor complex helicase XPD subunit Tumor suppressor p53-binding protein 1* Tyrosine-protein kinase JAK2 WD repeat-containing protein 48 Werner syndrome ATP-dependent helicase*
Legend: *Brown: Implicated in leukemia (according to HPO).* *Orange: Implicated in any disease (according to UniProtKB/Swiss-Prot disease annotation).* **: Implicated in DSB repair (according to UniProtKB/Swiss-Prot GO annotations).*

The link between leukemia and DSB repair is known through association between mutation in proteins implicated in DSB repair and predisposition to leukemia, as well as leukemia following cancer therapies inducing DSB (Casorelli *et al.*, 2012). The precise mechanism is however not yet fully understood.

In the subnetwork, we found the CREB-binding protein (CREBBP), a histone acetylase whose mutation leads to the Rubinstein-Taybi syndrome, which among many other features predisposes to leukemia (MIM number 180849; Orphanet number ORPHA783).
The CREB-binding protein did not interact directly with the Bloom syndrome protein and was not known to be directly implicated in DSB repair but as represented in Figure 12:

- Its participation in this process is being investigated (Ogiwara *et al.*, 2011), as is histone acetylation (Vempati et al., 2010) and more generally chromatin modification (Liu *et al.*, 2013; Karagiannis & El-Osta, 2006).
- A functional interaction with BRCA1 had been described in 2000 in a context not directly related to DSB repair (Pao *et al.*, 2000). This particular transcriptional activation of BRCA1 by the CREB-binding protein had been finally described in the DSB repair context in 2012 (Ogiwara & Kohno, 2012).

Moreover, the protein linking the CREBBP with the BLM protein is BRCA1 that is not directly known to be implicated in leukemia but is suspected to have a role in it (Friedenson, 2007). The interactions provided here could be of interest for the exploration of the role of BRCA1 in the pathogenesis of leukemia. Moreover, it could also help understand why mutations in the CREB-binding protein modifies the response to leukemia treatment (Mullighan *et al.*, 2011) and why histone deacetylase inhibition work as a therapeutic agent for the treatment of leukemia (Fredly *et al.*, 2013).

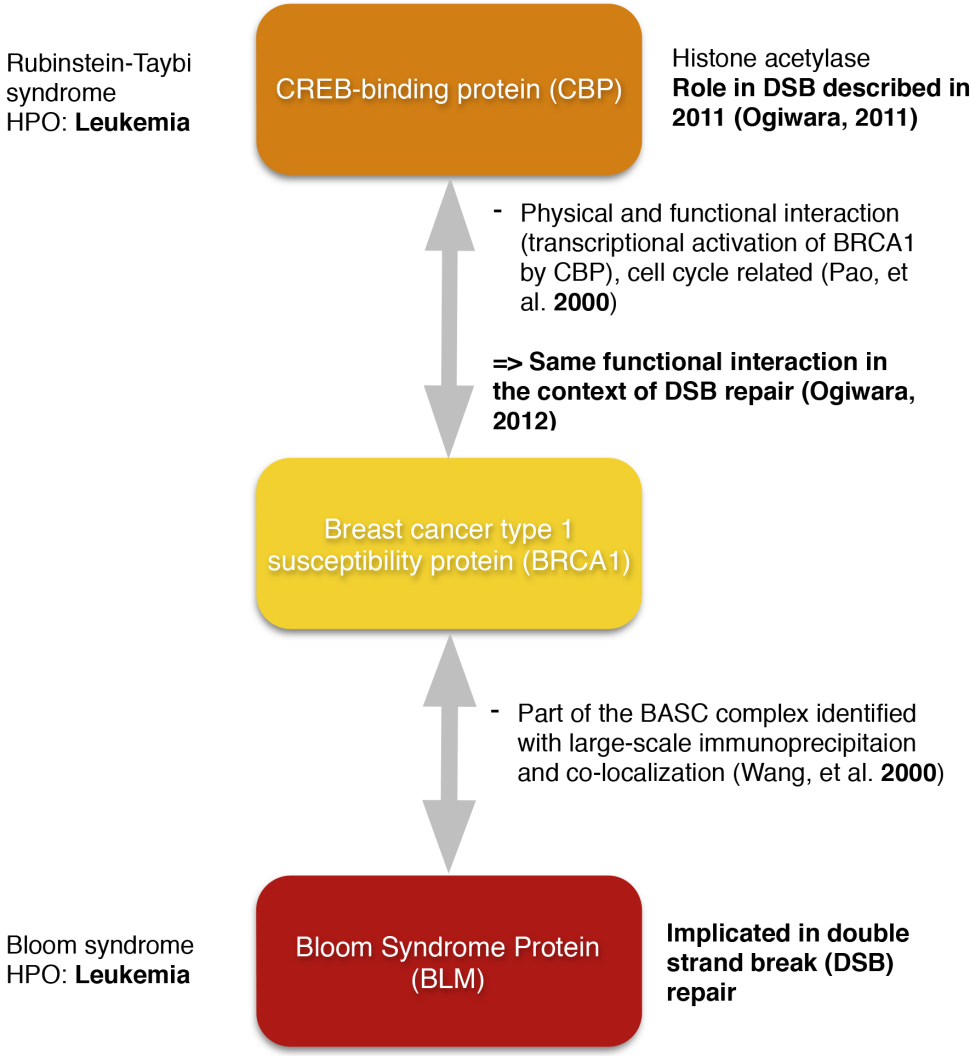


Figure 12. Example of predicted implication in process of protein and interaction found in the 'leukemia' network around the 'Bloom syndrome protein'.

4. Discussion and perspectives

Diseases, especially Mendelian, are highly valuable for the exploration of the link between the genotype and the phenotype. This link is important to establish a dialog between fundamental research and clinical applications and to implement translational genomics. But to fully exploit this relationship, it is essential that molecular and clinical resources be interoperable. Besides technical issues such as lack of compatibility between file formats or legal issues in term of data sharing, semantic interoperability is a key element that depends on the use of semantic standards, such as controlled vocabularies. Unfortunately in the life sciences, a large fraction of the current information has been captured as unstructured textual data. A preliminary step consists then in mapping existing data with controlled vocabularies. The aim of the presented work was therefore to increase interoperability between molecular data and related disease information through the mapping of a protein resource to a medical vocabulary.

The programs developed here enabled the automatic mapping of UniProtKB/Swiss-Prot, a central protein resource, to MeSH, a medical vocabulary used to index literature. This provides a direct link between a central molecular resource and medical information found in published literature, and potentially other resources using the same vocabulary. However it does not provide a direct access to clinical data. Indeed, ICD-10, which is used to code medical records in Geneva University Hospital, was also tested for the mapping but its coverage was too low for this purpose. At least two reasons can be found. Firstly, many Mendelian disorders affect only a few individuals world-wide. Secondly, they are rarely in themselves a reason for clinical care. Indeed, few specific treatments exist yet. People with such disorders are therefore treated for pathologies associated with their syndromes, such as diabetes or congenital heart defect. If nevertheless they had to be mapped to a clinical vocabulary, either directly or through UMLS, SNOMED-CT would probably be the best choice since it is the most extensive clinical resource.

The similarity score that we designed and that sums similar tokens from disease terms and subtracts different ones, weighted by a TFIDF related measure, could have been improved with synonyms and partial string matches such as the score developed by Ha-Thuc (Ha-Thuc & Srinivasan, 2007). However, this score did not yield better results, perhaps because our score was more fitted to our benchmark but more probably because the main issue in the lack of coverage was the relatively coarse granularity of the hierarchy in MeSH.

The mapping module that we developed can be independently reused. It has for example been employed to map tissue expression sites to MeSH anatomy terms for the creation of a tissue expression resource (Duek *et al.*, 2011). It can also be used to map terms to any other given vocabulary.

Our approach uses fully automated procedures. They are fast but even if very good, their precision is not perfect. The question of automatic versus manual expert mapping is important. Combining both is attractive. Automatic mapping for example can be a first step before a manual review. Also it can be useful for the maintenance of a mapping, by automatically warning for better matches in updated vocabularies. UniProtKB/Swiss-Prot disease vocabulary for example was meanwhile manually mapped to MeSH (http://www.uniprot.org/docs/humdisease) and the automatic mapping procedure could be used to warn for better mapping to newer MeSH versions.

The enhanced access to the literature provided by the mapping could be used to retrieve clinical data about Mendelian diseases in literature. Since MeSH is used to index MEDLINE articles there would be no need to parse all the texts searching for different synonyms to select publications of interest but only to query the MEDLINE database with an identifier. For example it could easily detect cooccurence of diseases in published literature to infer functional relation between proteins. Also, by taking advantage of the OMIM – MeSH mapping, new clinical traits associated to Mendelian diseases could be discovered in retrieved articles by mapping them to HPO and comparing them to already mapped phenotypes. It has been demonstrated indeed that phenotypic data about Mendelian diseases still lack coverage especially in HPO and Orphanet (Oti *et al.*, 2009).
Another advantage of the mapping to MeSH besides direct link to literature is the value added by its synonyms and hierarchy. Both are used for accessing proteins and variants through the SwissVar website, which enables to easily query variants implicated in disease categories combined with variant sequence or three-dimensional features.

It appeared rapidly yet that the hierarchy in MeSH was not representative of the wealth of information concerning Mendelian diseases. Indeed, a majority of them affect several traits, in relation to gene pleiotropy, difficult to represent in a hierarchy. For example the more recent Disease Ontology seems to have no better and even poorer hierarchy than MeSH concerning Mendelian diseases.
Recent accumulation of indices suggests that pleiotropy is mainly explained by the implication of a protein in several biological processes rather than proteins having several molecular functions. In parallel, protein functioning has been described as modular. However, modularity is in this context mostly interpreted as several proteins interacting for one given function and hardly ever as one protein implicated in several biological processes. Considering this aspect of modularity enables yet to interpret pleiotropy and has been acknowledged quite recently (Wagner & Zhang, 2011). This interpretation of pleiotropy has been chosen here to study further the use of Mendelian disorders in translational genomics.

Few efforts have been done using diseases to help study protein function. Considering separately phenotypic traits found in Mendelian disorders can theoretically be of great help. Indeed if the different traits are the consequence of separate processes that depend on specific interactions, each trait should enable to isolate spatio-temporal contexts and process specific interactions. Such spatio-temporal context filtering is highly needed especially for protein-protein interaction network interpretation. Moreover, interactions integrating specific

functional modules into higher level processes should be retrieved with such approach, particularly when considering different diseases having few phenotypes in common (Wilson *et al.*, 2011). These interactions are of high interest because they are essential for pathway cross-talk (Lu *et al.*, 2005).

In the third chapter, *Phenotype-based PPI contextualization*, a prototype tool filtering PPIs through clinical traits found in Mendelian diseases is presented. It uses HPO to filter interactions retrieved from STRING, taking advantage of the mapping between HPO and OMIM and the cross-references from UniProtKB/Swiss-Prot to OMIM and to STRING. It enables to retrieve any interaction, including those retrieved in other species or with high-throughput techniques, connecting two proteins implicated in one common clinical trait, either directly or through intermediate protein. It meant to illustrate that filtering nearly 15,000 interactions with one single phenotype could help predict new implications in the biological process behind the phenotype and associated interactions.

In the case study, the process and phenotype of interest have been chosen according to a previous knowledge about the association between the process, DSB repair, and the phenotype, leukemia. An alternative approach would be to search for any biological process enriched in the proteins implicated in a given phenotype.
At least one protein, the CREB-binding protein, and associated interaction, with the Breast cancer type 1 susceptibility protein, has been confirmed to be implicated in DSB repair and would have been found in the network before this confirmation (Ogiwara & Kohno, 2012). Importantly, far from all interactions have been investigated.
Systematic evaluations are necessary to validate this approach. A first assessment could consist in creating a network containing interactions found before a given date, and see the proportion of confirmed interactions since, compared to unfiltered network for example. Another approach could consist in testing the interactions in the predicted context in laboratory. Comparing networks obtained through single phenotypes with networks obtained through global disease similarity could be interesting as well. Single phenotypes should retrieve interactions between higher level biological process modules, which could be estimated with the number of interactions with hub proteins.
The choice of starting from a given protein was done with the thought of a manual case study. To obtain a similar global network, it should contain any protein implicated in a given phenotype as well as proteins interacting with at least two proteins implicated in the phenotype.
Of course, predicting process implication only from phenotype does not require interaction information. However, it seems that using it is worth since the enrichment in the biological process, DSB repair, not only concerned proteins directly implicated in the phenotype, leukemia, but mainly ‘intermediate’ proteins.

Data retrieved with this tool may seem obvious. But the difficulty resides in putting them together. For example, the CREBBP implication in leukemia is clear here. This information is however not as visible as it seems. The Rubinstein-Taybi syndrome in which it is found is classified as a craniofacial abnormality with mental

retardation and not a cancer syndrome in MeSH for example. A tool putting together this information with the information that the Bloom syndrome protein also predisposes to leukemia and how both proteins are related does not exist.
Importantly, this approach can be applied to any phenotype present in HPO with the result being available very rapidly (order of seconds).

Studying genes predisposing to cancer, as in the case study, may seem useless since somatic driver mutations are studied with whole genome sequencing of cancer tissue. But processes predisposing to cancer are not exactly equivalent to processes necessary for cancer. Only 10% of driver genes are known to predispose to cancer and only 40% of genes predisposing to cancer are known driver genes (Rahman, 2014). The majority thus do not overlap. Moreover even when predisposing and driver genes overlap, the type of cancer they are related with is not necessarily the same. For example germinal mutations in KRas predispose with a relatively low risk to a limited number of juvenile tumors (Hernández-Martín & Torrelo, 2011) while somatic mutation in KRas is a driver gene in many different types of adult cancers including a great majority of pancreatic cancers (Jones *et al.*, 2008). Differentiating somatic from germinal mutation is thus important.

Phenotypic traits found in Mendelian disorders often present a non-Mendelian transmission (Dipple & McCabe, 2000) and can be encountered outside syndromes as complex traits, including complex diseases such as cancer, diabetes or heart defect. For example heart septal defects are encountered in many syndromes while more than 90% of congenital heart diseases are multifactorial (Arnold *et al.*, 2006) The genes and interactions found with this kind of approaches could therefore be interesting also for complex disease understanding.
Besides, most approaches that prioritize disease genes use global disease similarity (Oellrich *et al.*, 2012). Using separately phenotypes found in Mendelian diseases would extend this approach, in particular for complex diseases. A recent study has indeed demonstrated that loci found through GWAS, especially replicated ones, were enriched with loci of Mendelian diseases that predispose to the corresponding complex traits (Blair *et al.*, 2013). Therefore Mendelian disease loci should be valid targets to predict complex diseases genes.

Phenotypes can themselves be mapped to other disease terminologies such as what Orphanet is doing now. Indeed it has developed a thesaurus of clinical traits, mapped to HPO and SNOMED-CT, enhancing the interoperability with clinical data (www.orphadata.org). Such mapping could for example help prioritizing variants in patients whose medical record has been indexed with SNOMED-CT for phenotypes and diseases found in Mendelian disorders, through Orphanet clinical traits or HPO.

More generally, the use of phenotypic data in translational research is useful in complementation to diagnoses that depend on the interpretation of clinical traits given past or current knowledge and treatments. Attempts are now being done in this direction such as the eMERGE effort to map clinical phenotypes from electronic medical records to SNOMED-CT (Pathak *et al.*, 2011). Consistent representation of phenotypes is thus needed. The

International Consortium for Human Phenotype Terminologies, created in 2012, aims for example at defining standard phenotypic terms for rare diseases to ensure interoperability between different phenotypic resources such as HPO and Orphanet. Interoperability with other species is also necessary (Schofield *et al.*, 2011; Collier *et al.*, 2013) and experience in biology with representation of phenotype can inspire their formal representation in human (Oellrich *et al.*, 2013).

5. Conclusion

There is a need for translational solutions to bring information from fundamental research toward clinical solutions and to bring clinical observation to fundamental research to better understand physiology and pathology, creating a virtuous circle.

Mendelian diseases offer a direct link from genotype to phenotype. By using disease semantic standards, better integration of molecular and clinical data are possible but automatic procedures are needed. The development of such tool was presented here and enabled the mapping of UniProtKB/Swiss-Prot, a central molecular data resource, with disease concepts from MeSH. Moreover a web interface was made available to query variants and proteins according to their implication in disease combined with features of the variant for exploring the link between a change at the molecular level and its consequences.

To investigate further the use of controlled vocabularies in genomic translational research, a literature survey of pleiotropy in Mendelian diseases enabled to relate it to protein modularity. This makes distinct clinical traits highly valuable for isolating spatio-temporal contexts and biological processes, for example in PPIs network. A prototype tool which uses a phenotype controlled vocabulary to filter PPIs was therefore developed. This kind of approach have theoretically the potential to extract biological data of high value to understand processes behind given phenotypes, often corresponding to complex traits or diseases, improving knowledge about them.

Translational genomic efforts need semantic standards and should not disregard Mendelian disorders and their clinical features.

6. References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., … McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. doi:10.1038/nature11632

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., & Pickard, B. S. (2006). SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, *22*(6), 773–774. doi:10.1093/bioinformatics/btk031

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., … Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, *24*(5), 537–544. doi:10.1038/nbt1203

Ala, U., Piro, R. M., Grassi, E., Damasco, C., Silengo, L., Oti, M., … Di Cunto, F. (2008). Prediction of Human Disease Genes by Human-Mouse Conserved Coexpression Analysis. *PLoS Comput Biol*, *4*(3), e1000043. doi:10.1371/journal.pcbi.1000043

Antezana, E., Egaña, M., Baets, B. D., Kuiper, M., & Mironov, V. (2008). ONTO-PERL: An API for supporting the development and analysis of bio-ontologies. *Bioinformatics*, *24*(6), 885–887. doi:10.1093/bioinformatics/btn042

Antonarakis, S. E., & Beckmann, J. S. (2006). Mendelian disorders deserve more attention. *Nature Reviews Genetics*, *7*(4), 277–282. doi:10.1038/nrg1826

Antonarakis, S. E., & Cooper, D. N. (2001). Mutations in Human Genetic Disease. In *eLS*. John Wiley & Sons, Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1038/npg.els.0005471/abstract

Arnold, C., Christopher, P. H., & Bernadette, M. (2006). March of Dimes: Global Report on Birth Defects, the Hidden Toll of Dying and Disabled Children. *White Plains, New York*.

Aronson, A. R., & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229–236. doi:10.1136/jamia.2009.002733

Badano, J. L., Kim, J. C., Hoskins, B. E., Lewis, R. A., Ansley, S. J., Cutler, D. J., … Katsanis, N. (2003). Heterozygous mutations in BBS1, BBS2 and BBS6 have a potential epistatic effect on Bardet–Biedl patients with two mutations at a second BBS locus. *Human Molecular Genetics*, *12*(14), 1651–1659.

Badano, J. L., Leitch, C. C., Ansley, S. J., May-Simera, H., Lawson, S., Lewis, R. A., … Katsanis, N. (2006). Dissection of epistasis in oligogenic Bardet-Biedl syndrome. *Nature*, *439*(7074), 326–330. doi:10.1038/nature04370

Badano, J. L., Mitsuma, N., Beales, P. L., & Katsanis, N. (2006). The ciliopathies: an emerging class of human genetic disorders. *Annual Review of Genomics and Human Genetics*, *7*, 125–148. doi:10.1146/annurev.genom.7.080505.115610

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113. doi:10.1038/nrg1272

Batzer, M. a, & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews. Genetics*, *3*(5), 370–9. doi:10.1038/nrg798

Bergeron, E., Simons, R., Linton, C., Yang, F., Tallon, J. M., Stewart, T. C., … Stephens, M. (2007). Canadian benchmarks in trauma. *The Journal of Trauma and Acute Care Surgery*, *62*(2), 491–497.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web: Scientific American. Retrieved January 4, 2014, from http://www.scientificamerican.com/article.cfm?id=the-semantic-web

Blair, D. R., Lyttle, C. S., Mortensen, J. M., Bearden, C. F., Jensen, A. B., Khiabanian, H., … Rzhetsky, A. (2013). A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell*, *155*(1), 70–80. doi:10.1016/j.cell.2013.08.030

Blake, J. A., & Bult, C. J. (2006). Beyond the data deluge: Data integration and bio-ontologies. *Journal of Biomedical Informatics*, *39*(3), 314–320. doi:10.1016/j.jbi.2006.01.003

Blanchard, E., Harzallah, M., Briand, H., & Kuntz, P. (2005). A Typology Of Ontology-Based Semantic Measures. In *EMOI-INTEROP*. Retrieved from http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-160/paper26.pdf

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(Database issue), D267–D270. doi:10.1093/nar/gkh061

Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform*, *47*, 67–79.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). UniProtKB/Swiss-Prot. *Methods in Molecular Biology (Clifton, N.J.)*, *406*, 89–112.

