Tools and annotations for variation

Schaafsma, Gerard C.P.

2017

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

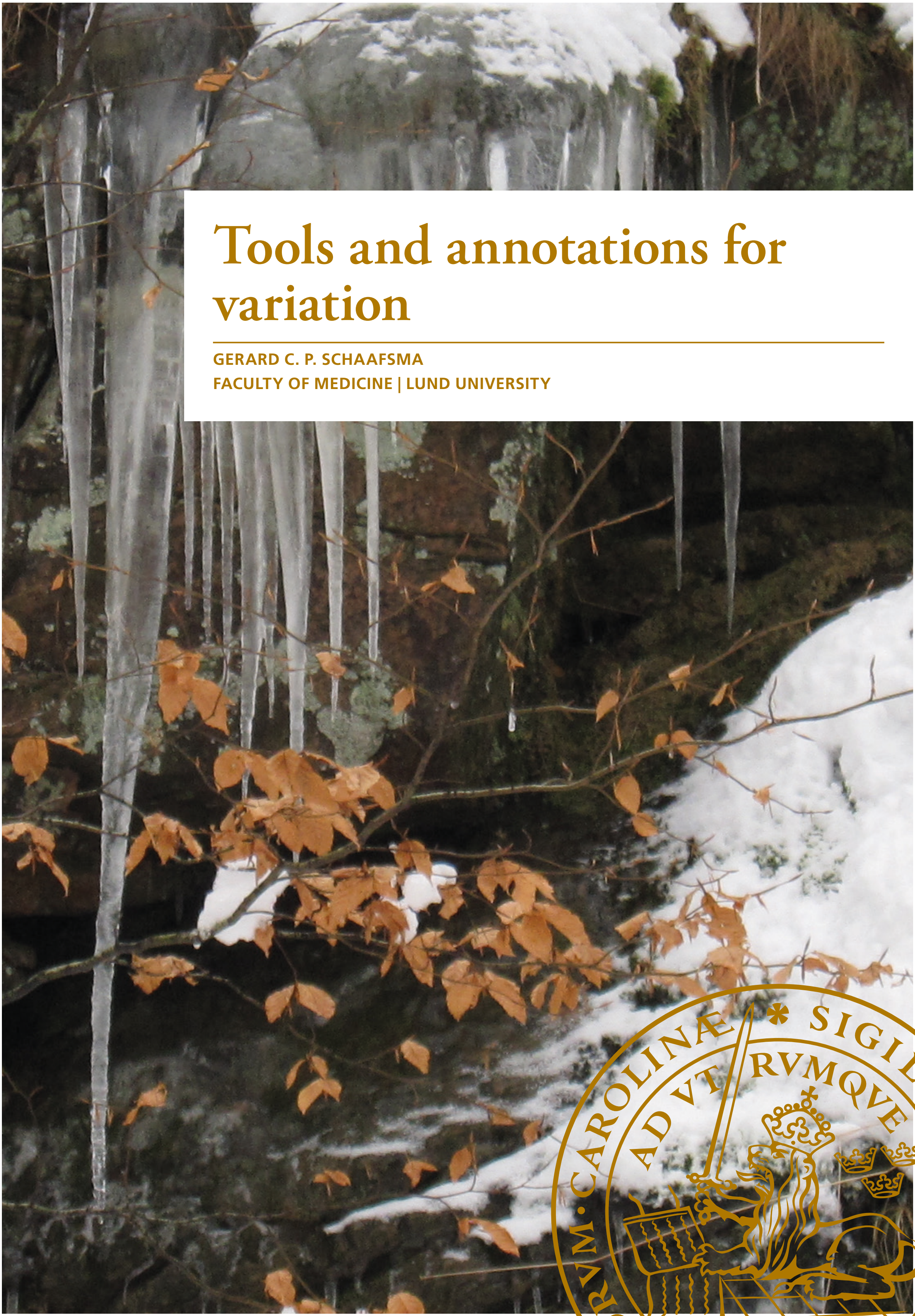
Citation for published version (APA):
Schaafsma, G. C. P. (2017). *Tools and annotations for variation*. [Doctoral Thesis (compilation), Department of Experimental Medical Science]. Lund University: Faculty of Medicine.

Total number of authors:
1

Tools and annotations for variation

Schaafsma, Gerard C.P.

Published: 2017-01-01



Tools and annotations for variation

GERARD C. P. SCHAAFSMA

FACULTY OF MEDICINE | LUND UNIVERSITY

Tools and annotations for variation

Gerard C. P. Schaafsma

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Medicine, Lund University, Sweden.