Brinkman, R. R., Dubé, M.-P., Rouleau, G. A., Orr, A. C., & Samuels, M. E. (2006). Human monogenic disorders - a source of novel drug targets. *Nature Reviews. Genetics*, *7*(4), 249–60. doi:10.1038/nrg1828

Bromberg, Y. (2013). Chapter 15: Disease Gene Prioritization. *PLoS Comput Biol*, *9*(4), e1002902. doi:10.1371/journal.pcbi.1002902

Bromberg, Y., & Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, *35*(11), 3823–3835. doi:10.1093/nar/gkm238

Buske, O. J., Manickaraj, A., Mital, S., Ray, P. N., & Brudno, M. (2013). Identification of deleterious synonymous variants in human genomes. *Bioinformatics*, btt308. doi:10.1093/bioinformatics/btt308

Capriotti, E., Calabrese, R., & Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics (Oxford, England)*, *22*(22), 2729–2734. doi:10.1093/bioinformatics/btl423

Casorelli, I., Bossa, C., & Bignami, M. (2012). DNA Damage and Repair in Human Cancer: Molecular Mechanisms and Contribution to Therapy-Related Leukemias. *International Journal of Environmental Research and Public Health*, *9*(8), 2636–2657. doi:10.3390/ijerph9082636

Cheatham, M., & Hitzler, P. (2013). String similarity metrics for ontology alignment. In *The Semantic Web–ISWC 2013* (pp. 294–309). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-41338-4_19

Chen, J., Aronow, B. J., & Jegga, A. G. (2009). Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, *10*, 73. doi:10.1186/1471-2105-10-73

Cirstea, I. C., Kutsche, K., Dvorsky, R., Gremer, L., Carta, C., Horn, D., … Zenker, M. (2010). A restricted spectrum of NRAS mutations causes Noonan syndrome. *Nature Genetics*, *42*(1), 27–29. doi:10.1038/ng.497

Cohen, W. W., Ravikumar, P. D., & Fienberg, S. E. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks. In *IIWeb* (Vol. 2003, pp. 73–78). Retrieved from http://dc-pubs.dbs.uni-leipzig.de/files/Cohen2003Acomparisonofstringdistance.pdf

Collier, N., Oellrich, A., & Groza, T. (2013). Toward knowledge support for analysis and interpretation of complex traits. *Genome Biology*, *14*(9), 214. doi:10.1186/gb-2013-14-9-214

Conrad, D. F., Keebler, J. E. M., Depristo, M. A., Lindsay, S. J., Cassals, F., Idaghdour, Y., … Kiran, V. (2012). Europe PMC Funders Group Variation in genome-wide mutation rates within and between human families, *43*(7), 712–714. doi:10.1038/ng.862.Variation

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., … Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704–712. doi:10.1038/nature08516

Consortium, G. O. (2006). The Gene Ontology (GO) project in 2006. Nucleic Acids Research, 34(suppl 1), D322–D326. doi:10.1093/nar/gkj021

Consortium, T. 1000 G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. doi:10.1038/nature11632

Cordero, P., & Ashley, E. A. (2012). Whole-genome sequencing in personalized therapeutics. *Clinical Pharmacology and Therapeutics*, *91*(6), 1001–1009. doi:10.1038/clpt.2012.51

Craig, J., & others. (2008). Complex diseases: Research and applications. *Nature Education*, *1*(1), 184.

David, F. P. A., & Yip, Y. L. (2008). SSMap: a new UniProt-PDB mapping resource for the curation of structural-related information in the UniProt/Swiss-Prot Knowledgebase. *BMC Bioinformatics*, *9*, 391. doi:10.1186/1471-2105-9-391

Den Dunnen, J. T., Antonarakis, S. E., & others. (2000). Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Human Mutation*, *15*(1), 7–12.

Dipple, K. M., & McCabe, E. R. B. (2000). Phenotypes of Patients with "Simple" Mendelian Disorders Are Complex Traits: Thresholds, Modifiers, and Systems Dynamics. *The American Journal of Human Genetics*, *66*(6), 1729–1735. doi:10.1086/302938

Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, *426*(6968), 884–90. doi:10.1038/nature02261

Doll, R. (1995). Chronic and degenerative disease: major causes of morbidity and death. *The American Journal of Clinical Nutrition*, *62*(6 Suppl), 1301S–1305S.

Duek, P. D., Gleizes, A., Zwahlen, C., Mottaz, A., Bairoch, A., & Lane, L. (2011). CALOHA: A new human anatomical ontology as a support for complex queries and tissue expression display in neXtProt. *Bio-Ontologies 2011*. Retrieved from http://bio-ontologies.knowledgeblog.org/196

Erlandsen, H., Patch, M. G., Gamez, A., Straub, M., & Stevens, R. C. (2003). Structural Studies on Phenylalanine Hydroxylase and Implications Toward Understanding and Treating Phenylketonuria. *Pediatrics*, *112*(Supplement 4), 1557–1565.

Erten, S., Bebek, G., & Koyutürk, M. (2011). Disease gene prioritization based on topological similarity in protein-protein interaction networks. In *Research in Computational Molecular Biology* (pp. 54–68). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-20036-6_7

Feldman, I., Rzhetsky, A., & Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences*, *105*(11), 4323–4328. doi:10.1073/pnas.0701722105

Fortin, A. S., Underhill, D. A., & Gros, P. (1997). Reciprocal Effect of Waardenburg Syndrome Mutations on DNA Binding by the Pax-3 Paired Domain and Homeodomain. *Human Molecular Genetics*, *6*(11), 1781–1790. doi:10.1093/hmg/6.11.1781

Franke, L., van Bakel, H., Fokkens, L., de Jong, E. D., Egmont-Petersen, M., & Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American Journal of Human Genetics*, *78*(6), 1011–1025. doi:10.1086/504300

Fredly, H., Gjertsen, B. T., & Bruserud, Ø. (2013). Histone deacetylase inhibition in the treatment of acute myeloid leukemia: the effects of valproic acid on leukemic cells, and the clinical and experimental evidence for combining valproic acid with other antileukemic agents. *Clinical Epigenetics*, *5*(1), 12. doi:10.1186/1868-7083-5-12

Friedberg, E. C. (2003). DNA damage and repair. Nature, 421(6921), 436–440. doi:10.1038/nature01408

Friedenson, B. (2007). The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. *BMC Cancer*, *7*, 152. doi:10.1186/1471-2407-7-152

Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., … others. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, *38*(3), 285–293.

Genin, E., Feingold, J., & Clerget-Darpoux, F. (2008). Identifying modifier genes of monogenic disease: strategies and difficulties. *Human Genetics*, *124*(4), 357–368. doi:10.1007/s00439-008-0560-2

George, R. A., Liu, J. Y., Feng, L. L., Bryson-Richardson, R. J., Fatkin, D., & Wouters, M. A. (2006). Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Research*, *34*(19), e130. doi:10.1093/nar/gkl707

Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. A. (2012). Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*, *20*(5), 490–497. doi:10.1038/ejhg.2011.258

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, *104*(21), 8685–8690. doi:10.1073/pnas.0701361104

Gonzalez, M. W., & Kann, M. G. (2012). Chapter 4: Protein Interactions and Disease. *PLoS Comput Biol*, *8*(12), e1002819. doi:10.1371/journal.pcbi.1002819

Greenes, R. A. (2011). *Clinical decision support: the road ahead*. Academic Press. Retrieved from http://books.google.ch/books?hl=en&lr=&id=_f4u1I-c6PIC&oi=fnd&pg=PP1&dq=Clinical+decision+support+:+the+road+ahead.+Amsterdam+%3B+Boston:+Elsevier+Academic+Press%3B+2007&ots=lEWY2CJ3Hf&sig=-z-WcfnNyNykfy6IathveiC2qzY

Groth, P., Weiss, B., Pohlenz, H.-D., & Leser, U. (2008). Mining phenotypes for gene function prediction. *BMC Bioinformatics*, *9*(1), 136. doi:10.1186/1471-2105-9-136

Guex, N., & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, *18*(15), 2714–2723. doi:10.1002/elps.1150181505

He, X., & Zhang, J. (2006). Toward a molecular understanding of pleiotropy. *Genetics*, *173*(4), 1885–1891. doi:10.1534/genetics.106.060269

Hermjakob, H., Fleischmann, W., & Apweiler, R. (1999). Swissknife - “lazy parsing” of SWISS-PROT entries. *Bioinformatics*, *15*(9), 771–772. doi:10.1093/bioinformatics/15.9.771

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, *33*(suppl 1), D514–D517. doi:10.1093/nar/gki033

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, *402*(6761 Suppl), C47–52. doi:10.1038/35011540

Ha-Thuc, V., & Srinivasan, P. (2007). Exploiting synonym relationships in biomedical named entity matching. *BioLINK SIG 2007, ISMB/ECCB Vienna*. Retrieved from https://www.cs.uiowa.edu/~psriniva/Papers/Biolink07.doc

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919.

Hernández-Martín, A., & Torrelo, A. (2011). Rasopathies: Developmental Disorders That Predispose to Cancer and Skin Manifestations. *Actas Dermo-Sifiliográficas (English Edition)*, *102*(6), 402–416. doi:10.1016/j.adengl.2011.02.002

Hodgkin, J. (1998). Seven types of pleiotropy. *International Journal of Developmental Biology*, *42*, 501–505.

Hung, M.-C., & Link, W. (2011). Protein localization in disease and therapy. *Journal of Cell Science*, *124*(Pt 20), 3381–92. doi:10.1242/jcs.089110

Ideker, T., & Sharan, R. (2008). Protein networks in disease. *Genome Research*, *18*(4), 644–652. doi:10.1101/gr.071852.107

Ingram, V. M. (1956). A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature*, *178*(4537), 792–794.

Jaccard, P. (1901). *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., … Mering, C. von. (2009). STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, *37*(suppl 1), D412–D416. doi:10.1093/nar/gkn760

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., Angenendt, P., … Kinzler, K. W. (2008). Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science*, *321*(5897), 1801–1806. doi:10.1126/science.1164368

Jonsson, P. F., & Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, *22*(18), 2291–2297. doi:10.1093/bioinformatics/btl390

Kahraman, A., Avramov, A., Nashev, L. G., Popov, D., Ternes, R., Pohlenz, H.-D., & Weiss, B. (2005). PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. *Bioinformatics*, *21*(3), 418–420. doi:10.1093/bioinformatics/bti010

Kahraman, A., Morris, R. J., Laskowski, R. A., & Thornton, J. M. (2007). Variation of geometrical and physicochemical properties in protein binding pockets and their ligands. *BMC Bioinformatics*, *8*(Suppl 8), S1. doi:10.1186/1471-2105-8-S8-S1

Karagiannis, T. C., & El-Osta, A. (2006). Chromatin modifications and DNA double-strand breaks: the current state of play. *Leukemia*, *21*(2), 195–200. doi:10.1038/sj.leu.2404478

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

Katsanis, N. (2004). The oligogenic properties of Bardet–Biedl syndrome. *Human Molecular Genetics*, *13*(suppl 1), R65–R71. doi:10.1093/hmg/ddh092

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell*, *155*(1), 27–38. doi:10.1016/j.cell.2013.09.006

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., … Robinson, P. N. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, *42*(D1), D966–D974. doi:10.1093/nar/gkt1026

Lage, K., Karlberg, E. O., Størling, Z. M., Ólason, P. Í., Pedersen, A. G., Rigina, O., … Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, *25*(3), 309–316. doi:10.1038/nbt1295

Lan, A., Ziv-Ukelson, M., & Yeger-Lotem, E. (2013). A context-sensitive framework for the analysis of human signalling pathways in molecular interaction networks. *Bioinformatics*, *29*(13), i210–i216. doi:10.1093/bioinformatics/btt240

Li, S., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2010). Loss of post-translational modification sites in disease. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 337–47.

Lin, H., Du, J., Jiang, H., & Begley, T. P. (2007). Post-Translational Modifications to Regulate Protein Function. In *Wiley Encyclopedia of Chemical Biology*. John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470048672.wecb467/abstract

Linde, L., Boelz, S., Nissim-Rafinia, M., Oren, Y. S., Wilschanski, M., Yaacov, Y., … Kerem, B. (2007). Nonsense-mediated mRNA decay affects nonsense transcript levels and governs response of cystic fibrosis patients to gentamicin. *Journal of Clinical Investigation*, *117*(3), 683–692. doi:10.1172/JCI28523

Liu, J., Kim, J., & Oberdoerffer, P. (2013). Metabolic modulation of chromatin: implications for DNA repair and genomic integrity. *Frontiers in Genetics*, *4*. doi:10.3389/fgene.2013.00182

Lobo, I. (2008). Multifactorial inheritance and genetic disease. *Nat Educ*, *1*(1).

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000). Mutations: Types and Causes. Text. Retrieved January 3, 2014, from http://www.ncbi.nlm.nih.gov/books/NBK21578/

López-Bigas, N., & Ouzounis, C. A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Research*, *32*(10), 3108–3114. doi:10.1093/nar/gkh605

Lu, L. J., Xia, Y., Yu, H., Rives, A., Lu, H., Schubert, F., & Gerstein, M. (2005). Protein interaction prediction by integrating genomic features and protein interaction network analysis. *Data Analysis and Visualization in Genomics and Proteomics*, 61.

Machado, C. M., Rebholz-Schuhmann, D., Freitas, A. T., & Couto, F. M. (2013). The semantic web in translational medicine: current applications and future directions. *Briefings in Bioinformatics*, bbt079. doi:10.1093/bib/bbt079

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, *35*(Database issue), D26–31. doi:10.1093/nar/gkl993

Monge, A. E., & Elkan, C. (1996). The Field Matching Problem: Algorithms and Applications. In *KDD* (pp. 267–270). Retrieved from http://www.aaai.org/Papers/KDD/1996/KDD96-044.pdf

Moreau, Y., & Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*. Retrieved from http://www.nature.com/nrg/journal/vaop/ncurrent/full/nrg3253.html

Mottaz, A., David, F. P. A., Veuthey, A.-L., & Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics*, *26*(6), 851–852.

Mottaz, A., Yip, Y. L., Ruch, P., & Veuthey, A.-L. (2008). Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics*, *9*(Suppl 5), S3–S3.

Mullighan, C. G., Zhang, J., Kasper, L. H., Lerach, S., Payne-Turner, D., Phillips, L. A., … Downing, J. R. (2011). CREBBP mutations in relapsed acute lymphoblastic leukaemia. *Nature*, *471*(7337), 235–239. doi:10.1038/nature09727

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., & Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, *13*(1), R5. doi:10.1186/gb-2012-13-1-r5

Natarajan, M., Lin, K.-M., Hsueh, R. C., Sternweis, P. C., & Ranganathan, R. (2006). A global analysis of cross-talk in a mammalian cellular signalling network. *Nature Cell Biology*, *8*(6), 571–580. doi:10.1038/ncb1418

Oellrich, A., Grabmüller, C., & Rebholz-Schuhmann, D. (2013). Automatically transforming pre- to post-composed phenotypes: EQ-lising HPO and MP. *Journal of Biomedical Semantics*, *4*(1), 29. doi:10.1186/2041-1480-4-29

Oellrich, A., Hoehndorf, R., Gkoutos, G. V., & Rebholz-Schuhmann, D. (2012). Improving Disease Gene Prioritization by Comparing the Semantic Similarity of Phenotypes in Mice with Those of Human Diseases. *PLoS ONE*, *7*(6), e38937. doi:10.1371/journal.pone.0038937

Ogiwara, H., & Kohno, T. (2012). CBP and p300 histone acetyltransferases contribute to homologous recombination by transcriptionally activating the BRCA1 and RAD51 genes. *PloS One*, *7*(12), e52810. doi:10.1371/journal.pone.0052810

Ogiwara, H., Ui, A., Otsuka, A., Satoh, H., Yokomi, I., Nakajima, S., … Kohno, T. (2011). Histone acetylation by CBP and p300 at double-strand break sites facilitates SWI/SNF chromatin remodeling and the recruitment of non-homologous end joining factors. *Oncogene*, *30*(18), 2135–2146. doi:10.1038/onc.2010.592

Olry, A., Urbero, B., Choquet, R., & Charlet, J. (2011). OntoOrpha: An Ontology to Support the Editing and Audit of Knowledge of Rare Diseases in ORPHANET. Retrieved from http://ceur-ws.org/Vol-833/paper35.pdf

Omran, A. R. (1971). The epidemiologic transition: a theory of the epidemiology of population change. *The Milbank Memorial Fund Quarterly*, 509–538.

On beyond GWAS. (2010). *Nature Genetics*, *42*(7), 551–551. doi:10.1038/ng0710-551

Ortutay, C., & Vihinen, M. (2009). Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Research*, *37*(2), 622–628. doi:10.1093/nar/gkn982

Osborne, J. D., Lin, S., Kibbe, W., Zhu, L., Danila, M., & Rex, C. (2007). GeneRIF is a more comprehensive, current and computationally tractable source of gene-disease relationships than OMIM. *Bioinformatics Core, Northwestern University Technical Report*. Retrieved from http://www.basic.northwestern.edu/publications/generifdo/geneRIFDO16.pdf

Oti, M., Snel, B., Huynen, M. A., & Brunner, H. G. (2006). Predicting disease genes using protein–protein interactions. *Journal of Medical Genetics*, *43*(8), 691–698. doi:10.1136/jmg.2006.041376

Oti, M., Huynen, M. A., & Brunner, H. G. (2009). The biological coherence of human phenome databases. *American Journal of Human Genetics*, *85*(6), 801–808. doi:10.1016/j.ajhg.2009.10.026

Paaby, A. B., & Rockman, M. V. (2012). The many faces of pleiotropy. *Trends in Genetics*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0168952512001692

Panoutsopoulou, K., Tachmazidou, I., & Zeggini, E. (2013). In search of low-frequency and rare variants affecting complex traits. *Human Molecular Genetics*, *22*(R1), R16–R21. doi:10.1093/hmg/ddt376

Pao, G. M., Janknecht, R., Ruffner, H., Hunter, T., & Verma, I. M. (2000). CBP/p300 interact with and function as transcriptional coactivators of BRCA1. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(3), 1020–1025.

Pathak, J., Wang, J., Kashyap, S., Basford, M., Li, R., Masys, D. R., & Chute, C. G. (2011). Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *Journal of the American Medical Informatics Association*, *18*(4), 376–386. doi:10.1136/amiajnl-2010-000061

Pawson, T., & Nash, P. (2000). Protein–protein interactions define specificity in signal transduction. *Genes & Development*, *14*(9), 1027–1047. doi:10.1101/gad.14.9.1027

Perez-Iratxeta, C., Bork, P., & Andrade-Navarro, M. A. (2007). Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Research*, *35*(suppl 2), W212–W216. doi:10.1093/nar/gkm223

Pericak-Vance, M. A., Bebout, J. L., Gaskell, P. C., Yamaoka, L. H., Hung, W.-Y., Alberts, M. J., … Roses, A. D. (1991). Linkage studies in familial Alzheimer disease: Evidence for chromosome 19 linkage. *American Journal of Human Genetics*, *48*(6), 1034–1050.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. *PLoS Comput Biol*, *5*(7), e1000443. doi:10.1371/journal.pcbi.1000443

Pillet, V., Zehnder, M., Seewald, A. K., Veuthey, A.-L., & Petrak, J. (2005). GPSDB: a new database for synonyms expansion of gene and protein names. *Bioinformatics*, *21*(8), 1743–1744. doi:10.1093/bioinformatics/bti235

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. *Nature*, *505*(7483), 302–308. doi:10.1038/nature12981

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. doi:10.1038/nature05329

Robinson, P. N., & Mundlos, S. (2010). The human phenotype ontology. *Clinical Genetics*, *77*(6), 525–534. doi:10.1111/j.1399-0004.2010.01436.x

Sabherwal, N., Schneider, K. U., Blaschke, R. J., Marchini, A., & Rappold, G. (2004). Impairment of SHOX nuclear localization as a cause for Léri-Weill syndrome. *Journal of Cell Science*, *117*(14), 3041–3048. doi:10.1242/jcs.01152

Sanner, M. F., Olson, A. J., & Spehner, J. C. (1996). Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, *38*(3), 305–320. doi:10.1002/(SICI)1097-0282(199603)38:3<305::AID-BIP4>3.0.CO;2-Y

Schaefer, M. H., Lopes, T. J. S., Mah, N., Shoemaker, J. E., Matsuoka, Y., Fontaine, J.-F., … Andrade-Navarro, M. A. (2013). Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Computational Biology*, *9*(1), e1002860. doi:10.1371/journal.pcbi.1002860

Schlicker, A., Lengauer, T., & Albrecht, M. (2010). Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics*, *26*(18), i561–i567. doi:10.1093/bioinformatics/btq384

Schneider, A., Dessimoz, C., & Gonnet, G. H. (2007). OMA Browser—Exploring orthologous relations across 352 complete genomes. *Bioinformatics*, *23*(16), 2180–2182. doi:10.1093/bioinformatics/btm295

Schofield, P. N., Sundberg, J. P., Hoehndorf, R., & Gkoutos, G. V. (2011). New approaches to the representation and analysis of phenotype knowledge in human diseases and their animal models. *Briefings in Functional Genomics*, *10*(5), 258–265. doi:10.1093/bfgp/elr031

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., … Kibbe, W. A. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, *40*(D1), D940–D946. doi:10.1093/nar/gkr972

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, *40*(W1), W452–W457. doi:10.1093/nar/gks539

Sioutos, N., Coronado, S. de, Haber, M. W., Hartel, F. W., Shaiu, W.-L., & Wright, L. W. (2007). NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, *40*(1), 30–43. doi:10.1016/j.jbi.2006.02.013

Smith, C. L., & Eppig, J. T. (2009). The Mammalian Phenotype Ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, *1*(3), 390–399. doi:10.1002/wsbm.44

Spirin, V., & Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, *100*(21), 12123–12128. doi:10.1073/pnas.2032324100

Stearns, F. W. (2010). One hundred years of pleiotropy: a retrospective. *Genetics*, *186*(3), 767–73. doi:10.1534/genetics.110.122549

Su, Z., Zeng, Y., & Gu, X. (2010). A preliminary analysis of gene pleiotropy estimated from protein sequences. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *314B*(2), 115–122. doi:10.1002/jez.b.21315

Strachan, T., & Read, A. P. (2011). *Human Molecular Genetics 4*. Garland Science/Taylor & Francis Group.

Terry, S. F., & Terry, P. F. (2011). Power to the People: Participant Ownership of Clinical Trial Data. *Science Translational Medicine*, *3*(69), 69cm3–69cm3. doi:10.1126/scitranslmed.3001857

Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B., & Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*, *33*(5), 1544–1552. doi:10.1093/nar/gki296

Turner, F. S., Clutterbuck, D. R., & Semple, C. A. M. (2003). POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biology*, *4*(11), R75. doi:10.1186/gb-2003-4-11-r75

UniProt Consortium. (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Research*, *37*(Database issue), D169–174. doi:10.1093/nar/gkn664

Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins: Structure, Function, and Bioinformatics*, *48*(2), 227–241. doi:10.1002/prot.10146

Van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., & Leunissen, J. A. M. (2006). A text-mining analysis of the human phenome. *European Journal of Human Genetics*, *14*(5), 535–542. doi:10.1038/sj.ejhg.5201585

Vempati, R. K., Jayani, R. S., Notani, D., Sengupta, A., Galande, S., & Haldar, D. (2010). p300-mediated acetylation of histone H3 lysine 56 functions in DNA damage response in mammals. *The Journal of Biological Chemistry*, *285*(37), 28553–28564. doi:10.1074/jbc.M110.149393

Von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., … Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, *33*(Database issue), D433–D437. doi:10.1093/nar/gki005

Wachi, S., Yoneda, K., & Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, *21*(23), 4205–4208. doi:10.1093/bioinformatics/bti688

Wagner, G. P., & Zhang, J. (2011). The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics*, *12*(3), 204–213. doi:10.1038/nrg2949

Wang, J., Zhou, X., Zhu, J., Zhou, C., & Guo, Z. (2010). Revealing and avoiding bias in semantic similarity scores for protein pairs. *BMC Bioinformatics*, *11*(1), 290. doi:10.1186/1471-2105-11-290

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, *17*(4), 263–70. doi:10.1002/humu.22

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., & Lewis, S. E. (2009). Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation. *PLoS Biol*, *7*(11), e1000247. doi:10.1371/journal.pbio.1000247

Waters, P. J., Parniak, M. A., Akerman, B. R., & Scriver, C. R. (2000). Characterization of Phenylketonuria Missense Substitutions, Distant from the Phenylalanine Hydroxylase Active Site, Illustrates a Paradigm for Mechanism and Potential Modulation of Phenotype. *Molecular Genetics and Metabolism*, *69*(2), 101–110. doi:10.1006/mgme.2000.2965

Wilson, K., Rocha, A. M., Padmanabhan, K., Wang, K., Chen, Z., Jin, Y., … Samatova, N. F. (2011). Detecting Pathway Cross-Talks by Analyzing Conserved Functional Modules across Multiple Phenotype-Expressing Organisms. In *2012 IEEE International Conference on Bioinformatics and Biomedicine* (Vol. 0, pp. 443–449). Los Alamitos, CA, USA: IEEE Computer Society. doi:10.1109/BIBM.2011.35

Winkler, W. E. (1999). The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.39.4336

Xu, J., & Li, Y. (2006). Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, *22*(22), 2800–2805. doi:10.1093/bioinformatics/btl467

Yip, Y. L., Scheib, H., Diemand, A. V., Gattiker, A., Famiglietti, L. M., Gasteiger, E., & Bairoch, A. (2004). The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Human Mutation*, *23*(5), 464–470. doi:10.1002/humu.20021

Yue, P., Melamud, E., & Moult, J. (2006). SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics*, *7*(1), 166. doi:10.1186/1471-2105-7-166

Zhang, Z., Miteva, M. A., Wang, L., & Alexov, E. (2012). Analyzing Effects of Naturally Occurring Missense Mutations. *Computational and Mathematical Methods in Medicine*, *2012*. doi:10.1155/2012/805827

7. Supplementary material

Figure S1

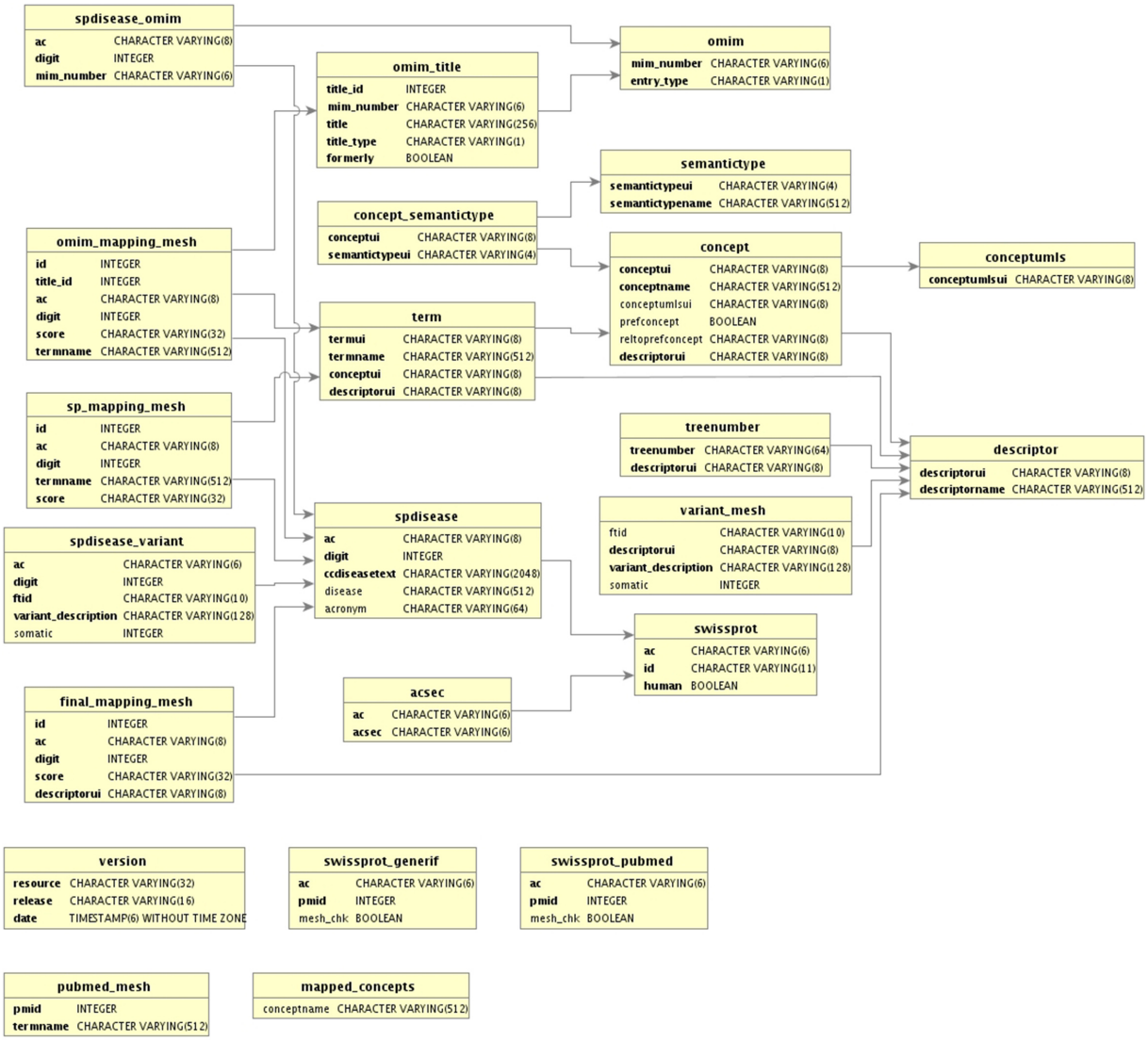


Figure S1: Database schema used for the mapping of UniProtKB/Swiss-Prot to MeSH.

Figure S2

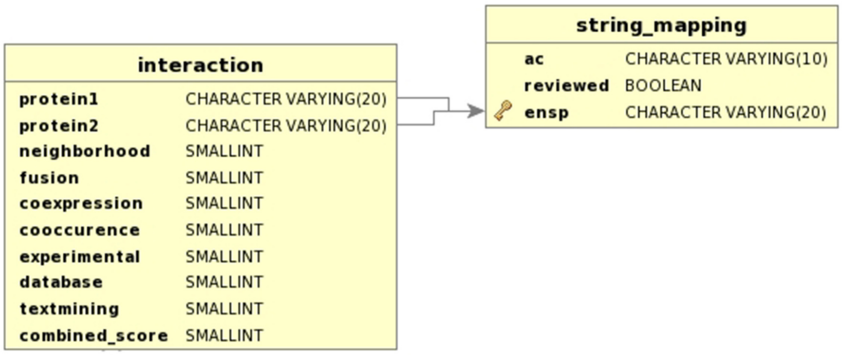


Figure S2: Database schema used for the PPI contextualization tool.

Additional figure 1, Mottaz *et al.*, 2008

Mapping of the Swiss-Prot disease comment lines to MeSH descriptors: result on the benchmark									
Origin	**True / False**	**Score**	**Disease**	**Mapped term**	**Mapped DescriptorUI**	**Correct term**	**Correct DescriptorUI**	**AC**	**CC line**
SP	T	3.1126	Idiopathic generalized epilepsy	Epilepsy, generalized	D004829	Generalized epilepsy	D004829	O00305	1
OMIM	T	3.1126	Epilepsy, idiopathic generalized	Epilepsy, generalized	D004829	Generalized epilepsy	D004829	O00305	1
SP	T	exact	Juvenile myoclonic epilepsy	Juvenile myoclonic epilepsy	D020190	Juvenile myoclonic epilepsy	D020190	O00305	2
OMIM	T	exact	Epilepsy, juvenile myoclonic	Epilepsy, juvenile myoclonic	D020190	Juvenile myoclonic epilepsy	D020190	O00305	2
SP	T	5.2698	Torsion dystonia 1	Dystonias, torsion	D004422	Idiopathic torsion dystonia	D004422	O14656	1
OMIM	T	8.5451	Dystonia musculorum deformans 1	Dystonia musculorum feformans	D004422	Idiopathic torsion dystonia	D004422	O14656	1
SP	T	1.2915	Squamous cell carcinoma of the head and neck	Carcinoma, cquamous cell	D002294	Squamous cell carcinoma | Head and neck cancer	D002294 D006258	O14763	1
OMIM	T	1.2915	Squamous cell carcinoma, head and neck	Carcinoma, squamous cell	D002294	Squamous cell carcinoma | Head and neck cancer	D002294 D006258	O14763	1
SP	T	-1.2171	Progressive familial intrahepatic cholestasis type 1	Cholestases, intrahepatic	D002780	Genetic disease, inborn | Intrahepatic cholestasis	D030342 D002780	O43520	1

OMIM	T	4.4199	Cholestasis, fatal intrahepatic	Cholestases, intrahepatic	D002780	Genetic disease, inborn | Intrahepatic cholestasis	D030342 D002780	O43520	1
SP	T	1.1308	Benign recurrent intrahepatic cholestasis	Cholestases, intrahepatic	D002780	Genetic disease, inborn | Intrahepatic cholestasis	D030342 D002780	O43520	2
OMIM	T	-0.7496	Cholestasis, benign recurrent intrahepatic 1	Cholestases, intrahepatic	D002780	Genetic disease, inborn | Intrahepatic cholestasis	D030342 D002780	O43520	2
SP	T	0.8412	Recurrent intrahepatic cholestasis of pregnancy	Cholestases, intrahepatic	D002780	Intrahepatic cholestasis | Pregnancy complications	D002780 D011248	O43520	3
OMIM	T	4.7828	Cholestasis, intrahepatic, of pregnancy	Cholestases, intrahepatic	D002780	Intrahepatic cholestasis | Pregnancy complications	D002780 D011248	O43520	3
SP	T	0.6919	Autosomal recessive limb girdle muscular dystrophy type 2b	Muscular dystrophies, limb girdle	D049288	Limb-girdle muscular dystrophy	D049288	O75923	1
OMIM	T	4.205	Muscular dystrophy, limb-girdle, type 3	Muscular dystrophies, limb girdle	D049288	Limb-girdle muscular dystrophy	D049288	O75923	1
SP	T	exact	Bladder cancer	Bladder cancer	D001749	Bladder cancer	D001749	P01112	3
OMIM	T	exact	Bladder cancer	Bladder cancer	D001749	Bladder cancer	D001749	P01112	3
SP	T	1.9776	Spondyloepiphyseal dysplasia congenital type	Spondyloepiphyseal dysplasias	D010009	Genetic disease, inborn | abnormalities, multiple | spondyloepiphyseal dysplasia | dwarfism	D030342 D000015 D010009 D004392	P02458	2
OMIM	T	3.7594	Spondyloepiphyseal dysplasia congenita	Spondyloepiphyseal dysplasias	D010009	Genetic disease, inborn | Abnormalities, multiple | Spondyloepiphyseal dysplasia | Dwarfism	D030342 D000015 D010009 D004392	P02458	2
SP	T	0.1118	Primary avascular necrosis of femoral head	Necrosis, avascular, of femur head	D005271	Avascular necrosis of femur head	D005271	P02458	6
OMIM	T	2.4825	Femoral head, avascular necrosis of	Necrosis, avascular, of femur head	D005271	Avascular necrosis of femur head	D005271	P02458	6

SP	T	-0.8968	Multiple epiphyseal dysplasia with myopia and conductive deafness	Dysplasias, multiple epiphyseal	D010009	Genetic disease, inborn | Abnormalities, multiple | Multiple epiphyseal dysplasia | Eye diseases | Hearing loss, conductive | Dwarfism	D030342 D000015 D010009 D005128 D006314 D004392	P02458	9
OMIM	T	-0.8968	Epiphyseal dysplasia, multiple, with myopia and conductive deafness	Dysplasias, multiple epiphyseal	D010009	Genetic disease, inborn | Abnormalities, multiple | Multiple epiphyseal dysplasia | Eye diseases | Hearing loss, conductive | Dwarfism	D030342 D000015 D010009 D005128 D006314 D004392	P02458	9
SP	T	-2.0082	Autosomal dominant rhegmatogenous retinal detachment	Detachments, retinal	D012163	Eye disease, hereditary | Retinal detachment	D015785 D012163	P02458	13
OMIM	T	-2.0082	Rhegmatogenous retinal detachment, autosomal dominant	Detachments, retinal	D012163	Eye disease, hereditary | Retinal detachment	D015785 D012163	P02458	13
SP	T	-1.4586	Low hdl levels observed in high density lipoprotein deficiency type 1	High density lipoprotein deficiency, type I	D013631	Tangier disease	D013631	P02647	2
OMIM	T	exact	Tangier disease	Tangier disease	D013631	Tangier disease	D013631	P02647	2
SP	T	exact	Congenital insensitivity to pain with anhidrosis	Congenital insensitivity to pain with anhidrosis	D009477	Congenital insensitivity to pain with anhidrosis	D009477	P04629	1
OMIM	T	exact	Insensitivity to pain, congenital, with anhidrosis	Insensitivity to pain with anhidrosis, congenital	D009477	Congenital insensitivity to pain with anhidrosis	D009477	P04629	1
SP	T	3.4424	Thyroid papillary carcinoma	Carcinoma, papillary	D002291	Thyroid carcinoma | papillary carcinoma	D013964 D002291	P04629	2
OMIM	T	3.4424	Thyroid carcinoma, papillary	Carcinoma, papillary	D002291	Thyroid carcinoma | papillary carcinoma	D013964 D002291	P04629	2
SP	T	3.4424	Thyroid papillary carcinoma	Carcinoma, papillary	D002291	Thyroid carcinoma | papillary carcinoma	D013964 D002291	P04629	3

OMIM	T	3.4424	Thyroid carcinoma, papillary	Carcinoma, papillary	D002291	Thyroid carcinoma | papillary carcinoma	D013964 D002291	P04629	3
SP	T	2.1909	Hypophosphatasia infantile	Hypophosphatasias	D007014	Hypophosphatasia	D007014	P05186	1
OMIM	T	2.1909	Hypophosphatasia, infantile	Hypophosphatasias	D007014	Hypophosphatasia	D007014	P05186	1
SP	T	2.3017	Hypophosphatasia childhood	Hypophosphatasias	D007014	Hypophosphatasia	D007014	P05186	2
OMIM	T	2.3017	Hypophosphatasia, childhood	Hypophosphatasias	D007014	Hypophosphatasia	D007014	P05186	2
SP	T	-0.8011	Hypophosphatasia adult type	Hypophosphatasias	D007014	Hypophosphatasia	D007014	P05186	3
OMIM	T	1.9116	Hypophosphatasia, mild	Hypophosphatasias	D007014	Hypophosphatasia	D007014	P05186	3
SP	T	1.9252	Nemaline myopathy type 1	Myopathies, nemaline	D017696	Childhood onset nemaline myopathy | Autosomal dominant nemaline myopathy	D017696 D017696	P06753	1
OMIM	T	4.825	Nemaline myopathy 1	Myopathies, nemaline	D017696	Childhood onset nemaline myopathy | Autosomal dominant nemaline myopathy	D017696 D017696	P06753	1
SP	T	3.4424	Thyroid papillary carcinoma	Carcinoma, papillary	D002291	Thyroid carcinoma | Papillary carcinoma	D013964 D002291	P06753	2
OMIM	T	3.4424	Thyroid carcinoma, papillary	Carcinoma, papillary	D002291	Thyroid carcinoma | Papillary carcinoma	D013964 D002291	P06753	2
SP	T	-1.7911	Autosomal dominant cataract	Cataract	D002386	Genetic disease, inborn | Cataract	D030342 D002386	P07315	1
OMIM	T	-1.7911	Cataract, autosomal dominant	Cataract	D002386	Genetic disease, inborn | Cataract	D030342 D002386	P07315	1
SP	T	2.3801	Familial hypertrophic cardiomyopathy type 8	Cardiomyopathies, familial hypertrophic	D024741	Cardiomyopathy, hypertrophic, familial	D024741	P08590	1
OMIM	T	exact	Cardiomyopathy, familial hypertrophic	Cardiomyopathy, familial hypertrophic	D024741	Cardiomyopathy, hypertrophic, familial	D024741	P08590	1
SP	T	exact	Fructose-1,6-bisphosphatase deficiency	Fructose 1,6 bisphosphatase deficiency	D015319	Fructose-1,6-bisphosphatase deficiency	D015319	P09467	1
OMIM	T	exact	Fructose-1,6-bisphosphatase deficiency	Fructose-1,6-bisphosphatase deficiency	D015319	Fructose-1,6-bisphosphatase deficiency	D015319	P09467	1

SP	T	exact	Pfeiffer syndrome	Pfeiffer syndrome	D000168	Pfeiffer syndrome	D000168	P11362	1
OMIM	T	exact	Pfeiffer syndrome	Pfeiffer syndrome	D000168	Pfeiffer syndrome	D000168	P11362	1
SP	T	4.6839	Isolated hypogonadotropic hypogonadism	Hypogonadotropic hypogonadism	D007006	Hypogonadotropic hypogonadism	D007006	P11362	2
OMIM	T	exact	Hypogonadotropic hypogonadism	Hypogonadotropic hypogonadism	D007006	Hypogonadotropic hypogonadism	D007006	P11362	2
SP	T	3.2191	Kallmann syndrome type 2	Kallmann syndrome 2	D017436	Kallmann syndrome 2	D017436	P11362	3
OMIM	T	exact	Kallmann syndrome 2	Kallmann syndrome 2	D017436	Kallmann syndrome 2	D017436	P11362	3
SP	T	2.9655	Epidermolysis bullosa simplex Dowling-Meara type	Epidermolysis bullosa herpetiformis Dowling Meara	D016110	Epidermolysis bullosa herpetiformis Dowling-Meara	D016110	P13647	2
OMIM	T	10.39	Epidermolysis bullosa herpetiformis, Dowling-Meara type	Epidermolysis bullosa herpetiformis Dowling Meara	D016110	Epidermolysis bullosa herpetiformis Dowling-Meara	D016110	P13647	2
SP	T	-1.671	Epidermolysis bullosa simplex with migratory circinate erythema	Epidermolysis bullosa simplex	D016110	Epidermolysis bullosa simplex	D016110	P13647	3
OMIM	T	-1.671	Epidermolysis bullosa simplex with migratory circinate erythema	Epidermolysis bullosa simplex	D016110	Epidermolysis bullosa simplex	D016110	P13647	3
SP	T	2.3614	Epidermolysis bullosa simplex Koebner type	Epidermolysis bullosa simplex	D016110	Epidermolysis bullosa simplex kobner	D016110	P13647	5
OMIM	T	2.3614	Epidermolysis bullosa simplex, Koebner type	Epidermolysis bullosa simplex	D016110	Epidermolysis bullosa simplex kobner	D016110	P13647	5
SP	T	1.4035	Epidermolysis bullosa simplex with mottled pigmentation	Epidermolysis bullosa simplex	D016110	Epidermolysis bullosa simplex	D016110	P13647	6
OMIM	T	1.4035	Epidermolysis bullosa simplex with mottled pigmentation	Epidermolysis bullosa simplex	D016110	Epidermolysis bullosa simplex	D016110	P13647	6
SP	T	4.5728	Acute hepatic porphyria	Porphyria, hepatic	D017094	Hepatic porphyria	D017094	P13716	1
OMIM	T	4.5728	Porphyria, acute hepatic	Porphyria, hepatic	D017094	Hepatic porphyria	D017094	P13716	1
SP	T	exact	Metachromatic leukodystrophy	Metachromatic leukodystrophy	D007966	Metachromatic leukodystrophy	D007966	P15289	1

OMIM	T	exact	Metachromatic leukodystrophy	Metachromatic leukodystrophy	D007966	Metachromatic leukodystrophy	D007966	P15289	1
SP	T	exact	Multiple sulfatase deficiency	Multiple sulfatase deficiency	D052517	Multiple sulfatase deficiency disease	D052517	P15289	2
OMIM	T	exact	Multiple sulfatase deficiency	Multiple sulfatase deficiency	D052517	Multiple sulfatase deficiency disease	D052517	P15289	2
SP	T	-1.1222	Autosomal dominant nocturnal frontal lobe epilepsy type 3	Frontal lobe epilepsies	D017034	Genetic disease, inborn | Epilepsy, frontal lobe	D030342 D017034	P17787	1
OMIM	T	1.184	Epilepsy, nocturnal frontal lobe, type 3	Frontal lobe epilepsies	D017034	Genetic disease, inborn | epilepsy, frontal lobe	D030342 D017034	P17787	1
SP	T	-2.1573	Familial erythrocytosis type 1	Erythrocytoses	D011086	Genetic disease, inborn | Erythrocytosis	D030342 D011086	P19235	1
OMIM	T	-0.6183	Erythrocytosis, familial, 1	Erythrocytoses	D011086	Genetic disease, inborn | Erythrocytosis	D030342 D011086	P19235	1
SP	T	exact	Melnick-Needles syndrome	Melnick-Needles syndrome	D010009	Melnick-Needles syndrome	D010009	P21333	6
OMIM	T	exact	Melnick-Needles syndrome	Melnick-Needles syndrome	D010009	Melnick-Needles syndrome	D010009	P21333	6
SP	T	2.6771	X-linked congenital idiopathic intestinal pseudoobstruction	Idiopathic intestinal pseudo-obstructions	D007418	Genetic disease, X-linked | Digestive system abnormalities | Intestinal pseudo-obstruction	D040181 D004065 D007418	P21333	8
OMIM	T	7.3371	Congenital idiopathic intestinal pseudoobstruction	Idiopathic intestinal pseudo-obstructions	D007418	Genetic disease, x-linked | Digestive system abnormalities | Intestinal pseudo-obstruction	D040181 D004065 D007418	P21333	8
SP	T	2.3769	Autosomal recessive lamellar ichthyosis	Ichthyosis, lamellar	D017490	Lamellar ichthyosis	D017490	P22735	1
OMIM	T	exact	Lamellar ichthyosis	Lamellar ichthyosis	D017490	Lamellar ichthyosis	D017490	P22735	1
SP	T	0.678	Waardenburg syndrome type i	Waardenburg syndrome	D014849	Waardenburg's syndrome	D014849	P23760	1
OMIM	T	0.678	Waardenburg syndrome, type i	Waardenburg syndrome	D014849	Waardenburg's syndrome	D014849	P23760	1
SP	T	0.3996	Waardenburg syndrome type iii	Waardenburg syndrome	D014849	Waardenburg-Klein syndrome	D014849	P23760	2