To be defended in room Dora Jacobsohn, BMC, Lund, on 29 September 2017 at 13:00h.

Faculty opponent

Prof. Dr. Andreas Keller

Chair for Clinical Bioinformatics,

Saarland University, Saarbrücken, Germany

Organization: LUND UNIVERSITY	Document name: Doctoral dissertation
	Date of issue: 2017-08-24
Author: Gerard C. P. Schaafsma	Sponsoring organization
Title: Tools and annotations for variation	

Abstract: Since the finishing of the Human Genome Project, many next-generation (NGS) or high-throughput sequencing platforms have emerged. One of the applications of NGS technology, variant discovery, can serve as a basis for precision medicine. Large sequencing projects are generating huge amounts of genetic variation data, which are stored in databases, either large central databases such as dbSNP, or gene- or disease-centered locus-specific databases (LSDBs). There are many variation databases with many different formats and varying quality. Apart from storage and analysis pipeline capacity problems, the interpretation of the variation is also an issue. Computational methods for predicting the effects of variants have been and are being developed, since experimental assessment of variation effects is often not feasible. Benchmark datasets are needed for the development and for performance assessment of such prediction methods.

We studied quality related aspects of variant databases and benchmark datasets. The online tool called VariOtator was developed to aid in the consistent use of the Variation Ontology, which was specifically developed to describe variation. Standardization is one aspect of database quality; the use of an ontology for variant annotation will contribute to the enhancement of it.

BTKbase is a locus-specific database containing information on variants in *BTK*, the gene involved in X-linked agammaglobulinemia (XLA), a primary immunodeficiency. If available, phenotypic data, i.e. the variant effects, are also provided. Statistics on variants and variation types showed that there is a wide spectrum of variants and variation types, and that the distribution of protein variants in the different BTK domains is not even.

The VariSNP database containing datasets with neutral (non-pathogenic) variants was generated by selecting variants from dbSNP and filtering for variants found in the ClinVar, PhenCode and SwissProt databases. Variants in these three databases are considered to be disease-related. The VariSNP database contains 13 datasets following the functional classification of dbSNP, and is updated on a regular basis.

To study the sensitivity to variation in different protein and disease groups, we predicted the pathogenicity of all possible single amino acid substitutions (SAASs) in all proteins in these groups, using the well-performing prediction method PON P2. Large differences in the proportions of harmful, benign and unknown variants were found, and distinctive patterns of SAAS types were found, both in the original and variant amino acids.

Representativeness is one quality aspect of variation benchmark datasets, and relates to the representation of the space of variants and their effects. We studied the coverage and distribution of protein features, including structure (CATH) and enzyme classification (EC), Pfam domains and Gene Ontology terms, in established benchmark datasets. None of the datasets is fully representative. Coverage of the features is in general better in the larger datasets, and better in the neutral datasets. At the higher levels of the CATH and EC classifications, all datasets were unbiased, but for the lower levels and other features, all datasets were biased.

Key words Annotation, genetic variation, benchmarks, databases, disease groups, pathogenicity, proteins, representativeness, sensitivity, variant effect analysis		
Classification system and/or index terms (if any)		
Supplementary bibliographical information Lund University, Faculty of Medicine Doctoral Dissertation Series 2017:132		Language English
ISSN 1652-8220		ISBN 978-91-7619-515-4
Recipient’s notes	Number of pages 58	Price
	Security classification	

Signature G Schaafsma Date 2017-08-24

Tools and annotations for variation

Gerard C. P. Schaafsma

LUND
UNIVERSITY

Cover: Vinter i Söderåsen

Department of Experimental Medical Science

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2017:132

ISBN 978-91-7619-515-4
ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University
Lund 2017

ISO 14001:2004 CERTIFICATION™
Intertek

MILJÖMÄRKT
341903

fti

Content

Papers included in this thesis

Paper I

Gerard C. P. Schaafsma and Mauno Vihinen

VariOtator, a software tool for variation annotation with the Variation Ontology

Human Mutation (2016) 37 (4): 344-349

Paper II

Gerard C. P. Schaafsma and Mauno Vihinen

Genetic variation in Bruton tyrosine kinase

In: Springer International Publishing Switzerland, 2015. A. Plebani, V. Lougaris (eds.), Agammaglobulinemia, Rare Diseases of the Immune System 4, p. 75-85