OMIM	T	exact	Klein-Waardenburg syndrome	Klein-Waardenburg syndrome	D014849	Waardenburg-Klein syndrome	D014849	P23760	2
SP	T	6.8316	Fuchs endothelial corneal dystrophy	Fuchs endothelial dystrophy	D005642	Fuchs endothelial dystrophy	D005642	P25067	2
OMIM	T	3.8111	Corneal dystrophy, Fuchs endothelial, 1	Fuchs endothelial dystrophy	D005642	Fuchs endothelial dystrophy	D005642	P25067	2
SP	T	-0.3954	Adrenal hyperplasia type 2	Adrenal hyperplasias, congenital	D000312	Congenital adrenal hyperplasia	D000312	P26439	1
OMIM	T	1.5101	Adrenal hyperplasia II	Adrenal hyperplasias, congenital	D000312	Congenital adrenal hyperplasia	D000312	P26439	1
SP	T	6.8255	Congenital pulmonary alveolar proteinosis	Alveolar proteinosis, pulmonary	D011649	Genetic disease, inborn | Pulmonary alveolar proteinosis	D030342 D011649	P32927	1
OMIM	T	-1.8404	Pulmonary alveolar proteinosis due to surfactant protein b deficiency	Alveolar proteinosis, pulmonary	D011649	Genetic disease, inborn | Pulmonary alveolar proteinosis	D030342 D011649	P32927	1
SP	T	exact	Pilomatrixoma	Pilomatrixoma	D018296	Pilomatrixoma	D018296	P35222	2
OMIM	T	exact	Pilomatrixoma	Pilomatrixoma	D018296	Pilomatrixoma	D018296	P35222	2
SP	T	exact	Medulloblastoma	Medulloblastoma	D008527	Medulloblastoma	D008527	P35222	3
OMIM	T	exact	Medulloblastoma	Medulloblastoma	D008527	Medulloblastoma	D008527	P35222	3
SP	T	-0.2049	Paramyotonia congenita of von Eulenburg	Congenita, paramyotonia	D020967	Eulenburg disease	D020967	P35499	1
OMIM	T	-0.2049	paramyotonia congenita of von Eulenburg	Congenita, paramyotonia	D020967	Eulenburg disease	D020967	P35499	1
SP	T	exact	Hypokalemic periodic paralysis	Hypokalemic periodic paralysis	D020514	Hypokalemic periodic paralysis	D020514	P35499	2
OMIM	T	exact	Hypokalemic periodic paralysis	Hypokalemic periodic paralysis	D020514	Hypokalemic periodic paralysis	D020514	P35499	2
SP	T	exact	Hyperkalemic periodic paralysis	Hyperkalemic periodic paralysis	D020513	Hyperkalemic periodic paralysis	D020513	P35499	3
OMIM	T	exact	Hyperkalemic periodic paralysis	Hyperkalemic periodic paralysis	D020513	Hyperkalemic periodic paralysis	D020513	P35499	3
SP	T	exact	Marfan syndrome	Marfan syndrome	D008382	Marfan syndrome	D008382	P35555	1

OMIM	T	exact	Marfan syndrome	Marfan syndrome	D008382	Marfan syndrome	D008382	P35555	1
SP	T	6.1018	Isolated ectopia lentis	Ectopia lentis	D004479	Genetic disease, inborn | Ectopia lentis	D030342 D004479	P35555	2
OMIM	T	7.1052	Ectopia lentis, familial	Ectopia lentis	D004479	Genetic disease, inborn | Ectopia lentis	D030342 D004479	P35555	2
SP	T	exact	Hereditary coproporphyria	Hereditary coproporphyria	D046349	Hereditary coproporphyria	D046349	P36551	1
OMIM	T	exact	Coproporphyrinogen oxidase deficiency	Coproporphyrinogen oxidase deficiency	D046349	Hereditary coproporphyria	D046349	P36551	1
SP	T	exact	Cowden disease	Cowden disease	D006223	Cowden disease	D006223	P36894	2
OMIM	T	exact	Cowden disease	Cowden disease	D006223	Cowden disease	D006223	P36894	2
SP	T	2.7763	X-linked alpha-thalassemia/mental retardation syndrome	X-linked mental retardation syndromes	D038901	Mental retardation, X-linked | Alpha-thalassemia | Abnormalities, multiple | Urogenital Abnormalities | Craniofacial abnormalities	D038901 D017085 D000015 D014564 D019465	P46100	1
OMIM	T	-0.5131	Alpha-thalassemia/mental retardation syndrome, nondeletion type, X-linked	X-linked mental retardation syndromes	D038901	Mental retardation, X-linked | alpha-thalassemia | Abnormalities, multiple | Urogenital abnormalities | Craniofacial abnormalities	D038901 D017085 D000015 D014564 D019465	P46100	1
SP	T	-0.0746	Mental retardation X-linked with hypotonic facies syndrome type 1	X-linked mental retardation syndromes	D038901	Mental retardation, X-linked | Abnormalities, multiple | Craniofacial abnormalities	D038901 D000015 D019465	P46100	2
OMIM	T	0.8827	Mental retardation, X-linked, with growth retardation, deafness, and microgenitalism	Mental retardation, X linked	D038901	Mental retardation, X-linked | Abnormalities, multiple | Craniofacial abnormalities	D038901 D000015 D019465	P46100	2
SP	T	0.5159	alpha-thalassemia myelodysplasia syndrome	Alpha-thalassemia	D017085	Alpha-thalassemia | Hematologic diseases | Genetic diseases, X-Linked	D017085 D006402 D040181	P46100	3

OMIM	T	2.1204	hemoglobin h disease, acquired	Hemoglobin h diseases	D017085	Alpha-thalassemia | Hematologic diseases | Genetic diseases, X-linked	D017085 D006402 D040181	P46100	3
SP	T	1.8158	familial early-onset alzheimer disease type 3	Early onset alzheimer disease	D000544	Early onset alzheimer disease | Genetic disease, inborn	D000544 D030342	P49768	1
OMIM	T	5.6049	Alzheimer disease 3, early-onset	Early onset alzheimer disease	D000544	Early onset alzheimer disease | Genetic disease, inborn	D000544 D030342	P49768	1
SP	T	6.5376	Familial hemiplegic migraine 2	Familial hemiplegic migraines	D020325	Hemiplegic migraine, familial	D020325	P50993	1
OMIM	T	6.5376	Migraine, familial hemiplegic, 2	Familial hemiplegic migraines	D020325	Hemiplegic migraine, familial	D020325	P50993	1
SP	T	2.0627	Long QT syndrome type 1	Long QT syndrome 1	D029597	Long QT syndrome 1	D029597	P51787	1
OMIM	T	exact	Long QT syndrome 1	Long QT syndrome 1	D029597	Long QT syndrome 1	D029597	P51787	1
SP	T	exact	Jervell and Lange-Nielsen syndrome	Jervell and Lange-Nielsen syndrome	D029593	Jervell and Lange-Nielsen syndrome	D029593	P51787	2
OMIM	T	exact	Jervell and Lange-Nielsen syndrome	Jervell and Lange-Nielsen syndrome	D029593	Jervell and Lange-Nielsen syndrome	D029593	P51787	2
SP	T	exact	Atrial fibrillation	Atrial fibrillation	D001281	Genetic disease, inborn | Atrial fibrillation	D030342 D001281	P51787	3
OMIM	T	2.2346	Atrial fibrillation, autosomal dominant	Fibrillations, atrial	D001281	Genetic disease, inborn | Atrial fibrillation	D030342 D001281	P51787	3
SP	T	0.8697	Bartter syndrome type 3	Syndrome, Bartter	D001477	Bartter syndrome	D001477	P51801	1
OMIM	T	2.3043	Bartter syndrome, classic	Syndrome, Bartter	D001477	Bartter syndrome	D001477	P51801	1
SP	T	-2.0166	Hemochromatosis type 2b	Hemochromatoses	D006432	Hemochromatosis	D006432	P81172	1
OMIM	T	1.8331	Hemochromatosis, juvenile	Hemochromatoses	D006432	Hemochromatosis	D006432	P81172	1
SP	T	exact	Pachyonychia congenita type 2	Type 2 pachyonychia congenita	D053549	Pachyonychia congenita, type 2	D053549	Q04695	1
OMIM	T	exact	Pachyonychia congenita, type 2	Pachyonychia congenita, type 2	D053549	Pachyonychia congenita, type 2	D053549	Q04695	1
SP	T	exact	Thyroid dysgenesis	Thyroid dysgenesis	D050033	Thyroid dysgenesis	D050033	Q06710	1
OMIM	T	exact	Thyroid dysgenesis	Thyroid dysgenesis	D050033	Thyroid dysgenesis	D050033	Q06710	1

SP	T	exact	Zellweger syndrome	Zellweger syndrome	D015211	Zellweger syndrome	D015211	Q13608	2
OMIM	T	exact	Zellweger syndrome	Zellweger syndrome	D015211	Zellweger syndrome	D015211	Q13608	2
SP	T	exact	Hypokalemic periodic paralysis	Hypokalemic periodic paralysis	D020514	Hypokalemic periodic paralysis	D020514	Q13698	1
OMIM	T	exact	Hypokalemic periodic paralysis	Hypokalemic periodic paralysis	D020514	Hypokalemic periodic paralysis	D020514	Q13698	1
SP	T	1.124	Malignant hyperthermia susceptibility 5	Malignant hyperthermias	D008305	Malignant hyperthermia | Genetic predisposition to disease	D008305 D020022	Q13698	2
OMIM	T	1.124	Malignant hyperthermia, susceptibility to, 5	Malignant hyperthermias	D008305	Malignant hyperthermia | Genetic predisposition to disease	D008305 D020022	Q13698	2
SP	T	exact	Cleidocranial dysplasia	Cleidocranial dysplasia	D002973	Cleidocranial dysplasia	D002973	Q13950	1
OMIM	T	exact	Cleidocranial dysplasia	Cleidocranial dysplasia	D002973	Cleidocranial dysplasia	D002973	Q13950	1
SP	T	1.3122	Ocular coloboma	Colobomas	D003103	Coloboma	D003103	Q15465	1
OMIM	T	1.3122	Coloboma, ocular	Colobomas	D003103	Coloboma	D003103	Q15465	1
SP	T	-1.0034	Holoprosencephaly type 3	Holoprosencephalies	D016142	Holoprosencephaly	D016142	Q15465	2
OMIM	T	1.882	Holoprosencephaly 3	Holoprosencephalies	D016142	Holoprosencephaly	D016142	Q15465	2
SP	T	3.2182	Hereditary multiple exostoses type 1	Exostosis, hereditary multiple	D005097	Hereditary multiple exostoses	D005097	Q16394	1
OMIM	T	exact	Multiple cartilaginous exostoses	Multiple cartilaginous exostoses	D005097	Hereditary multiple exostoses	D005097	Q16394	1
SP	T	exact	Chondrosarcoma	Chondrosarcoma	D002813	Chondrosarcoma	D002813	Q16394	3
OMIM	T	exact	Chondrosarcoma	Chondrosarcoma	D002813	Chondrosarcoma	D002813	Q16394	3
SP	T	exact	Dyslexia	Dyslexia	D004410	Dyslexia | Genetic predisposition to disease	D004410 D020022	Q5VV43	1
OMIM	T	-1.3579	Reading disability, specific, 2	Developmental reading disabilities	D004410	Dyslexia | Genetic predisposition to disease	D004410 D020022	Q5VV43	1
SP	T	-0.7092	Autosomal recessive osteopetrosis	Osteopetrosis	D010022	Osteopetrosis | Genetic disease, inborn	D010022 D030342	Q86WC4	1
OMIM	T	4.0659	Albers-Schonberg disease, autosomal recessive	Disease, Albers-Schoenberg	D010022	Osteopetrosis | Genetic disease, inborn	D010022 D030342	Q86WC4	1

SP	T	2.7339	Bardet-Biedl syndrome type 10	Bardet-Biedl syndrome	D020788	Bardet-Biedl syndrome	D020788	Q8TAM1	1
OMIM	T	exact	Bardet-Biedl syndrome	Bardet-Biedl syndrome	D020788	Bardet-Biedl syndrome	D020788	Q8TAM1	1
SP	T	0.7775	Primary open angle glaucoma type 1E	Glaucoma, open-angle	D005902	Open angle glaucoma	D005902	Q96CV9	1
OMIM	T	6.7831	Glaucoma, primary open angle	Glaucoma, open-angle	D005902	Open angle glaucoma	D005902	Q96CV9	1
SP	T	-1.4367	Severe combined immunodeficiency with sensitivity to ionizing radiation	Immunodeficiencies, severe combined	D016511	Severe combined immunodeficiency	D016511	Q96SD1	1
OMIM	T	4.9303	Severe combined immunodeficiency, partial	Immunodeficiencies, severe combined	D016511	Severe combined immunodeficiency	D016511	Q96SD1	1
SP	T	-1.5294	Athabascan SCID	SCID	D016511	Severe combined immunodeficiency	D016511	Q96SD1	2
OMIM	T	4.9303	Severe combined immunodeficiency, partial	Immunodeficiencies, severe combined	D016511	Severe combined immunodeficiency	D016511	Q96SD1	2
SP	T	4.9303	Partial severe combined immunodeficiency	Immunodeficiencies, severe combined	D016511	Severe combined immunodeficiency	D016511	Q96SD1	3
OMIM	T	4.9303	Severe combined immunodeficiency, partial	Immunodeficiencies, severe combined	D016511	Severe combined immunodeficiency	D016511	Q96SD1	3
SP	T	exact	Macular corneal dystrophy	Macular dystrophy, corneal	D003317	Macular dystrophy, corneal	D003317	Q9GZX3	1
OMIM	T	6.1515	Corneal dystrophy, macular type	Macular dystrophy, corneal	D003317	Macular dystrophy, corneal	D003317	Q9GZX3	1
SP	T	-2.2768	Congenital muscular dystrophy type 1C	Dystrophies, muscular	D009136	Muscular dystrophy	D009136	Q9H9S5	1
OMIM	T	-1.1525	Muscular dystrophy, congenital, 1C	Dystrophies, muscular	D009136	Muscular dystrophy	D009136	Q9H9S5	1
SP	T	3.0717	Limb-girdle muscular dystrophy type 2I	Muscular dystrophies, limb girdle	D049288	Limb-girdle muscular dystrophy	D049288	Q9H9S5	2
OMIM	T	3.0717	Muscular dystrophy, limb-girdle, type 2I	Muscular dystrophies, limb girdle	D049288	Limb-girdle muscular dystrophy	D049288	Q9H9S5	2
SP	T	-1.0666	Hemochromatosis type 4	Hemochromatoses	D006432	Hemochromatosis	D006432	Q9NP59	1

OMIM	T	-0.7757	Hemochromatosis, autosomal dominant	Hemochromatoses	D006432	Hemochromatosis	D006432	Q9NP59	1
SP	T	exact	Sialuria	Sialuria	D029461	Sialuria	D029461	Q9Y223	1
OMIM	T	exact	Sialuria	Sialuria	D029461	Sialuria	D029461	Q9Y223	1
SP	F	-3.7089	Miyoshi myopathy	Myopathy	D009135	Distal muscular dystrophy	D049310	O75923	2
OMIM	T	-0.8909	Muscular dystrophy, distal, late-onset, autosomal recessive	Distal muscular dystrophy	D049310	Distal muscular dystrophy	D049310	O75923	2
SP	T	0.6097	Stem cell myeloproliferative disorder	Disorder, myeloproliferative	D009196	Neoplastic syndrome, hereditary | Lymphoma, lymphoblastic | Myeloproliferative disorder | Eosinophilia	D009386 D016401 D009196 D004802	O95684	1
OMIM	no OMIM					Neoplastic syndrome, hereditary | Lymphoma, lymphoblastic | Myeloproliferative disorder | Eosinophilia	D009386 D016401 D009196 D004802	O95684	1
SP	T	4.5277	Oral squamous cell carcinoma	Carcinoma, squamous cell	D002294	Squamous cell carcinoma | Oral cancer	D002294 D009062	P01112	4
OMIM	no OMIM					Squamous cell carcinoma | Oral cancer	D002294 D009062	P01112	4
SP	F	5.3478	High density lipoprotein deficiency type 2	High density lipoprotein deficiency, type I	D013631	Hypoalphalipoproteinemia, familial	D052456	P02647	1
OMIM	T	exact	Hypoalphalipoproteinemia, familial	Hypoalphalipoproteinemia, familial	D052456	Hypoalphalipoproteinemia, familial	D052456	P02647	1
SP	F	-5.9912	Systemic non-neuropathic amyloidosis	Amyloidoses	D000686	Amyloidosis, familial	D028226	P02647	3
OMIM	T	3.5746	Amyloidosis, familial renal	Familial amyloidoses	D028226	Amyloidosis, familial	D028226	P02647	3
SP	T	-1.0552	Senile cataract	Cataract	D002386	Genetic predisposition to disease | Cataract	D020022 D002386	P07315	3
OMIM	no OMIM					Genetic predisposition to disease | Cataract	D020022 D002386	P07315	3

SP	F	-5.5317	Hypertrophic cardiomyopathy with mid-left ventricular chamber type 1	Hypertrophic cardiomyopathies	D002312	Cardiomyopathy, hypertrophic, familial	D024741	P08590	2
OMIM	T	4.6686	Cardiomyopathy, familial hypertrophic, 8	Cardiomyopathies, familial hypertrophic	D024741	Cardiomyopathy, hypertrophic, familial	D024741	P08590	2
SP	F	3.4296	Glycogen storage disease type 2	Glycogen storage disease type I	D005953	Glycogen storage disease type II	D006009	P10253	1
OMIM	T	exact	Pompe disease	Pompe disease	D006009	Glycogen storage disease type II	D006009	P10253	1
SP	F	-8.4762	Osteoglophonic dysplasia	Bone dysplasia	D001848	Dwarfism | Osteochondrodysplasia | Craniosynostosis	D004392 D010009 D003398	P11362	4
OMIM	T	-2.0749	Osteoglophonic dwarfism	Dwarfism	D004392	Dwarfism | Osteochondrodysplasia | Craniosynostosis	D004392 D010009 D003398	P11362	4
SP	F	-22.885	Non-syndromic trigonocephaly	Lymphoma, non hodgkin's	D008228	Craniosynostosis	D003398	P11362	5
OMIM	T	-2.2773	Craniosynostosis, metopic	Craniosynostoses	D003398	Craniosynostosis	D003398	P11362	5
SP	T	0.6097	Stem cell myeloproliferative disorder	Disorder, myeloproliferative	D009196	Neoplastic syndrome, hereditary | Lymphoma, lymphoblastic | Myeloproliferative disorder | Eosinophilia	D009386 D016401 D009196 D004802	P11362	7
OMIM	no OMIM					Neoplastic syndrome, hereditary | Lymphoma, lymphoblastic | Myeloproliferative disorder | Eosinophilia	D009386 D016401 D009196 D004802	P11362	7
SP	T	exact	Epidermolysis bullosa simplex	Epidermolysis bullosa simplex	D016110	Epidermolysis bullosa simplex	D016110	P13647	1
OMIM	F	exact	Epidermolysis bullosa dystrophica, cockayne-touraine type	Epidermolysis bullosa dystrophica, cockayne-touraine type	D016108	Epidermolysis bullosa simplex	D016110	P13647	1
SP	T	exact	Schizophrenia	Schizophrenia	D012559	Schizophrenia	D012559	P21918	1
OMIM	no OMIM					Schizophrenia	D012559	P21918	1
SP	T	exact	Blepharospasm	Blepharospasm	D001764	Blepharospasm	D001764	P21918	2
OMIM	F	-0.7357	Blepharospasm, benign essential	essential tremors, benign	D020329	Blepharospasm	D001764	P21918	2

SP	F	2.3495	Non-bullous congenital ichthyosiform erythroderma	Congenital ichthyosiform erythroderma	D016113	Nonbullous congenital ichthyosiform erythroderma	D017490	P22735	2
OMIM	T	8.1821	Ichthyosiform erythroderma, congenital, nonbullous, 1	Ichthyosiform erythroderma, nonbullous congenital	D017490	Nonbullous congenital ichthyosiform erythroderma	D017490	P22735	2
SP	F	2.4413	Rhabdomyosarcoma 2	Rhabdomyosarcoma	D012208	Alveolar rhabdomyosarcoma	D018232	P23760	4
OMIM	T	exact	Rhabdomyosarcoma, alveolar	Rhabdomyosarcoma, alveolar	D018232	Alveolar rhabdomyosarcoma	D018232	P23760	4
SP	T	-3.8405	Posterior polymorphous corneal dystrophy	Hereditary corneal dystrophies	D003317	Corneal dystrophy, hereditary	D003317	P25067	1
OMIM	T	0.5274	Corneal dystrophy, hereditary polymorphous posterior	Hereditary corneal dystrophies	D003317	Corneal dystrophy, hereditary	D003317	P25067	1
SP	T	-0.8576	Tumor development	Tumors	D009369	Neoplasms	D009369	P35222	1
OMIM	no OMIM					Neoplasms	D009369	P35222	1
SP	F	-6.6606	Autosomal dominant potassium-aggravated myotonia	Myotonias	D009222	Myotonic disorder	D020967	P35499	4
OMIM	T	exact	Myotonia fluctuans	Myotonia fluctuans	D020967	Myotonic disorder	D020967	P35499	4
SP	T	exact	Congenital myasthenic syndrome	Congenital myasthenic syndrome	D020294	Congenital myasthenic syndrome	D020294	P35499	5
OMIM	T	-5.1729	Myasthenic syndrome due to mutation in scn4a	Congenital myasthenic syndrome	D020294	Congenital myasthenic syndrome	D020294	P35499	5
SP	F	-1.1235	MASS syndrome	MASS behaviors	D008399	Bone diseases, developmental | Heart defects, congenital | Abnormalities, multiple | Genetic diseases, inborn | Connective tissue diseases	D001848 D006330 D000015 D030342 D003240	P35555	5
OMIM	T	3.8553	Overlap connective tissue disease	Diseases, connective tissue	D003240	Bone diseases, developmental | Heart defects, congenital | Aabnormalities, multiple | Genetic diseases, inborn | Connective tissue diseases	D001848 D006330 D000015 D030342 D003240	P35555	5

SP	F	-0.1875	Juvenile polyposis syndrome	Familial polyposis syndrome	D011125	Intestinal polyposis | Neoplastic syndrome, hereditary | Gastrointestinal neoplasms | Hamartomas	D044483 D009386 D005770 D006222	P36894	1
OMIM	T	4.4418	Polyposis, juvenile intestinal	Intestinal polyposis	D044483	Intestinal polyposis | Neoplastic syndrome, hereditary | Gastrointestinal neoplasms | Hamartomas	D044483 D009386 D005770 D006222	P36894	1
SP	T	-1.9767	Maternal acute fatty liver of pregnancy	Liver, fatty	D005234	Pregnancy complications | Fatty liver	D011248 D005234	P40939	3
OMIM	F	-7.5206	lchad deficiency	Deficiency diseases	D003677	Pregnancy complications | Fatty liver	D011248 D005234	P40939	3
SP	T	-0.0964	Pro-lymphocytic T-cell leukemia	T-cell leukemia	D015458	Leukemia, T-cell	D015458	P46736	1
OMIM	no OMIM					Leukemia, T-cell	D015458	P46736	1
SP	T	-2.9147	Frontotemporal dementia	Lobar degenerations, frontotemporal	D003704	Frontotemporal lobar degeneration	D003704	P49768	2
OMIM	T	exact	Frontotemporal lobar degeneration	Frontotemporal lobar degeneration	D003704	Frontotemporal lobar degeneration	D003704	P49768	2
SP	F	4.3752	Myotonic dystrophy 2	Dystrophies, myotonic	D009223	Proximal myotonic myopathy	D020967	P62633	1
OMIM	T	exact	Proximal myotonic myopathy	Proximal myotonic myopathy	D020967	Proximal myotonic myopathy	D020967	P62633	1
SP	T	1.617	Non-syndromal X-linked mental retardation	Mental retardation, X linked	D038901	Mental retardation, X-linked	D038901	P98174	2
OMIM	no OMIM					Mental retardation, X-linked	D038901	P98174	2
SP	F	-8.7353	Steatocystoma multiplex	Mononeuropathy multiplex	D020422	Skin disease, genetic | Sebaceous cysts	D012873 D004814	Q04695	2
OMIM	T	4.6464	Sebaceous cysts, multiple	Cysts, sebaceous	D004814	Skin disease, genetic | Sebaceous cysts	D012873 D004814	Q04695	2
SP	T	exact	Acute promyelocytic leukemia	Acute promyelocytic leukemia	D015473	Acute promyelocytic leukemia	D015473	Q05516	1
OMIM	no OMIM					Acute promyelocytic leukemia	D015473	Q05516	1

SP	T	-2.7175	Myokymia with periodic ataxia	Myokymia	D020385	Genetic disease, inborn | Neuromuscular disease | Ataxia | Myokymia	D030342 D009468 D001259 D020385	Q09470	1
OMIM	T	exact	Myokymia	Myokymia	D020385	Genetic disease, inborn | Neuromuscular disease | Ataxia | Myokymia	D030342 D009468 D001259 D020385	Q09470	1
SP	T	3.8352	X-linked mental retardation in Xq13	Mental retardation, X linked	D038901	Mental retardation, X-linked	D038901	Q14202	1
OMIM	no OMIM					Mental retardation, X-linked	D038901	Q14202	1
SP	T	exact	Endometrial stromal tumors	Endometrial stromal tumors	D036821	Endometrial stromal tumors	D036821	Q15022	1
OMIM	no OMIM					Endometrial stromal tumors	D036821	Q15022	1
SP	T	4.228	Form of B-cell leukemia	B-cell leukemias	D015448	B-cell leukemia	D015448	Q16633	1
OMIM	no OMIM					B-cell leukemia	D015448	Q16633	1
SP	T	-4.7981	Ataxia-oculomotor apraxia 1	Apraxia	D001072	Early onset cerebellar ataxia | Peripheral neuropathies | Apraxia, motor | Hypoalbuminemia	D013132 D010523 D001072 D034141	Q7Z2E3	1
OMIM	T	4.4515	Cerebellar ataxia, early-onset, with hypoalbuminemia	Cerebellar ataxia, early onset	D013132	Early onset cerebellar ataxia | Peripheral neuropathies | Apraxia, motor | Hypoalbuminemia	D013132 D010523 D001072 D034141	Q7Z2E3	1
SP	T	-5.0948	Female-specific osteoarthritis susceptibility	Osteoarthritis	D010003	Genetic predisposition to disease | Osteoarthritis	D020022 D010003	Q92765	1
OMIM	T	exact	Osteoarthritis	Osteoarthritis	D010003	Genetic predisposition to disease | Osteoarthritis	D020022 D010003	Q92765	1
SP	F	-5.223	Omenn syndrome	Syndromes	D013577	Severe combined immunodeficiency	D016511	Q96SD1	4
OMIM	T	3.4379	Severe combined immunodeficiency with hypereosinophilia	Immunodeficiencies, severe combined	D016511	Severe combined immunodeficiency	D016511	Q96SD1	4
SP	T	exact	Zellweger syndrome	Zellweger syndrome	D015211	Zellweger syndrome	D015211	Q99424	1
OMIM	no OMIM					Zellweger syndrome	D015211	Q99424	1

SP	T	exact	Breast cancer	Breast cancer	D001943	Breast cancer	D001943	Q9H6U6	1
OMIM	no OMIM					Breast cancer	D001943	Q9H6U6	1
SP	F	-7.6241	Walker-Warburg syndrome	Syndromes	D013577	Genetic disease, inborn | Abnormalities, multiple | Muscular dystrophy | Brain diseases | Retinal dysplasia	D030342 D000015 D009136 D001927 D015792	Q9H9S5	4
OMIM	T	-2.2899	Hydrocephalus, agyria, and retinal dysplasia	Retinal dysplasia	D015792	Genetic disease, inborn | Abnormalities, multiple | Muscular dystrophy | Brain diseases | Retinal dysplasia	D030342 D000015 D009136 D001927 D015792	Q9H9S5	4
SP	T	-2.2617	Azoospermia or oligospermia	Oligospermia	D009845	Azoospermia | Oligospermia	D053713 D009845	Q9NQZ3	1
OMIM	no OMIM					Azoospermia | Oligospermia	D053713 D009845	Q9NQZ3	1
SP	T	exact	Chronic neutrophilic leukemia	Chronic neutrophilic leukemia	D015467	Chronic neutrophilic leukemia	D015467	Q9NVA2	1
OMIM	no OMIM					Chronic neutrophilic leukemia	D015467	Q9NVA2	1
SP	F	-3.2089	Nonaka myopathy	Myopathy	D009135	Distal myopathy	D049310	Q9Y223	3
OMIM	T	1.4836	Nonaka distal myopathy	Myopathies, distal	D049310	Distal myopathy	D049310	Q9Y223	3
SP	F	-6.139	Lacticacidemia	Acidosis, lactic	D000140	Pyruvate dehydrogenase complex deficiency disease	D015325	O00330	1
OMIM	T	-3.8486	Pyruvate dehydrogenase E3-binding protein deficiency	Pyruvate dehydrogenase complex deficiency disease	D015325	Pyruvate dehydrogenase complex deficiency disease	D015325	O00330	1
SP	T	-4.5341	Variety of human tumors	Tumors	D009369	Neoplasms	D009369	P01112	2
OMIM	no OMIM					Neoplasms	D009369	P01112	2
SP	F	-5.4305	Kniest syndrome	Syndromes	D013577	Genetic disease, inborn | Abnormalities, multiple | Osteochondrodysplasia | Dwarfism | Craniofacial abnormalities	D030342 D000015 D010009 D004392 D019465	P02458	5

OMIM	T	-5.1473	Metatropic dwarfism, type II	Dwarfism	D004392	Genetic disease, inborn | Abnormalities, multiple | Osteochondrodysplasia | Dwarfism | Craniofacial abnormalities	D030342 D000015 D010009 D004392 D019465	P02458	5
SP	T	-3.1863	Osteoarthritis with mild chondrodysplasia	Osteoarthritis	D010003	Genetic disease, inborn | Osteochondrodysplasia | Osteoarthritis	D030342 D010009 D010003	P02458	7
OMIM	T	-3.1863	Osteoarthritis with mild chondrodysplasia	Osteoarthritis	D010003	Genetic disease, inborn | Osteochondrodysplasia | Osteoarthritis	D030342 D010009 D010003	P02458	7
SP	T	-5.5957	Coppock-like cataract	Cataract	D002386	Genetic disease, inborn | Cataract	D030342 D002386	P07315	2
OMIM	T	-5.2227	Cataract, embryonic nuclear	Cataract	D002386	Genetic disease, inborn | Cataract	D030342 D002386	P07315	2
SP	T	-5.319	Autosomal recessive severe combined immunodeficiency T-cell-negative/B-cell-positive/NK cell-positive	Immunodeficiencies, severe combined	D016511	Severe combined immunodeficiency	D016511	P08575	1
OMIM	T	-5.319	Severe combined immunodeficiency, autosomal recessive, T cell-negative, B cell-positive, NK cell-positive	Immunodeficiencies, severe combined	D016511	Severe combined immunodeficiency		P08575	1
SP	F	-8.3797	Trismus-pseudocamptodactyly syndrome	Syndromes	D013577	Abnormalities, multiple | Arthrogryposis	D000015 D001176	P13535	2
OMIM	T	-3.9776	Arthrogryposis, distal, type 7	Arthrogryposis	D001176	Abnormalities, multiple | Arthrogryposis	D000015 D001176	P13535	2
SP	T	-4.79	Posterior polymorphous corneal dystrophy 2	Hereditary corneal dystrophies	D003317	Corneal dystrophy, hereditary	D003317	P25067	3
OMIM	T	-4.79	Corneal dystrophy, posterior polymorphous, 2	Hereditary corneal dystrophies	D003317	Corneal dystrophy, hereditary	D003317	P25067	3
SP	T	-5.269	Familial hyperinsulinemic hypoglycemia type 3	Hypoglycemia	D007003	Metabolism, inborn errors | Hyperinsulinism | Hypoglycemia	D008661 D006946 D007003	P35557	2

OMIM	T	-4.8927	Hyperinsulinemic hypoglycemia, familial, 3	Hypoglycemia	D007003	Metabolism, inborn errors | Hyperinsulinism | Hypoglycemia	D008661 D006946 D007003	P35557	2
SP	T	-3.8688	Alternating hemiplegia of childhood	Hemiplegia	D006429	Genetic disease, inborn | Hemiplegia, infantile	D030342 D006429	P50993	2
OMIM	T	-3.8688	Alternating hemiplegia of childhood	Hemiplegia	D006429	Genetic disease, inborn | Hemiplegia, infantile	D030342 D006429	P50993	2
SP	T	-2.9542	The uterine cervix and in psoriasis vulgaris	Cancer of the uterine cervix	D002583	Psoriasis | Uterine cervical diseases | Uterine cervical neoplasms	D011565 D002577 D002583	Q04695	3
OMIM	no OMIM					Psoriasis | Uterine cervical diseases | Uterine cervical neoplasms	D011565 D002577 D002583	Q04695	3
SP	F	-8.77	Peroxisome biogenesis disorder complementation group 4	T-group	D012681	Peroxisomal disorder	D018901	Q13608	1
OMIM	T	-7.628	Peroxisomal assembly factor 2	Peroxisomal disorder	D018901	Peroxisomal disorder	D018901	Q13608	1
SP	F	-9.4008	Triphalangeal thumb-polysyndactyly syndrome	Syndromes	D013577	Limb deformities, congenital | Genetic disease, inborn | Polydactyly | Syndactyly	D017880 D030342 D017689 D013576	Q15465	4
OMIM	T	-4.0164	Triphalangeal thumb with polysyndactyly	Polysyndactyly	D013576	Limb deformities, congenital | Genetic disease, inborn | Polydactyly | Syndactyly	D017880 D030342 D017689 D013576	Q15465	4
SP	F	-9.7325	AMME complex	WAGR complex	D017624	Genetic disease, X-linked | Abnormalities, multiple | Nephritis, hereditary | Elliptocytosis, hereditary | Craniofacial abnormalities | Mental retardation, X-linked	D040181 D000015 D009394 D004612 D019465 D038901	Q9Y4X0	1

OMIM	T	-5.5813	Alport syndrome, mental retardation, midface hypoplasia, and elliptocytosis	Alport syndromes	D009394	Genetic disease, x-linked | Abnormalities, multiple | Nephritis, hereditary | Elliptocytosis, hereditary | Craniofacial abnormalities | Mental retardation, X-linked	D040181 D000015 D009394 D004612 D019465 D038901	Q9Y4X0	1
SP	F	-7.9264	Peeling skin syndrome acral type	Skin diseases	D012871	Skin disease, genetic | Skin abnormalities | Skin disease, vesiculobullous	D012873 D012868 D012872	O43548	1
OMIM	F	-7.9264	Peeling skin syndrome, acral type	Skin diseases	D012871	Skin disease, genetic | Skin abnormalities | Skin disease, vesiculobullous	D012873 D012868 D012872	O43548	1
SP	F	-5.4305	Costello syndrome	Syndromes	D013577	Genetic disease, inborn | Abnormalities, multiple | Craniofacial abnormalities | Skin abnormalities | Heart defects, congenital	D030342 D000015 D019465 D012868 D006330	P01112	1
OMIM	F	-5.4305	Costello syndrome	Syndromes	D013577	Genetic disease, inborn | Abnormalities, multiple | Craniofacial abnormalities | Skin abnormalities | Heart defects, congenital	D030342 D000015 D019465 D012868 D006330	P01112	1
SP	F	-7.412	Variety of chondrodysplasia including hypochondrogenesis and osteoarthritis	Osteoarthritis	D010003	Osteochondrodysplasia	D010009	P02458	1
OMIM	no OMIM					Osteochondrodysplasia	D010009	P02458	1
SP	F	-9.5244	Strudwick type spondyloepimetaphyseal dysplasia	Bone dysplasia	D001848	Genetic disease, inborn | Abnormalities, multiple | Osteochondrodysplasia | Dwarfism	D030342 D000015 D010009 D004392	P02458	3

OMIM	F	-4.9305	Strudwick syndrome	Syndromes	D013577	Genetic disease, inborn | Abnormalities, multiple | Osteochondrodysplasia | Dwarfism	D030342 D000015 D010009 D004392	P02458	3
SP	F	-9.1653	Achondrogenesis hypochondrogenesis type 2	2, neurofibromatosis type	D016518	Genetic disease, inborn | Osteochondrodysplasia | Dwarfism | Craniofacial abnormalities	D030342 D010009 D004392 D019465	P02458	4
OMIM	F	-6.656	Achondrogenesis, type II	Type II, neurofibromatosis	D016518	Genetic disease, inborn | Osteochondrodysplasia | Dwarfism | Craniofacial abnormalities	D030342 D010009 D004392 D019465	P02458	4
SP	F	-8.6837	Spondyloperipheral dysplasia	Bone dysplasia	D001848	Genetic disease, inborn | Abnormalities, multiple | Spondyloepiphyseal dysplasia | Sensorineural hearing loss | Dwarfism | Craniofacial abnormalities	D030342 D000015 D010009 D006319 D004392 D019465	P02458	10
OMIM	F	-8.0854	Spondyloperipheral dysplasia with short ulna	Fracture, ulna	D014458	Genetic disease, inborn | Abnormalities, multiple | Spondyloepiphyseal dysplasia | Sensorineural hearing loss | Dwarfism | Craniofacial abnormalities	D030342 D000015 D010009 D006319 D004392 D019465	P02458	10
SP	F	-3.9683	Wagner syndrome type II	Usher syndrome, type II	D052245	Eye disease, hereditary	D015785	P02458	11
OMIM	no OMIM					Eye disease, hereditary	D015785	P02458	11
SP	F	-5.7579	Stickler syndrome type 1	Syndromes	D013577	Genetic disease, inborn | Abnormalities, multiple | Eye disease, hereditary | Bone diseases | Sensorineural hearing loss | Craniofacial abnormalities	D030342 D000015 D015785 D001847 D006319 D019465	P02458	12

OMIM	F	-3.4339	Stickler syndrome, type I	Usher syndrome, type I	D052245	Genetic disease, inborn | Abnormalities, multiple | Eye disease, hereditary | Bone diseases | Sensorineural hearing loss | Craniofacial abnormalities	D030342 D000015 D015785 D001847 D006319 D019465	P02458	12
SP	F	-3.7086	Acid phosphatase deficiency	Acid deficiency, folic	D005494	Lysosomal storage disease	D016464	P11117	1
OMIM	F	-3.7086	Acid phosphatase deficiency	Acid deficiency, folic	D005494	Lysosomal storage disease	D016464	P11117	1
SP	F	-8.6929	Carney complex variant	Migraine variants	D008881	Neoplastic syndromes, hereditary | Abnormalities, multiple	D009386 D000015	P13535	1
OMIM	F	-8.6929	Carney complex variant	Migraine variants	D008881	Neoplastic syndromes, hereditary | Abnormalities, multiple	D009386 D000015	P13535	1
SP	F	-8.1507	Dowling-Degos disease	Diseases	D004194	Skin disease, genetic | Hyperpigmentation	D012873 D017495	P13647	7
OMIM	F	-8.1507	Dowling-Degos disease	Diseases	D004194	Skin disease, genetic | Hyperpigmentation	D012873 D017495	P13647	7
SP	F	-9.3543	Muscle-specific enolase- beta deficiency	Deficiency diseases	D003677	Glycogen storage disease | Myopathy	D006008 D009135	P13929	1
OMIM	F	-7.7061	Enolase 3 deficiency	Antithrombin 3 deficiency	D020152	Glycogen storage disease | Myopathy	D006008 D009135	P13929	1
SP	F	-8.0911	Periventricular nodular heterotopia 1	Nodular lymphomas	D008224	Genetic disease, X-linked | Nervous system malformation	D040181 D009421	P21333	1
OMIM	F	-5.2688	Heterotopia, periventricular, x-linked dominant	Hypophosphatemic rickets, X linked dominant	D053098	Genetic disease, X-linked | Nervous system malformation	D040181 D009421	P21333	1
SP	F	-8.5683	Periventricular nodular heterotopia 4	Nodular lymphomas	D008224	Genetic disease, X-linked | Abnormalities, multiple | Nervous system malformation | Joint hypermobility	D040181 D000015 D009421 D007593	P21333	2

OMIM	F	-3.7967	Heterotopia, periventricular, ehlers-danlos variant	Syndrome, Ehlers-Danlos	D004535	Genetic disease, X-linked | Abnormalities, multiple | Nervous system malformation | Joint hypermobility	D040181 D000015 D009421 D007593	P21333	2
SP	F	-8.6973	Otopalatodigital syndrome type 1	Syndromes	D013577	Genetic disease, X-linked | Abnormalities, multiple | Osteochondrodysplasia | Craniofacial abnormalities	D040181 D000015 D010009 D019465	P21333	3
OMIM	F	-6.6228	OPD syndrome, type 1	Syndromes	D013577	Genetic disease, X-linked | Abnormalities, multiple | Osteochondrodysplasia | Craniofacial abnormalities	D040181 D000015 D010009 D019465	P21333	3
SP	F	-8.7501	Otopalatodigital syndrome type 2	Syndromes	D013577	Genetic disease, X-linked | Abnormalities, multiple | Osteochondrodysplasia | Craniofacial abnormalities	D040181 D000015 D010009 D019465	P21333	4
OMIM	F	-5.723	Cranioorodigital syndrome	Syndromes	D013577	Genetic disease, X-linked | Abnormalities, multiple | Osteochondrodysplasia | Craniofacial abnormalities	D040181 D000015 D010009 D019465	P21333	4
SP	F	-8.4762	Frontometaphyseal dysplasia	Bone dysplasia	D001848	Genetic disease, X-linked | Abnormalities, multiple | Osteochondrodysplasia | Craniofacial abnormalities	D040181 D000015 D010009 D019465	P21333	5
OMIM	F	-8.4762	Frontometaphyseal dysplasia	Bone dysplasia	D001848	Genetic disease, X-linked | Abnormalities, multiple | Osteochondrodysplasia | Craniofacial abnormalities	D040181 D000015 D010009 D019465	P21333	5

SP	F	-5.4305	Cerebrofrontofacial syndrome	Syndromes	D013577	Genetic disease, X-linked | Abnormalities, multiple | Nervous system malformation | Craniofacial abnormalities	D040181 D000015 D009421 D019465	P21333	7
OMIM	F	-5.4305	Cerebrofrontofacial syndrome	Syndromes	D013577	Genetic disease, X-linked | Abnormalities, multiple | Nervous system malformation | Craniofacial abnormalities	D040181 D000015 D009421 D019465	P21333	7
SP	F	-6.9626	Craniofacial-deafness- hand syndrome	Syndromes	D013577	Abnormalities, multiple | Craniofacial abnormalities | Sensorineural hearing loss	D000015 D019465 D006319	P23760	3
OMIM	F	-6.9626	Craniofacial-deafness-hand syndrome	Syndromes	D013577	Abnormalities, multiple | Craniofacial abnormalities | Sensorineural hearing loss	D000015 D019465 D006319	P23760	3
SP	F	-6.7679	Characteristic traits of polycystic ovary syndrome, such as insulin resistance and luteinizing hormon hypersecretion	Syndrome, polycystic ovary	D011085	Insulin resistance | Pituitary LH hypersecretion	D007333 D006964	P26439	2
OMIM	no OMIM					Insulin resistance | Pituitary LH hypersecretion	D007333 D006964	P26439	2
SP	F	-6.9894	Hyperprolinemia type II	Type II, neurofibromatosis	D016518	Amino acid metabolism, inborn error	D000592	P30038	1
OMIM	F	-6.9894	Hyperprolinemia, type II	Type II, neurofibromatosis	D016518	Amino acid metabolism, inborn error	D000592	P30038	1
SP	F	-4.2149	Certain cardiovascular and musculo-skeletal abnormalities observed in Williams-Beuren syndrome	Tyndrome, Williams-Beuren	D018980	Musculoskeletal abnormalities | Cardiovascular abnormalities	D009139 D018376	P35250	1
OMIM	no OMIM					Musculoskeletal abnormalities | Cardiovascular abnormalities	D009139 D018376	P35250	1