Paper III

Gerard C. P. Schaafsma and Mauno Vihinen

VariSNP, a benchmark database for variations from dbSNP

Human Mutation (2015) 36 (2): 161-166

Paper IV

Gerard C. P. Schaafsma and Mauno Vihinen

Large differences in proportions of harmful and benign amino acid substitutions in proteins and diseases

Human Mutation (2017) 38 (7): 839-848

Paper V

Gerard C. P. Schaafsma and Mauno Vihinen

Representativeness of variation benchmark datasets

Manuscript

Abstract

Since the finishing of the Human Genome Project, many next-generation (NGS) or high-throughput sequencing platforms have emerged. One of the applications of NGS technology, variant discovery, can serve as a basis for precision medicine. Large sequencing projects are generating huge amounts of genetic variation data, which are stored in databases, either large central databases such as dbSNP, or gene- or disease-centered locus-specific databases (LSDBs). There are many variation databases with many different formats and varying quality. Apart from storage and analysis pipeline capacity problems, the interpretation of the variation is also an issue. Computational methods for predicting the effects of variants have been and are being developed, since experimental assessment of variation effects is often not feasible. Benchmark datasets are needed for the development and for performance assessment of such prediction methods.

We studied quality related aspects of variant databases and benchmark datasets. The online tool called VariOtator was developed to aid in the consistent use of the Variation Ontology, which was specifically developed to describe variation. Standardization is one aspect of database quality; the use of an ontology for variant annotation will contribute to the enhancement of it.

BTKbase is a locus-specific database containing information on variants in *BTK*, the gene involved in X-linked agammaglobulinemia (XLA), a primary immunodeficiency. If available, phenotypic data, i.e. the variant effects, are also provided. Statistics on variants and variation types showed that there is a wide spectrum of variants and variation types, and that the distribution of protein variants in the different BTK domains is not even.

The VariSNP database containing datasets with neutral (non-pathogenic) variants was generated by selecting variants from dbSNP and filtering for variants found in the ClinVar, PhenCode and SwissProt databases. Variants in these three databases are considered to be disease-related. The VariSNP database contains 13 datasets following the functional classification of dbSNP, and is updated on a regular basis.

To study the sensitivity to variation in different protein and disease groups, we predicted the pathogenicity of all possible single amino acid substitutions (SAASs) in all proteins in these groups, using the well-performing prediction method PON-P2. Large differences in the proportions of harmful, benign and unknown

variants were found, and distinctive patterns of SAAS types were found, both in the original and variant amino acids.

Representativeness is one quality aspect of variation benchmark datasets, and relates to the representation of the space of variants and their effects. We studied the coverage and distribution of protein features, including structure (CATH) and enzyme (EC) classification, Pfam domains and Gene Ontology terms, in established benchmark datasets. None of the datasets is fully representative. Coverage of the features is in general better in the larger datasets, and better in the neutral datasets. At the higher levels of the CATH and EC classifications, all datasets were unbiased, but for the lower levels and other features, all datasets were biased.

List of abbreviations

ACMG	American College of Medical Genetics and Genomics
AAS	Amino Acid Substitution
BTK	Bruton Tyrosine Kinase
CATH	Class Architecture Topology Homology
dbSNP	Database of Single Nucleotide Polymorphisms
EBI	European Bioinformatics Institute
EC	Enzyme Commission
GO	Gene Ontology
GUI	Graphical User Interface
HGNC	HUGO Gene Nomenclature Committee
HGP	Human Genome Project
HGVS	Human Genome Variation Society
HVP	Human Variome Project
HUGO	Human Genome Organisation
IUIS	International Union of Immunological Societies
LOVD	Leiden Open Variation Database
LRG	Locus Reference Genomic
LSDB	Locus Specific Database
ML	Machine Learning
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
OMIM	Online Mendelian Inheritance in Man
OWL	Web Ontology Language

PDB	Protein Data Bank
PH	Pleckstrin Homology
PID	Primary Immunodeficiency
RCSB	Research Collaboratory for Structural Bioinformatics
RNA-seq	RNA sequencing
SAAS	Single Amino Acid Substitution
SH2	Src Homology 2
SH3	Src Homology 3
SNP	Single Nucleotide Polymorphism
SO	Sequence Ontology
SVD-WG	Sequence Variant Description Working Group
TH	Tec Homology
TK	Tyrosine Kinase
UCSC	University of California, Santa Cruz
VariO	Variation Ontology
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
XLA	X-linked agammaglobulinemia