SP	F	-4.6331	Autosomal dominant weill-marchesani syndrome	Alport syndrome, autosomal dominant	D009394	Abnormalities, multiple | Connective tissue disease | Bone disease, developmental | Eye disease, hereditary	D000015 D003240 D001848 D015785	P35555	3
OMIM	F	-4.6331	Weill-Marchesani syndrome, autosomal dominant	Alport syndrome, autosomal dominant	D009394	Abnormalities, multiple | Connective tissue disease | Bone disease, developmental | Eye disease, hereditary	D000015 D003240 D001848 D015785	P35555	3
SP	F	-3.8203	Hereditary mixed polyposis syndrome 2	Familial polyposis syndrome	D011125	Intestinal polyposis | Neoplastic syndrome, hereditary | Colonic neoplasms	D044483 D009386 D003110	P36894	3
OMIM	F	-3.8203	Polyposis syndrome, hereditary mixed, 2	Familial polyposis syndrome	D011125	Intestinal polyposis | Neoplastic syndrome, hereditary | Colonic neoplasms	D044483 D009386 D003110	P36894	3
SP	F	-8.2603	Long-chain 3-hydroxyl- coA dehydrogenase deficiency	Deficiencies, glucosephosphate dehydrogenase	D005955	Lipid metabolism, inborn error | Mitochondrial disease	D008052 D028361	P40939	2
OMIM	F	-7.5206	LCHAD deficiency	Deficiency diseases	D003677	Lipid metabolism, inborn error | Mitochondrial disease	D008052 D028361	P40939	2
SP	F	-4.8884	Short QT syndrome type 2	Bowel syndromes, short	D012778	Genetic disease, inborn | Arrhythmia	D030342 D001145	P51787	4
OMIM	F	-4.4169	Short QT syndrome 2	Bowel syndromes, short	D012778	Genetic disease, inborn | Arrhythmia	D030342 D001145	P51787	4
SP	F	-5.672	Norrie disease	Diseases	D004194	Genetic disease, X-linked | Eye disease, hereditary | Retinal dysplasia	D040181 D015785 D015792	Q00604	1
OMIM	F	-3.8874	Episkopi blindness	Blindness	D001766	Genetic disease, X-linked | Eye disease, hereditary | Retinal dysplasia	D040181 D015785 D015792	Q00604	1
SP	F	-6.0099	X-linked familial exudative vitreoretinopathy	Ichthyosis, x-linked	D016114	Genetic disease, X-linked | Eye disease, hereditary | Retinal Dysplasia	D040181 D015785 D015792	Q00604	2

OMIM	F	-2.8909	FEVR, X-linked	Ichthyosis, x-linked	D016114	Genetic disease, X-linked | Eye disease, hereditary | Retinal dysplasia	D040181 D015785 D015792	Q00604	2
SP	F	-8.8007	Axenfeld-Rieger syndrome	Syndromes	D013577	Abnormalities, multiple | Eye disease, hereditary | Glaucoma, angle-closure | Eye abnormalities | Craniofacial abnormalities	D000015 D015785 D015812 D005124 D019465	Q12948	1
OMIM	no OMIM					Abnormalities, multiple | Eye disease, hereditary | Glaucoma, angle-closure | Eye abnormalities | Craniofacial abnormalities	D000015 D015785 D015812 D005124 D019465	Q12948	1
SP	F	-8.7841	Iridogoniodysgenesis anomaly	Anomalies, pupillary	D011681	Abnormalities, multiple | Eye disease, hereditary | Glaucoma, angle-closure | Eye abnormalities	D000015 D015785 D015812 D005124	Q12948	2
OMIM	F	-5.0515	Glaucoma iridogoniodysplasia, familial	glaucoma	D005901	Abnormalities, multiple | Eye disease, hereditary | Glaucoma, angle-closure | Eye abnormalities	D000015 D015785 D015812 D005124	Q12948	2
SP	F	-8.3804	Peters anomaly	Anomalies, pupillary	D011681	Eye disease, hereditary | Eye abnormalities	D015785 D005124	Q12948	3
OMIM	F	-8.3804	Peters anomaly	Anomalies, pupillary	D011681	Eye disease, hereditary | Eye abnormalities	D015785 D005124	Q12948	3
SP	F	-4.16	Autosomal dominant filaminopathy	Dominant parkinsonism, autosomal	D020734	Muscular dystrophy	D009136	Q14315	1
OMIM	F	-4.16	Filaminopathy, autosomal dominant	Dominant parkinsonism, autosomal	D020734	Muscular dystrophy	D009136	Q14315	1
SP	F	-8.9993	Solitary median maxillary central incisor	Disease, maxillary	D008439	Tooth abnormality	D014071	Q15465	3
OMIM	F	-5.223	SMMCI syndrome	Syndromes	D013577	Tooth abnormality	D014071	Q15465	3
SP	F	-4.717	Multiple exostoses observed in Langer-Giedon syndrome	Multiple exostoses	D005097	Exostoses	D005096	Q16394	2

OMIM	no OMIM					Exostoses	D005096	Q16394	2
SP	F	-10.845	Bietti crystalline corneoretinal dystrophy	Diseases, retinal	D012164	Eye disease, hereditary | Retinal degeneration | Corneal dystrophy, hereditary	D015785 D012162 D003317	Q6ZWL3	1
OMIM	F	-4.4198	Bietti tapetoretinal degeneration with marginal corneal dystrophy	Degeneration, tapetoretinal	D012174	Eye disease, hereditary | Retinal degeneration | Corneal dystrophy, hereditary	D015785 D012162 D003317	Q6ZWL3	1
SP	F	-9.7806	Coenzyme Q10 deficiency	Deficiency diseases	D003677	Abnormalities, multiple | Brain diseases, metabolic, inborn | Cerebellar ataxia	D000015 D020739 D002524	Q7Z2E3	2
OMIM	F	-8.8041	CoQ10 deficiency, primary	Deficiency diseases	D003677	Abnormalities, multiple | Brain diseases, metabolic, inborn | Cerebellar ataxia	D000015 D020739 D002524	Q7Z2E3	2
SP	F	-7.7281	ICOS deficiency	Deficiency diseases	D003677	Common variable immunodeficiency	D017074	Q9Y6W8	1
OMIM	F	-7.7281	ICOS deficiency	Deficiency diseases	D003677	Common variable immunodeficiency	D017074	Q9Y6W8	1
SP	F	-2.3535	Distal myopathy with anterior tibial onset	Anterior tibial syndromes	D000868	Distal muscular dystrophy	D049310	O75923	3
OMIM	F	-2.3535	Myopathy, distal, with anterior tibial onset	Anterior tibial syndromes	D000868	Distal muscular dystrophy	D049310	O75923	3
SP	F	-10.055	Platyspondylic lethal skeletal dysplasia Torrance type	Lethal catatonia	D002389	Genetic disease, inborn | Osteochondrodysplasia | Dwarfism	D030342 D010009 D004392	P02458	8
OMIM	F	-1.0708	Thanatophoric dysplasia, Torrance variant	Dysplasia, thanatophoric	D013796	Genetic disease, inborn | Osteochondrodysplasia | Dwarfism	D030342 D010009 D004392	P02458	8
SP	F	-0.3982	Cryptogenic cirrhosis	Cirrhosis	D005355	Liver cirrhosis	D008103	P05783	1
OMIM	F	1.8468	Cirrhosis, familial	Cirrhosis	D005355	Liver cirrhosis	D008103	P05783	1
SP	F	-0.3982	Cryptogenic cirrhosis	Cirrhosis	D005355	Liver cirrhosis	D008103	P05787	1
OMIM	F	1.8468	Cirrhosis, familial	Cirrhosis	D005355	Liver cirrhosis	D008103	P05787	1
SP	T	-6.3083	Gelatinous drop-like corneal dystrophy	Hereditary corneal dystrophies	D003317	Corneal dystrophy, hereditary	D003317	P09758	1

OMIM	F	0.5231	Amyloidosis, corneal	Amyloidoses	D000686	Corneal dystrophy, hereditary	D003317	P09758	1
SP	F	-2.1702	Stem cell leukemia lymphoma syndrome	Adult T-cell leukemia-lymphoma	D015460	Neoplastic syndrome, hereditary | Lymphoma, lymphoblastic | Myeloproliferative disorder | Eosinophilia	D009386 D016401 D009196 D004802	P11362	6
OMIM	no OMIM					Neoplastic syndrome, hereditary | Lymphoma, lymphoblastic | Myeloproliferative disorder | Eosinophilia	D009386 D016401 D009196 D004802	P11362	6
SP	T	0.3964	Epidermolysis bullosa simplex Weber-Cockayne type	Epidermolysis bullosa simplex	D016110	Weber-Cockayne syndrome	D016110	P13647	4
OMIM	F	exact	Epidermolysis bullosa dystrophica, Cockayne-Touraine type	Epidermolysis bullosa dystrophica, Cockayne-Touraine type	D016108	Weber-Cockayne syndrome	D016110	P13647	4
SP	F	-1.9877	Hematopoietic tumors such as B-cell lymphomas	Lymphoma, B cell	D016393	Hematologic neoplasms	D019337	P26196	1
OMIM	no OMIM					Hematologic neoplasms	D019337	P26196	1
SP	T	-6.2552	Shprintzen-Goldberg craniosynostosis syndrome	Craniosynostoses	D003398	Abnormalities, multiple | Connective tissue disease | Bone disease, developmental | Craniosynostosis | Heart defects, congenital | Eye disease, hereditary	D000015 D003240 D001848 D003398 D006330 D015785	P35555	4
OMIM	F	-1.0442	Craniosynostosis with arachnodactyly and abdominal hernias	Abdominal hernias	D046449	Abnormalities, multiple | Connective tissue disease | Bone disease, developmental | Craniosynostosis | Heart defects, congenital | Eye disease, hereditary	D000015 D003240 D001848 D003398 D006330 D015785	P35555	4
SP	T	-0.974	Maturity onset diabetes of the young type 2	Maturity onset diabetes mellitus	D003924	Genetic disease, inborn | Maturity-onset diabetes mellitus	D030342 D003924	P35557	1

OMIM	F	exact	Diabetes, gestational	Diabetes, gestational	D016640	Genetic disease, inborn | Maturity-onset diabetes mellitus	D030342 D003924	P35557	1
SP	F	0.0504	Trifunctional protein deficiency	Protein deficiency	D011488	Lipid metabolism, inborn error | Mitochondrial disease	D008052 D028361	P40939	1
OMIM	F	0.0504	Trifunctional protein deficiency	Protein deficiency	D011488	Lipid metabolism, inborn error | Mitochondrial disease	D008052 D028361	P40939	1
SP	F	-8.8076	Microphthalmia syndromic type 3	Type 3 gaucher disease	D005776	Anophthalmia | Microphthalmos	D000853 D008850	P48431	1
OMIM	F	1.37	Microphthalmia and esophageal atresia syndrome	Esophageal atresia	D004933	Anophthalmia | Microphthalmos	D000853 D008850	P48431	1
SP	F	-5.8108	Leukoencephalopathy with vanishing white matter	Spongy disease of white matter	D017825	Hereditary central nervous system demyelinating diseases	D020279	P49770	1
OMIM	F	-2.2785	Childhood ataxia with central nervous system hypomyelinization	Central nervous system diseases	D002493	Hereditary central nervous system demyelinating diseases	D020279	P49770	1
SP	no result		Ovarioleukodystrophy			Hereditary central nervous system demyelinating diseases | Ovarian failure, premature	D020279 D016649	P49770	2
OMIM	F	-2.2785	Childhood ataxia with central nervous system hypomyelinization	Central nervous system diseases	D002493	Hereditary central nervous system demyelinating diseases | Ovarian failure, premature	D020279 D016649	P49770	2
SP	F	2.5812	Bleeding disorder	Bleeding	D006470	Genetic disease, inborn | Hemorrhagic disorder	D030342 D006474	P51575	1
OMIM	F	-5.5163	Bleeding disorder due to P2RX1 defect	Bleeding	D006470	Genetic disease, inborn | Hemorrhagic disorder	D030342 D006474	P51575	1
SP	F	-9.2414	Aarskog-Scott syndrome	Syndromes	D013577	Genetic disease, x-linked | Abnormalities, multiple | Craniofacial abnormalities | Urogenital abnormalities	D040181 D000015 D019465 D014564	P98174	1

OMIM	F	3.446	Faciogenital dysplasia with attention deficit-hyperactivity disorder	Attention deficit disorders with hyperactivity	D001289	Genetic disease, x-linked | Abnormalities, multiple | Craniofacial abnormalities | Urogenital abnormalities	D040181 D000015 D019465 D014564	P98174	1
SP	F	-8.4762	Gnathodiaphyseal dysplasia	Bone dysplasia	D001848	Genetic disease, inborn | Osteochondrodysplasia	D030342 D010009	Q75V66	1
OMIM	F	-2.1352	Osteogenesis imperfecta with unusual skeletal lesions	Osteogenesis imperfecta	D010013	Genetic disease, inborn | Osteochondrodysplasia	D030342 D010009	Q75V66	1
SP	F	-2.0012	Chromosome 22q13.3 deletion syndrome	Chromosome deletions	D002872	Genetic disease, inborn | Abnormalities, multiple	D030342 D000015	Q8NEU8	1
OMIM	F	-2.0012	Chromosome 22q13.3 deletion syndrome	Chromosome deletions	D002872	Genetic disease, inborn | Abnormalities, multiple	D030342 D000015	Q8NEU8	1
SP	F	-0.7282	Normal pressure glaucoma	Hydrocephalus, normal pressure	D006850	Glaucoma | Genetic predisposition to disease	D005901 D020022	Q96CV9	2
OMIM	F	-3.024	Glaucoma, normal pressure, susceptibility to	Hydrocephalus, normal pressure	D006850	Glaucoma | Genetic predisposition to disease	D005901 D020022	Q96CV9	2
SP	F	-0.5309	Muscle-eye-brain disease	Muscle-liver-brain-eye nanism	D050336	Genetic disease, inborn | Abnormalities, multiple | Muscular dystrophy | Eye diseases | Nervous system diseases	D030342 D000015 D009136 D005128 D009422	Q9H9S5	3
OMIM	F	-0.5309	Muscle-eye-brain disease	Muscle-liver-brain-eye nanism	D050336	Genetic disease, inborn | Abnormalities, multiple | Muscular dystrophy | Eye diseases | Nervous system diseases	D030342 D000015 D009136 D005128 D009422	Q9H9S5	3
SP	F	0.8696	Inclusion body myopathy type 2	Myopathy, inclusion body, sporadic	D018979	Myopathy	D009135	Q9Y223	2
OMIM	F	0.7779	Inclusion body myopathy, autosomal recessive	Myopathy, inclusion body, sporadic	D018979	Myopathy	D009135	Q9Y223	2

Colors correspond to a score threshold of -2.5
True positive SP ∩ OMIM
True Positive SP U OMIM
False negative
True negative
False positive

Additional figure 2, Mottaz *et al.*, 2008

UniProtKB/Swiss-Prot mapping to MeSH:

http://www.biomedcentral.com/content/supplementary/1471-2105-9-s5-s3-s2.html

Additional figure 3 , Mottaz *et al.*, 2008

Regular Expressions used to extract the disease names from the Swiss-Prot disease comment lines

(1) Starter expressions	(2) Specific stop words	(3) Termination term
Cause(s) of /a	susceptibility to	also known as
involved in	development of	but
(can) contribute(s) to	genetic predisposition for	which
associated/association with	developing	an
correlated with	pathogenesis of	due to
responsible for	subset of	in condition(s) such as
contributor to	various types of	.
result(s)/resulting in	some form of	[MIM:
lead(s) to	increased risk of	
induce(s)		
defective in		
individual(s) with		
patient(s) with/suffering from		
reduce(s)		
influence(s)		
deleted in		
down-regulated in		
found in		
implicated in		
predispose(s) to		
favor		
antigen of		
antigen for		
thought to be an		
role in		
could impart		
mediate(s)		
candidate (gene)		

(1) Expressions used to extract the part of the string containing the disease name. (2) Terms removed from the string extracted. (3) Expressions indicating the end of the disease name.

SwissVar documentation page

Global query

The global query enables the user to retrieve Swiss-Prot entries, diseases and variants from a disease, a protein/gene name, a Swiss-Prot accession number, or a variant identifier (FTID or rsID).
If the text entered corresponds to a MeSH disease or if it is a MeSH descriptor identifier (DUI), the returned Swiss-Prot entries and variants are those indexed with the given MeSH descriptors or its children.
If the text is a MIM number or a Swiss-Prot disease, the entries returned are those for which the given disease, or MIM number, has been extracted from the Swiss-Prot disease comment line.
If the text is a gene name, a protein name or an accession number, the entry returned is the protein, its diseases and variants, only if it corresponds to a human protein having at least one variant or one disease association.
If the text is a variant identifier (FTID (UniProtKB) or rsID (dbSNP)), the corresponding protein is returned with the diseases associated to this variant specifically.
If the text entered does not correspond to any identifier, protein or gene name or exact MeSH disease, the proteins returned are the one whose disease (MeSH or disease as extracted from the disease comment line) contains the text.

Disease query

The disease query enables the user to retrieve Swiss-Prot entries and variants from a disease.
If the disease entered corresponds to a MeSH disease or if it is a MeSH descriptor identifier (DUI), the returned Swiss-Prot entries and variants are those indexed with the given MeSH descriptors or its children.
If the disease entered does not correspond to a MeSH term, or if it is a MIM number, the entries returned are those for which the given disease, or MIM number, has been extracted from the Swiss-Prot disease comment line.

Disease textfield

The user can enter one disease or several MeSH descriptor identifiers (DUI) or several MIM numbers separated by spaces.

Disease file upload

The file can contain diseases or MeSH descriptor identifiers (DUI) or MIM numbers each on a new line.

Proteins and variants linked to disease

Proteins and variants linked to the disease are searched. It means that all the proteins implicated in the disease are returned even if no variants are known to be associated to the disease.

Variants linked to disease

Only proteins whose variants are known to be associated to the disease are searched.

MeSH

The Medical Subject Headings (MeSH) terminology is a controlled vocabulary thesaurus used for biomedical and health-related documents indexing. It is maintained and used by the National Library of Medicine. (MeSH Home Page).

About two third of the Swiss-Prot entries known to be implicated in a disease have been automatically mapped to the MeSH terminology (Mottaz *et al.*, 2008).

General query

The general query enables the user to retrieve Swiss-Prot entries and variants using Swiss-Prot accession number or identifier, protein name or gene name.

If the searched protein is not a Swiss-Prot human protein containing variant or disease annotation, it will not be found (see *Protein not found*).

General textfield

The user can enter one gene/protein name or several Swiss-Prot accession numbers or identifiers separated by spaces.

General file upload

The file can contain Swiss-Prot accession numbers, identifiers, protein names or gene names, each on a new line.

Variant query

The variant query enables the user to search for variants with specific molecular characteristics. The Swiss-Prot variants are systematically classified into three categories: "polymorphism", "disease" or "unclassified".

- Polymorphism: A variant is classified as "Polymorphism" if no disease-association has been reported;
- Disease: A variant is classified as "Disease" when it is found in patients and disease-association is reported in literature. However, this classification is not a definitive assessment of pathogenicity;
- Unclassified: A variant is "unclassified" if disease-association remains unclear.

Variant textfield

The user can enter one or several variants identifiers such as Swiss-Prot FTID or dbSNP rsID separated by spaces.

Variant filefield

The file can contain one or several variants identifiers such as Swiss-Prot FTID or dbSNP rsID each on a new line

Substitution amino acids

The user can specify for the desired variants the wild-type residue or the mutated residue or both. Polar amino acids include: Arginine, Lysine, Aspartate, Glutamate, Asparagine and Glutamine. Hydrophobic amino acids include: Valine, Isoleucine, Leucine, Methionine, Phenylalanine, Tryptophan and Cysteine.

Blosum Score

The user can specify for the desired variants a threshold for the blosum score. The Blosum score is the score within a Blosum matrix for the corresponding wild-type to variant amino acid change. The log-odds score measures the logarithm for the ratio of the likelihood of two amino acids appearing by chance. The Blosum62 substitution matrix is used. This substitution matrix contains scores for all possible exchanges of one amino acid with another.

Lowest score: -4 (low probability of substitution), highest score: 11 (high probability of substitution)

Information on Blosum matrix

Conservation Score

The user can specify for the desired variants a threshold for the conservation score. The score is a decimal number between 0 and 1. The score was calculated using orthologous sequences from the Orthologs Matrix Project (OMA) project (Schneider *et al.*, 2007). The computation involves several steps:

- Identify to which OMA group the UniProt sequence belongs;
- Perform multiple sequences alignment of all the sequences belonging to the OMA group identified above using MAFFT alignment program (Katoh *et al.*, 2002);
- Compute the diversity of the alignment as well as the conservation score of each residue (or position) of the UniProt sequence using the program (Valdar, 2002).

Protein features in sequence neighborhood

The user can find variants close in the sequence to a feature. He can specify the distance threshold between the mutated residue and the feature, distance that is a number of residue.

3D structure

The user can find variants that have been mapped on an experimental 3 dimensional structure.

3D homology models

The user can find variants for which an available protein homology model(s) exists. The models were constructed using PromodII, the core program of SWISS-MODEL (Guex & Peitsch, 1997).
Protein homology models were constructed only for proteins that have a suitable structural template deposited in the Protein Data Bank (PDB). The sequence identity between the Swiss-Prot protein sequence and the PDB template is at least 70%. In addition, only crystal structures with better than 2.5 A resolution are selected as templates. In cases where there are several suitable templates, an additional selection step will be performed to select only templates that are significantly different from each other, i.e. they display a root mean square deviation (rmsd) of more than 1.5 A.

Surface accessibility

The user can choose to retrieve variants whose wild type residue is surface accessible or buried, by specifying the solvent-accessible surface area (SAS). The SAS is calculated using the MSMS program. We can consider that the variant is surface accessible if the SAS is greater than 0 (Sanner *et al.*, 1996).

Protein-protein interface

The user can choose to retrieve variants whose wild type residue is involved in a protein-protein interface.
We consider that a residue is involved in the interface if one of its atoms is located within a distance r of an atom of a residue present in another protein chain. In the "carbon alpha" method, we only consider the atom carbon alpha of the residue and the distance r is set to 6 Å. In the "Van der Waal" method, all atoms are taken into consideration, and the distance r is set to 4.5 Å.

Protein features in 3D neighborhood

The user can specify for the desired variants a feature that is close to the wild type residue in the 3D structure. The distance radius between the wild type residue and the feature can vary between 3 to 6 angstroms and can be chosen by the user. The mapping of the Swiss-Prot features onto 3D structures was performed using SSMap (David & Yip, 2008). Only variants that have been mapped on an experimentally resolved 3D structure can be retrieved.

Download

The downloadable table contains:

Accession: The Swiss-Prot accession number.

Entry name: The Swiss-Prot entry name.

Disease: The Disease extracted from the Swiss-Prot disease comment line.

MeSH descriptor: MeSH descriptor Unique identifier (descriptorUI).

Feature identifier: The Swiss-Prot sequence feature identifier (ftid), identifying the variants.

Variant: The name of the variant, according to the HGVS recommendations.

rsID: The dbSNP variant identifier.

PDB structure identifier: The PDB structure which contains the variant residue, chosen according to the structural definition of the variant residue environment.

PDB chain: The chain of the PDB structure which contains the variant residue.

PDB position: The position in the PDB chain of the variant residue.

Protein not found

SwissVar gives access to Swiss-Prot human proteins with variants or disease annotation. Different reasons can explain that a protein is not found:

1. The protein does not have any variants or disease annotated in Swiss-Prot.
2. The protein is not a human protein.
3. The protein is in UniProtKB/TrEMBL and not in UniProtKB/Swiss-Prot.

OMIM not found

SwissVar only contains MIM numbers describing phenotypes (# and +).

MeSH descriptor not found

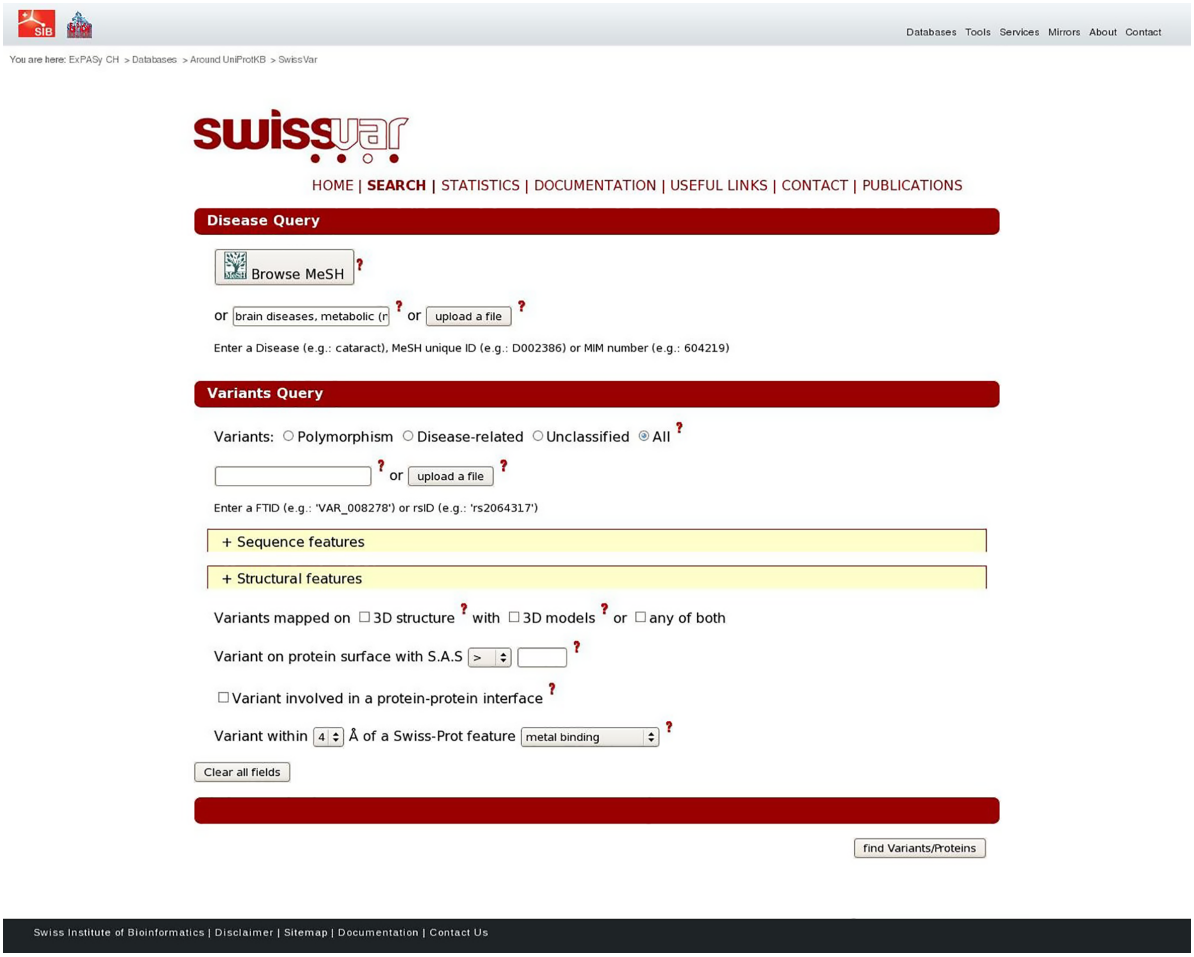
SwissVar only contains MeSH descriptors of the 'Diseases' and 'Psychiatry and Psychology' trees.

Programmatic access

You can directly access the results in the xml or tab delimited format by using the url 'http://swissvar.expasy.org/cgi-bin/swissvar/result' with parameter 'format' having the value xml, tab or html. Without other parameter, all the proteins, diseases and variants will be returned. You can also specify a value to the global_textfield parameter.

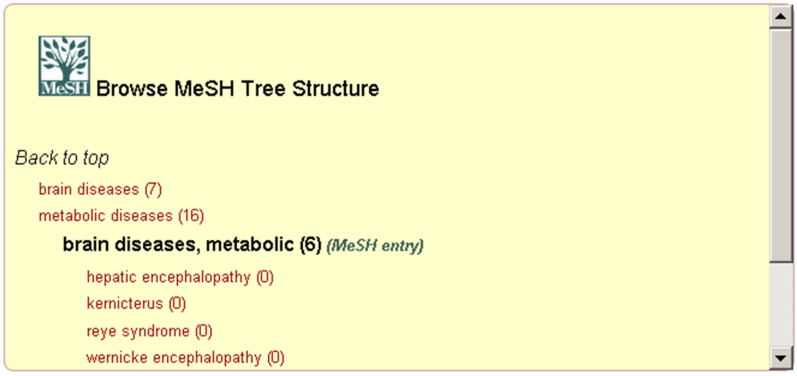
E.g. http://swissvar.expasy.org/cgi-bin/swissvar/result?format=xml&global_textfield=marfan

Supplementary figure 1, Mottaz *et al.*, 2010



Supplementary figure 1. Query combining implication in disease and variant structural feature, searching for variants implicated in any brain metabolic disorder and whose residue is close to a metal binding site in 3D space.

Supplementary figure 2, Mottaz *et al.*, 2010



Refine search | New search

6 Swiss-Prot human proteins found (28 variants)

download

with variants

implicated in brain diseases, metabolic

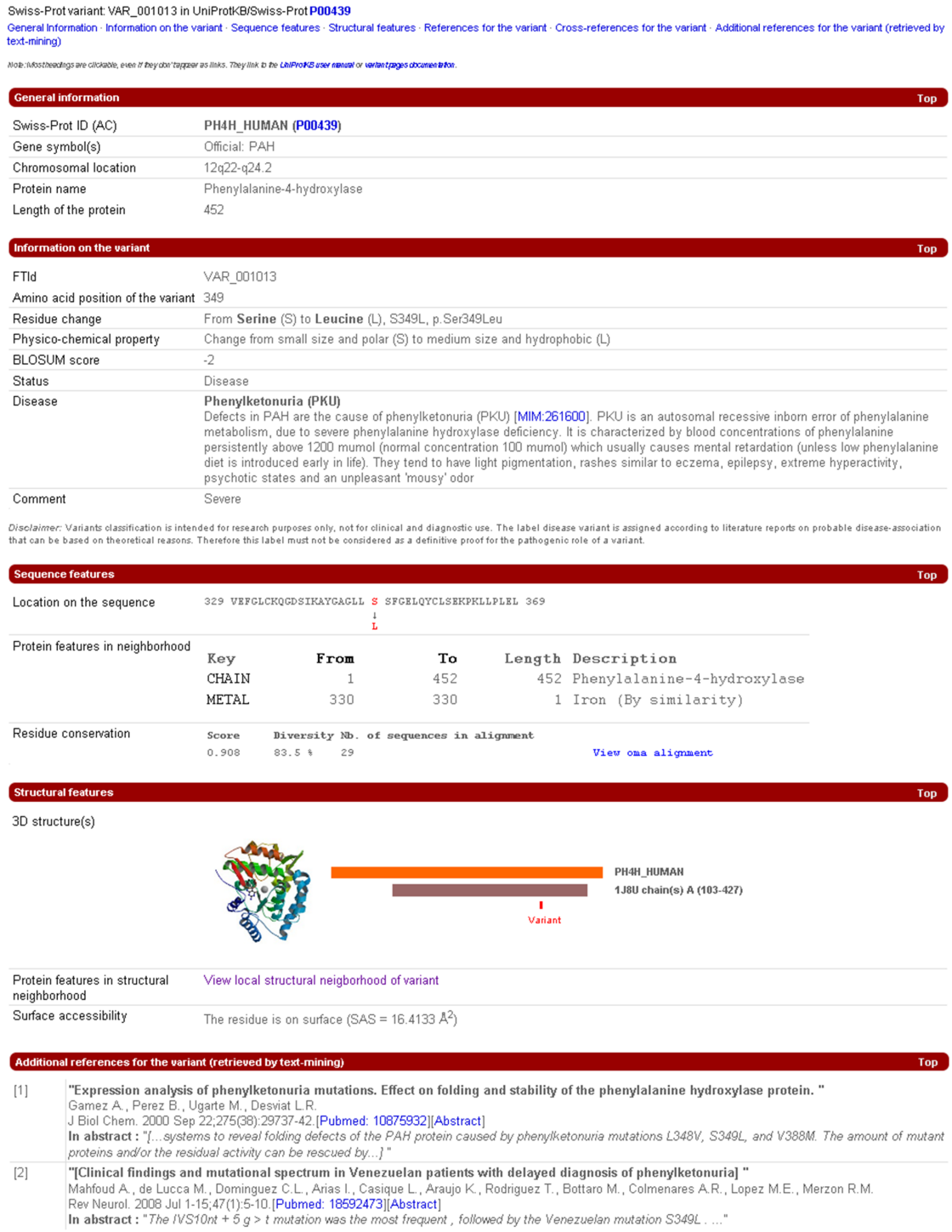
whose wild-type residue is at 4 Å or less of the Swiss-Prot feature metal binding

Disclaimer: The query results are intended for research purposes only, not for clinical and diagnostic use.

Accession	Entry name	Disease	Variants	3D mapping (variant position)
O14832	PAHX_HUMAN	refsum disease	p.Pro173Ser p.His175Arg p.Gln176Lys p.Asp177Gly p.His220Tyr	2A1XA (173) 2A1XA (175) 2A1XA (176) 2A1XA (177) 2A1XA (220)
P00439	PH4H_HUMAN	phenylketonuria	p.Phe263Leu p.Arg270Ser p.Arg270Lys p.Pro281Leu p.Asp282Asn p.Ile283Phe p.Ile283Asn p.Tyr325Cys p.Glu330Asp p.Phe331Leu p.Gly344Arg p.Gly344Val p.Ala345Thr p.Ala345Ser p.Leu347Phe p.Ser349Leu p.Ser349Pro	1KW0A (263) 1MMKA (270) 1MMKA (270) 1DMWA (281) 1MMKA (282) 1MMTA (283) 1MMTA (283) 1LRMA (325) 1KW0A (330) 1LRMA (331) 1DMWA (344) 1DMWA (344) 1MMTA (345) 1MMTA (345) 1DMWA (347) 1J8UA (349) 1J8UA (349)
P05089	ARGI1_HUMAN	argininemia	p.Gly235Arg	2AEBA (235)
P11498	PYC_HUMAN	pyruvate carboxylase deficiency	p.Met743Ile	3BG3D (743)
P15289	ARSA_HUMAN	leukodystrophy metachromatic	p.Gly309Ser	1E1ZP (309)
Q8NBK3	SUMF1_HUMAN	multiple sulfatase deficiency	p.Asn259Ile p.Ala279Val p.Cys336Arg	1Y1IX (259) 1Y1EX (279) 1Y1JX (336)

Supplementary figure 2. Result of the query presented in supplementary figure 1. Six proteins and 28 variants are found. The links give direct access to the original Swiss-Prot entry (column 'Accession'), the MeSH descriptor (column 'Disease'), the Swiss-Prot variants pages (column 'Variants') and the PDB structure with the corresponding position of the residue (column '3D mapping'). Variants related to diseases with a finer or coarser granularity can be searched. Results are downloadable.

Supplementary figure 3, Mottaz *et al.*, 2010



Swiss-Prot variant: VAR_001013 in UniProtKB/Swiss-Prot P00439

General Information · Information on the variant · Sequence features · Structural features · References for the variant · Cross-references for the variant · Additional references for the variant (retrieved by text-mining)

Note: Most headings are clickable, even if they don't appear as links. They link to the UniProtKB user manual or variant pages documentation.

General information Top

Swiss-Prot ID (AC)	PH4H_HUMAN (P00439)
Gene symbol(s)	Official: PAH
Chromosomal location	12q22-q24.2
Protein name	Phenylalanine-4-hydroxylase
Length of the protein	452

Information on the variant Top

FTId	VAR_001013
Amino acid position of the variant	349
Residue change	From **Serine** (S) to **Leucine** (L), S349L, p.Ser349Leu
Physico-chemical property	Change from small size and polar (S) to medium size and hydrophobic (L)
BLOSUM score	-2
Status	Disease
Disease	**Phenylketonuria (PKU)** Defects in PAH are the cause of phenylketonuria (PKU) [MIM:261600]. PKU is an autosomal recessive inborn error of phenylalanine metabolism, due to severe phenylalanine hydroxylase deficiency. It is characterized by blood concentrations of phenylalanine persistently above 1200 mumol (normal concentration 100 mumol) which usually causes mental retardation (unless low phenylalanine diet is introduced early in life). They tend to have light pigmentation, rashes similar to eczema, epilepsy, extreme hyperactivity, psychotic states and an unpleasant 'mousy' odor
Comment	Severe

Disclaimer: Variants classification is intended for research purposes only, not for clinical and diagnostic use. The label disease variant is assigned according to literature reports on probable disease-association that can be based on theoretical reasons. Therefore this label must not be considered as a definitive proof for the pathogenic role of a variant.

Sequence features Top

Location on the sequence

```
329 VEFGLCKQGDSIKAYGAGLL S SFGELQYCLSEKPKLLPLEL 369
                         ↓
                         L
```

Protein features in neighborhood

Key	From	To	Length	Description
CHAIN	1	452	452	Phenylalanine-4-hydroxylase
METAL	330	330	1	Iron (By similarity)

Residue conservation

Score	Diversity	Nb. of sequences in alignment	
0.908	83.5 %	29	View oma alignment

Structural features Top

3D structure(s)

Protein features in structural neighborhood	View local structural neigborhood of variant
Surface accessibility	The residue is on surface (SAS = 16.4133 Å^2)

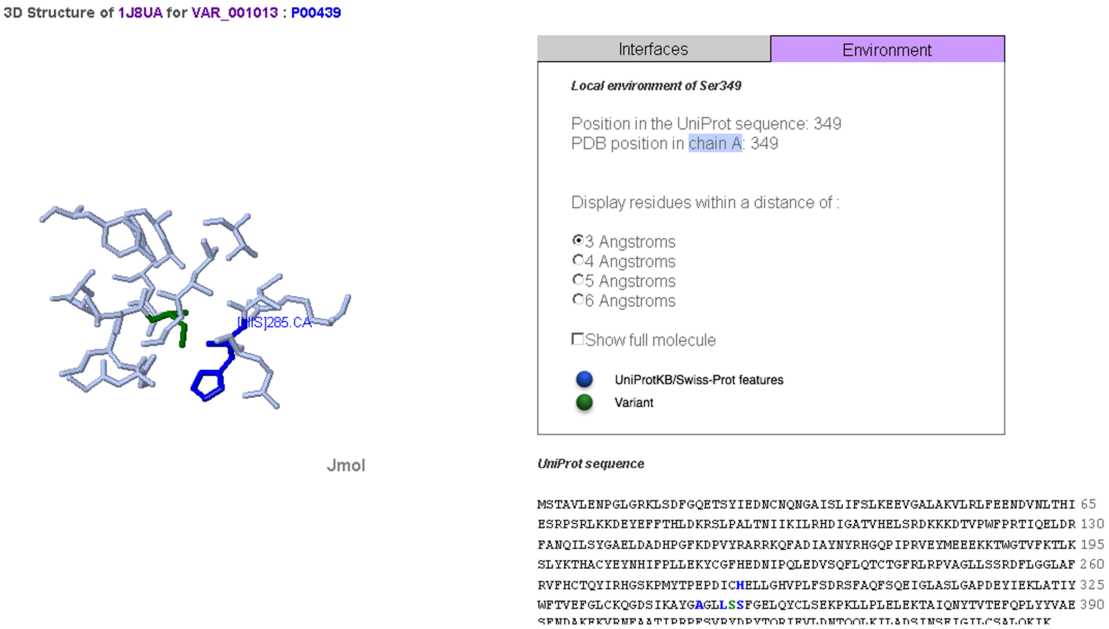
Additional references for the variant (retrieved by text-mining) Top

[1] **"Expression analysis of phenylketonuria mutations. Effect on folding and stability of the phenylalanine hydroxylase protein. "**
Gamez A., Perez B., Ugarte M., Desviat L.R.
J Biol Chem. 2000 Sep 22;275(38):29737-42.[Pubmed: 10875932][Abstract]
In abstract : *"[...systems to reveal folding defects of the PAH protein caused by phenylketonuria mutations L348V, S349L, and V388M. The amount of mutant proteins and/or the residual activity can be rescued by...] "*

[2] **"[Clinical findings and mutational spectrum in Venezuelan patients with delayed diagnosis of phenylketonuria] "**
Mahfoud A., de Lucca M., Dominguez C.L., Arias I., Casique L., Araujo K., Rodriguez T., Bottaro M., Colmenares A.R., Lopez M.E., Merzon R.M.
Rev Neurol. 2008 Jul 1-15;47(1):5-10.[Pubmed: 18592473][Abstract]
In abstract : *"The IVS10nt + 5 g > t mutation was the most frequent , followed by the Venezuelan mutation S349L . ..."*

Supplementary figure 3. Variant page accessed from the result table in supplementary figure 2. Sequence and structural features, general information as well as automatically retrieved references on the variant are presented.

Supplementary figure 4, Mottaz *et al.*, 2010



Key	Sequence range	Description	Structural environment		
VARIANT	350	S -> T (in PKU; haplotype 2)	UniProt sequence position	PDB structure position	Jmol link
			350	Chain A position 350	Show
METAL	285	Iron	UniProt sequence position	PDB structure position	Jmol link
			285	Chain A position 285	Show
VARIANT	348	L -> V (in PKU; mild haplotype 9)	UniProt sequence position	PDB structure position	Jmol link
			348	Chain A position 348	Show
VARIANT	345	A -> T (in PKU; haplotype 7)	UniProt sequence position	PDB structure position	Jmol link
			345	Chain A position 345	Show

Supplementary figure 4. From the variant page, it is possible to precisely visualize the variation and the surrounding features when an experimentally resolved 3D structure of the protein exists.

Table S1

Table S1: Interactions found in the 'Leukemia' subnetwork around the 'Bloom Syndrome Protein'.		
Protein 1	**Protein 2**	**STRING Global Score**
Meiotic recombination protein DMC1/LIM15 homolog	DNA topoisomerase 3-alpha	0.168
DNA mismatch repair protein Msh2*	Meiotic recombination protein DMC1/LIM15 homolog	0.224
DNA repair endonuclease XPF*	Serine/threonine-protein kinase Chk1	0.247
DNA repair protein RAD52 homolog*	DNA repair protein complementing XP-G cells	0.25
DNA topoisomerase 2-beta	Bloom syndrome protein*	0.26
Mismatch repair endonuclease PMS2	DNA repair protein complementing XP-G cells	0.27
Mitotic spindle assembly checkpoint protein MAD2A	DNA repair endonuclease XPF	0.272
DNA repair endonuclease XPF	Meiotic recombination protein DMC1/LIM15 homolog	0.321
CREB-binding protein	Adenomatous polyposis coli protein	0.324
Double-strand break repair protein MRE11A*	DNA repair protein complementing XP-G cells	0.384
Flap endonuclease 1*	Meiotic recombination protein DMC1/LIM15 homolog	0.399
DNA repair endonuclease XPF	DNA repair protein RAD50*	0.408
DNA repair protein RAD52 homolog*	Flap endonuclease 1*	0.421
Mitotic spindle assembly checkpoint protein MAD2A	Bloom syndrome protein*	0.431
Replication protein A 70 kDa DNA-binding subunit*	DNA topoisomerase 2-beta	0.434
Meiotic recombination protein DMC1/LIM15 homolog	Bloom syndrome protein*	0.438
DNA mismatch repair protein Msh2*	Serine-protein kinase ATM*	0.444
Flap endonuclease 1*	Structural maintenance of chromosomes protein 1A	0.446
DNA topoisomerase 1	Bloom syndrome protein*	0.448
TFIIH basal transcription factor complex helicase XPD subunit	DNA topoisomerase 3-alpha	0.45
DNA topoisomerase 2-alpha	Bloom syndrome protein*	0.454
DNA mismatch repair protein Msh6	Replication protein A 32 kDa subunit*	0.456
DNA repair protein RAD52 homolog*	Mismatch repair endonuclease PMS2	0.456
TFIIH basal transcription factor complex helicase XPD subunit	Replication factor C subunit 1	0.466
DNA mismatch repair protein Msh6	DNA topoisomerase 3-alpha	0.467
DNA repair protein RAD50*	Structural maintenance of chromosomes protein 1A	0.47
Flap endonuclease 1*	DNA topoisomerase 1	0.475
Flap endonuclease 1*	Exonuclease 1	0.488
Replication factor C subunit 1	Serine-protein kinase ATM*	0.489
Flap endonuclease 1*	DNA repair protein complementing XP-G cells	0.493
Mismatch repair endonuclease PMS2	DNA repair endonuclease XPF	0.497
DNA repair protein RAD52 homolog*	DNA mismatch repair protein Msh6	0.501
Serine-protein kinase ATM*	DNA mismatch repair protein Msh6	0.507
Mismatch repair endonuclease PMS2	Replication factor C subunit 1	0.53
DNA repair endonuclease XPF	DNA mismatch repair protein Msh6	0.531

TFIIH basal transcription factor complex helicase XPD subunit	DNA repair protein RAD50*	0.532
Chromatin assembly factor 1 subunit A	Telomeric repeat-binding factor 1	0.538
Interferon-induced GTP-binding protein Mx1	Bloom syndrome protein*	0.543
Replication factor C subunit 1	Mast/stem cell growth factor receptor Kit	0.548
CREB-binding protein	DNA mismatch repair protein Msh6	0.573
DNA mismatch repair protein Mlh1*	Flap endonuclease 1*	0.573
DNA mismatch repair protein Msh2*	CREB-binding protein	0.573
CREB-binding protein	Replication protein A 32 kDa subunit*	0.575
DNA repair endonuclease XPF	DNA topoisomerase 3-alpha	0.595
DNA mismatch repair protein Msh6	Replication protein A 70 kDa DNA-binding subunit*	0.598
Flap endonuclease 1*	DNA repair protein RAD51 homolog 1*	0.604
Mismatch repair endonuclease PMS2	Flap endonuclease 1*	0.604
Interferon-induced GTP-binding protein Mx1	Fanconi anemia group A protein	0.609
TFIIH basal transcription factor complex helicase XPD subunit	Flap endonuclease 1*	0.609
Double-strand break repair protein MRE11A*	DNA repair endonuclease XPF	0.614
DNA repair protein RAD52 homolog*	DNA mismatch repair protein Mlh1*	0.615
Chromatin assembly factor 1 subunit A	Bloom syndrome protein*	0.619
Chromatin assembly factor 1 subunit A	Werner syndrome ATP-dependent helicase*	0.619
Chromatin assembly factor 1 subunit A	Replication protein A 70 kDa DNA-binding subunit*	0.619
DNA topoisomerase 2-alpha	Retinoblastoma-associated protein	0.619
Replication factor C subunit 1	Nibrin*	0.619
Retinoblastoma-associated protein	Replication factor C subunit 1	0.631
DNA topoisomerase 2-alpha	DNA topoisomerase 2-beta	0.634
DNA repair endonuclease XPF	Bloom syndrome protein*	0.638
DNA repair protein RAD52 homolog*	Bloom syndrome protein*	0.647
DNA topoisomerase 1	DNA repair protein RAD50*	0.647
DNA repair protein RAD51 homolog 1*	DNA repair protein complementing XP-G cells	0.665
DNA repair protein RAD52 homolog*	Structural maintenance of chromosomes protein 1A	0.673
Flap endonuclease 1*	DNA repair protein RAD50*	0.673
H/ACA ribonucleoprotein complex subunit 4	WD repeat-containing protein 48	0.675
WD repeat-containing protein 48	Bloom syndrome protein*	0.675
WD repeat-containing protein 48	Werner syndrome ATP-dependent helicase*	0.675
Caspase-3	Bloom syndrome protein*	0.681
Cellular tumor antigen p53*	Replication factor C subunit 1	0.681
DNA mismatch repair protein Mlh1*	Caspase-3	0.681
Breast cancer type 1 susceptibility protein*	Tyrosine-protein kinase JAK2	0.682
Breast cancer type 1 susceptibility protein*	Caspase-3	0.682
Double-strand break repair protein MRE11A*	TFIIH basal transcription factor complex helicase XPD subunit	0.683
Caspase-3	DNA repair protein RAD51 homolog 1*	0.684
RecQ-mediated genome instability protein 1	Fanconi anemia group A protein	0.689
Cellular tumor antigen p53*	WD repeat-containing protein 48	0.695
Telomeric repeat-binding factor 2	Breast cancer type 1 susceptibility protein*	0.696
Fanconi anemia group M protein	Replication protein A 32 kDa subunit*	0.701

Fanconi anemia group M protein	Replication protein A 70 kDa DNA-binding subunit*	0.701
Telomeric repeat-binding factor 1	Nibrin*	0.702
DNA mismatch repair protein Msh6	Bloom syndrome protein*	0.703
Telomeric repeat-binding factor 2	DNA topoisomerase 3-alpha	0.712
Bloom syndrome protein*	Serine/threonine-protein kinase Chk1	0.713
DNA mismatch repair protein Msh2*	DNA repair protein complementing XP-G cells	0.719
Retinoblastoma-associated protein	Serine/threonine-protein kinase Chk1	0.723
DNA mismatch repair protein Msh6	Serine/threonine-protein kinase Chk1	0.724
Double-strand break repair protein MRE11A*	DNA topoisomerase 1	0.726
DNA mismatch repair protein Mlh1*	Meiotic recombination protein DMC1/LIM15 homolog	0.738
DNA topoisomerase 2-alpha	Replication protein A 70 kDa DNA-binding subunit*	0.738
Fanconi anemia group M protein	RecQ-mediated genome instability protein 1	0.749
DNA mismatch repair protein Msh2*	DNA repair endonuclease XPF	0.754
Histone H2AX*	Bloom syndrome protein*	0.755
Telomeric repeat-binding factor 1	Breast cancer type 1 susceptibility protein*	0.757
DNA mismatch repair protein Msh2*	Bloom syndrome protein*	0.765
DNA repair protein RAD50*	DNA topoisomerase 3-alpha	0.765
RecQ-mediated genome instability protein 1	Replication protein A 32 kDa subunit*	0.765
Replication factor C subunit 1	Bloom syndrome protein*	0.765
Meiotic recombination protein DMC1/LIM15 homolog	DNA repair protein RAD51 homolog 1*	0.768
DNA repair protein RAD52 homolog*	DNA repair endonuclease XPF	0.774
DNA mismatch repair protein Mlh1*	DNA repair endonuclease XPF	0.776
DNA mismatch repair protein Mlh1*	Structural maintenance of chromosomes protein 1A	0.777
DNA mismatch repair protein Mlh1*	DNA repair protein RAD51 homolog 1*	0.779
DNA repair endonuclease XPF	DNA repair protein RAD51 homolog 1*	0.779
Tumor suppressor p53-binding protein 1*	Bloom syndrome protein*	0.781
DNA repair protein RAD52 homolog*	DNA topoisomerase 3-alpha	0.783
Exonuclease 1	DNA repair protein RAD51 homolog 1*	0.785
DNA repair endonuclease XPF	Replication protein A 70 kDa DNA-binding subunit*	0.788
Werner syndrome ATP-dependent helicase*	Bloom syndrome protein*	0.798
DNA mismatch repair protein Mlh1*	Serine-protein kinase ATM*	0.8
Exonuclease 1	Bloom syndrome protein*	0.81
Chromatin assembly factor 1 subunit A	Serine-protein kinase ATM*	0.812
DNA mismatch repair protein Msh2*	DNA repair protein RAD51 homolog 1*	0.815
Werner syndrome ATP-dependent helicase*	Cyclin-dependent kinase inhibitor 2A, isoforms ½/3	0.816
Double-strand break repair protein MRE11A*	DNA topoisomerase 3-alpha	0.818
DNA mismatch repair protein Msh2*	Nibrin*	0.821
DNA repair protein RAD52 homolog*	DNA topoisomerase 1	0.826
Flap endonuclease 1*	Replication protein A 32 kDa subunit*	0.83
Meiotic recombination protein DMC1/LIM15 homolog	DNA repair protein RAD50*	0.83
DNA mismatch repair protein Msh2*	Replication protein A 70 kDa DNA-binding subunit*	0.832
DNA repair protein RAD51 homolog 1*	DNA topoisomerase 3-alpha	0.832

Replication protein A 70 kDa DNA-binding subunit*	DNA repair protein RAD50*	0.833
Serine-protein kinase ATM*	DNA repair protein RAD51 homolog 1*	0.833
Serine-protein kinase ATM*	Bloom syndrome protein*	0.834
Double-strand break repair protein MRE11A*	Replication factor C subunit 1	0.835
Replication protein A 70 kDa DNA-binding subunit*	Tumor suppressor p53-binding protein 1*	0.838
Double-strand break repair protein MRE11A*	Flap endonuclease 1*	0.845
Retinoblastoma-associated protein	Caspase-3	0.846
Telomeric repeat-binding factor 1	Bloom syndrome protein*	0.849
DNA topoisomerase 1	DNA topoisomerase 3-alpha	0.851
TFIIH basal transcription factor complex helicase XPD subunit	Bloom syndrome protein*	0.857
DNA mismatch repair protein Msh2*	DNA repair protein RAD52 homolog*	0.863
Replication factor C subunit 1	Flap endonuclease 1*	0.865
DNA mismatch repair protein Msh6	Structural maintenance of chromosomes protein 1A	0.867
Telomeric repeat-binding factor 1	Serine-protein kinase ATM*	0.871
DNA repair protein RAD52 homolog*	TFIIH basal transcription factor complex helicase XPD subunit	0.875
Double-strand break repair protein MRE11A*	Meiotic recombination protein DMC1/LIM15 homolog	0.875
Tumor suppressor p53-binding protein 1*	Replication protein A 32 kDa subunit*	0.876
DNA repair protein RAD50*	Bloom syndrome protein*	0.879
DNA mismatch repair protein Msh2*	TFIIH basal transcription factor complex helicase XPD subunit	0.88
Meiotic recombination protein DMC1/LIM15 homolog	Replication protein A 70 kDa DNA-binding subunit*	0.881
Fanconi anemia group A protein	Replication protein A 70 kDa DNA-binding subunit*	0.882
DNA repair protein RAD52 homolog*	Meiotic recombination protein DMC1/LIM15 homolog	0.884
DNA repair endonuclease XPF	Flap endonuclease 1*	0.887
Nibrin*	Fanconi anemia group D2 protein	0.891
Cellular tumor antigen p53*	DNA topoisomerase 2-beta	0.894
Serine-protein kinase ATM*	Werner syndrome ATP-dependent helicase*	0.9
DNA mismatch repair protein Msh6	Nibrin*	0.901
Bloom syndrome protein*	DNA repair protein complementing XP-G cells	0.904
Breast cancer type 1 susceptibility protein*	Structural maintenance of chromosomes protein 1A	0.906
RecQ-mediated genome instability protein 1	Replication protein A 70 kDa DNA-binding subunit*	0.906
DNA repair protein RAD51 homolog 1*	DNA repair protein RAD50*	0.907
Bloom syndrome protein*	Structural maintenance of chromosomes protein 1A	0.909
Histone H2AX*	Werner syndrome ATP-dependent helicase*	0.911
Histone H2AX*	DNA repair protein RAD51 homolog 1*	0.911
DNA mismatch repair protein Msh2*	Structural maintenance of chromosomes protein 1A	0.912
Mitotic spindle assembly checkpoint protein MAD2A	Structural maintenance of chromosomes protein 1A	0.912
Exonuclease 1	DNA repair protein RAD50*	0.915
Flap endonuclease 1*	Replication protein A 70 kDa DNA-binding subunit*	0.915
Nibrin*	Bloom syndrome protein*	0.917

DNA mismatch repair protein Mlh1*	Nibrin*	0.918
DNA topoisomerase 2-alpha	Breast cancer type 1 susceptibility protein*	0.919
DNA topoisomerase 1	Cyclin-dependent kinase inhibitor 2A, isoforms ½/3	0.925
Fanconi anemia group M protein	Bloom syndrome protein*	0.925
Replication factor C subunit 1	Breast cancer type 1 susceptibility protein*	0.925
Breast cancer type 1 susceptibility protein*	WD repeat-containing protein 48	0.927
Replication factor C subunit 1	DNA mismatch repair protein Msh6	0.929
Mismatch repair endonuclease PMS2	Exonuclease 1	0.93
DNA mismatch repair protein Msh2*	Flap endonuclease 1*	0.931
Replication factor C subunit 1	Structural maintenance of chromosomes protein 1A	0.932
Replication factor C subunit 1	DNA repair protein RAD50*	0.933
DNA mismatch repair protein Mlh1*	Replication factor C subunit 1	0.935
Double-strand break repair protein MRE11A*	DNA repair protein RAD51 homolog 1*	0.935
Fanconi anemia group A protein	DNA topoisomerase 3-alpha	0.937
DNA repair protein RAD51 homolog 1*	Werner syndrome ATP-dependent helicase*	0.939
Histone H2AX*	DNA repair protein RAD50*	0.94
Double-strand break repair protein MRE11A*	Exonuclease 1	0.941
TFIIH basal transcription factor complex helicase XPD subunit	DNA repair protein RAD51 homolog 1*	0.941
Double-strand break repair protein MRE11A*	DNA repair protein RAD52 homolog*	0.944
Bloom syndrome protein*	Replication protein A 32 kDa subunit*	0.946
DNA mismatch repair protein Msh2*	Replication factor C subunit 1	0.946
DNA mismatch repair protein Msh2*	Double-strand break repair protein MRE11A*	0.947
DNA repair protein RAD51 homolog 1*	Fanconi anemia group D2 protein	0.948
Double-strand break repair protein MRE11A*	Bloom syndrome protein*	0.949
DNA repair protein RAD52 homolog*	DNA repair protein RAD50*	0.953
Double-strand break repair protein MRE11A*	DNA mismatch repair protein Msh6	0.954
Breast cancer type 1 susceptibility protein*	Bloom syndrome protein*	0.955
Fanconi anemia group A protein	Bloom syndrome protein*	0.956
Telomeric repeat-binding factor 2	Nibrin*	0.959
Breast cancer type 1 susceptibility protein*	Serine/threonine-protein kinase Chk1	0.962
Caspase-3	DNA topoisomerase 1	0.965
Double-strand break repair protein MRE11A*	Histone H2AX*	0.966
Replication factor C subunit 1	Caspase-3	0.966
DNA mismatch repair protein Msh6	DNA repair protein RAD50*	0.967
DNA repair endonuclease XPF	Fanconi anemia group A protein	0.967
DNA topoisomerase 1	DNA topoisomerase 2-beta	0.967
DNA mismatch repair protein Mlh1*	DNA repair protein RAD50*	0.968
Mitotic spindle assembly checkpoint protein MAD2A	Adenomatous polyposis coli protein	0.968
DNA mismatch repair protein Msh2*	Serine/threonine-protein kinase Chk1	0.972
Mismatch repair endonuclease PMS2	DNA mismatch repair protein Msh6	0.974
DNA mismatch repair protein Msh2*	DNA repair protein RAD50*	0.975
Double-strand break repair protein MRE11A*	Fanconi anemia group D2 protein	0.975
Cellular tumor antigen p53*	DNA topoisomerase 2-alpha	0.977
Bloom syndrome protein*	Fanconi anemia group D2 protein	0.98
DNA repair protein RAD52 homolog*	Werner syndrome ATP-dependent helicase*	0.981

Double-strand break repair protein MRE11A*	DNA mismatch repair protein Mlh1*	0.981
Flap endonuclease 1*	Bloom syndrome protein*	0.981
DNA mismatch repair protein Mlh1*	Breast cancer type 1 susceptibility protein*	0.982
DNA repair protein RAD51 homolog 1*	Bloom syndrome protein*	0.982
DNA repair protein RAD51 homolog 1*	Serine/threonine-protein kinase Chk1	0.983
Double-strand break repair protein MRE11A*	Replication protein A 70 kDa DNA-binding subunit*	0.983
Tumor suppressor p53-binding protein 1*	Serine/threonine-protein kinase Chk1	0.984
Mitotic checkpoint serine/threonine-protein kinase BUB1 beta	Adenomatous polyposis coli protein	0.987
Serine-protein kinase ATM*	Fanconi anemia group D2 protein	0.987
Cellular tumor antigen p53*	Histone H2AX*	0.988
Breast cancer type 1 susceptibility protein*	Nibrin*	0.989
Telomeric repeat-binding factor 2	Serine-protein kinase ATM*	0.989
Tyrosine-protein kinase JAK2	Mast/stem cell growth factor receptor Kit	0.989
DNA repair protein RAD51 homolog 1*	Replication protein A 70 kDa DNA-binding subunit*	0.99
DNA topoisomerase 2-alpha	DNA topoisomerase 1	0.99
DNA repair protein RAD52 homolog*	Replication protein A 32 kDa subunit*	0.991
CREB-binding protein	Breast cancer type 1 susceptibility protein*	0.992
Retinoblastoma-associated protein	Breast cancer type 1 susceptibility protein*	0.992
Cellular tumor antigen p53*	TFIIH basal transcription factor complex helicase XPD subunit	0.993
Fanconi anemia group M protein	Fanconi anemia group E protein	0.993
Replication protein A 70 kDa DNA-binding subunit*	Werner syndrome ATP-dependent helicase*	0.993
Serine-protein kinase ATM*	Structural maintenance of chromosomes protein 1A	0.993
Telomeric repeat-binding factor 2	Werner syndrome ATP-dependent helicase*	0.993
Breast cancer type 1 susceptibility protein*	DNA mismatch repair protein Msh6	0.994
Fanconi anemia group M protein	Fanconi anemia group C protein	0.994
DNA mismatch repair protein Mlh1*	Bloom syndrome protein*	0.995
DNA mismatch repair protein Msh2*	Breast cancer type 1 susceptibility protein*	0.995
Telomeric repeat-binding factor 2	Bloom syndrome protein*	0.995
Breast cancer type 1 susceptibility protein*	Tumor suppressor p53-binding protein 1*	0.996
DNA repair endonuclease XPF	DNA repair protein complementing XP-G cells	0.996
Histone H2AX*	Nibrin*	0.996
Telomeric repeat-binding factor 2	5′ exonuclease Apollo	0.996
Breast cancer type 1 susceptibility protein*	Fanconi anemia group A protein	0.997
Cellular tumor antigen p53*	DNA topoisomerase 1	0.997
DNA repair protein RAD52 homolog*	Replication protein A 70 kDa DNA-binding subunit*	0.997
Double-strand break repair protein MRE11A*	Telomeric repeat-binding factor 2	0.997
Serine-protein kinase ATM*	DNA repair protein RAD50*	0.997
Serine-protein kinase ATM*	Tumor suppressor p53-binding protein 1*	0.997
Cellular tumor antigen p53*	Werner syndrome ATP-dependent helicase*	0.998
Cellular tumor antigen p53*	DNA repair protein RAD51 homolog 1*	0.998
DNA mismatch repair protein Mlh1*	DNA mismatch repair protein Msh6	0.998
DNA mismatch repair protein Msh2*	Cellular tumor antigen p53*	0.998
DNA mismatch repair protein Msh2*	Mismatch repair endonuclease PMS2	0.998

Mismatch repair endonuclease PMS2	DNA mismatch repair protein Mlh1*	0.998
Replication protein A 70 kDa DNA-binding subunit*	Bloom syndrome protein*	0.998
Telomeric repeat-binding factor 2	DNA repair protein RAD50*	0.998
TFIIH basal transcription factor complex helicase XPD subunit	DNA repair protein complementing XP-G cells	0.998
Bloom syndrome protein*	DNA topoisomerase 3-alpha	0.999
Breast cancer type 1 susceptibility protein*	Fanconi anemia group D2 protein	0.999
Breast cancer type 1 susceptibility protein*	Serine-protein kinase ATM*	0.999
Breast cancer type 1 susceptibility protein*	DNA repair protein RAD50*	0.999
Breast cancer type 1 susceptibility protein*	DNA repair protein RAD51 homolog 1*	0.999
Cellular tumor antigen p53*	Serine/threonine-protein kinase Chk1	0.999
Cellular tumor antigen p53*	Bloom syndrome protein*	0.999
Cellular tumor antigen p53*	Cyclin-dependent kinase inhibitor 2A, isoforms ½/3	0.999
Cellular tumor antigen p53*	CREB-binding protein	0.999
Cellular tumor antigen p53*	Serine-protein kinase ATM*	0.999
Cellular tumor antigen p53*	Replication protein A 70 kDa DNA-binding subunit*	0.999
Cellular tumor antigen p53*	Breast cancer type 1 susceptibility protein*	0.999
Cellular tumor antigen p53*	Tumor suppressor p53-binding protein 1*	0.999
DNA mismatch repair protein Mlh1*	Exonuclease 1	0.999
DNA mismatch repair protein Msh2*	DNA mismatch repair protein Msh6	0.999
DNA mismatch repair protein Msh2*	DNA mismatch repair protein Mlh1*	0.999
DNA mismatch repair protein Msh2*	Exonuclease 1	0.999
DNA repair protein RAD50*	Nibrin*	0.999
DNA repair protein RAD52 homolog*	DNA repair protein RAD51 homolog 1*	0.999
Double-strand break repair protein MRE11A*	Serine-protein kinase ATM*	0.999
Double-strand break repair protein MRE11A*	DNA repair protein RAD50*	0.999
Double-strand break repair protein MRE11A*	Nibrin*	0.999
Double-strand break repair protein MRE11A*	Breast cancer type 1 susceptibility protein*	0.999
Fanconi anemia group A protein	Fanconi anemia group C protein	0.999
Fanconi anemia group A protein	Fanconi anemia group E protein	0.999
Fanconi anemia group C protein	Fanconi anemia group D2 protein	0.999
Fanconi anemia group C protein	Fanconi anemia group E protein	0.999
Fanconi anemia group E protein	Fanconi anemia group D2 protein	0.999
Fanconi anemia group M protein	Fanconi anemia group A protein	0.999
Flap endonuclease 1*	Werner syndrome ATP-dependent helicase*	0.999
Histone H2AX*	Serine-protein kinase ATM*	0.999
Histone H2AX*	Breast cancer type 1 susceptibility protein*	0.999
Histone H2AX*	Tumor suppressor p53-binding protein 1*	0.999
Mitotic checkpoint serine/threonine-protein kinase BUB1 beta	Mitotic spindle assembly checkpoint protein MAD2A	0.999
RecQ-mediated genome instability protein 1	Bloom syndrome protein*	0.999
RecQ-mediated genome instability protein 1	DNA topoisomerase 3-alpha	0.999
Replication protein A 70 kDa DNA-binding subunit*	Replication protein A 32 kDa subunit*	0.999
Serine-protein kinase ATM*	Serine/threonine-protein kinase Chk1	0.999
Serine-protein kinase ATM*	Nibrin*	0.999

Legend:
Brown: Implicated in leukemia (according to HPO).
Orange: Implicated in any disease (according to UniProtKB/Swiss-Prot disease annotation).
**: Implicated in DSB repair (according to UniProtKB/Swiss-Prot GO annotations).*