Introduction

Background

The Human Genome Project (HGP) was initiated in 1990, with the aim of determining the complete sequence of DNA bases in the human genome, and to disclose all human genes and make them accessible for further study (Lander, et al., 2001). The HGP officially ended in 2003 with a near-complete sequence containing ~99.7% of the euchromatic genome, only interrupted by ~300 gaps and an error rate of one nucleotide per 100 000 bases (Consortium, 2004; Lander, 2011). Sanger sequencing was essentially the sequencing method used in the HGP (Sanger and Coulson, 1975; Sanger, et al., 1977). Next-generation sequencing (NGS) or high-throughput sequencing technologies have since emerged and evolved, increasing the capacity and decreasing the cost of human genome sequencing (Goodwin, et al., 2016). NGS platforms are now a routine part of biological research and are becoming more widespread in the clinical sector (Goodwin, et al., 2016). The NGS platforms can be classified into two groups, short- and long-read sequencing, both with their advantages and drawbacks, depending on the objectives. Applications of sequencing include whole-genome sequencing (WGS), whole-exome sequencing (WES), RNA sequencing (RNA-seq) and targeted sequencing, among many others. Through WGS, which is becoming one of the most widely used NGS applications, it is possible to obtain the most comprehensive view of genomic information and associated biological implications (Goodwin, et al., 2016).

All these NGS platforms generate huge amounts of data, which is challenging for both analysis and infrastructure. The world capacity was in 2013 estimated at ~15 petabytes of sequencing data per year, and at a rate increasing three- to five-fold per year (Schatz and Langmead, 2013). Projected needs for data acquisition and data storage are estimated at 1 zetta-bases/year and 2-40 exa-bytes/year for the year 2025 (Stephens, et al., 2015).

One of the applications of NGS technology, variant discovery, is becoming common in medical genetics and can serve as a basis for personalized medicine (Nekrutenko and Taylor, 2012). Although often used interchangeably, the term precision medicine is now preferred, since it covers the current approach better (Carrasco-Ramiro, et al., 2017).

Sequencing errors, which are thought to be mostly caused by polymerase chain reaction mistakes or sequencing miscalls, are a challenge. Preprocessing of NGS data to improve quality can be done e.g. by simply trimming based on quality scores, but there are also more advanced error-correction methods (Yang, et al., 2013). Next to sequencing errors and variant calling errors, interpretation of these DNA sequence changes regarding their functional consequences can also be problematic. Although many variants have been associated with rare and common genetic disorders correctly, false assignments exist at a substantial level (MacArthur, et al., 2014). Incorrect assignment of pathogenicity can have serious consequences, both for medicine and research. A recent study (Chen, et al., 2017) indicated that mutagenic DNA damage occurring during sequencing is a cause of sequencing errors, previously thought to occur only in specialized samples, causing erroneous variant identification. Guidelines for the evaluation of causality of variants have been published (MacArthur, et al., 2014).

Variation data are being submitted to variation databases, such as gene variant or locus-specific variation databases (LSDBs) or large depositories such as the Database for Short Genetic Variations (dbSNP) (Sherry, et al., 2001), UniProt (UniProt Consortium, 2017) and Ensembl (Aken, et al., 2017). Making variation data publicly available is a prerequisite for further analysis, such as variation interpretation. To illustrate the growth of variation data, dbSNP, the largest variation database, is taken as an example. dbSNP contains both benign and disease-causing cases. Despite of the abbreviation for single nucleotide polymorphism (SNP) in dbSNP, it is a public archive for all short sequence variation, and includes of broad collection of simple genetic variations such as single-base nucleotide substitutions, small-scale multi-base deletions or insertions, and microsatellite repeats (National Center for Biotechnology Information, 2014). Human Build 150, released in April 2017, contains more than 336 million records on Homo sapiens. Inclusion of data from large sequencing projects such as the 1000 Genomes Project (Genomes Project, et al., 2015), and recently from Human Longevity Inc. (Telenti, et al., 2016) and the Trans-Omics for Precision Medicine (TOPMed; https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed) programme contributed substantially to the growth of this database, it more than doubled its contents, from 154 to 324 million reference SNPs in April 2017. The growth of human variation data in dbSNP in the last 10 years is illustrated in Fig. 1. Ongoing large-scale projects such as the 100,000 Genomes Project (https://www.genomicsengland.co.uk/the-100000-genomes-project/) (Peplow, 2016) and the All of Us Research Program (formerly Precision Medicine Initiative Cohort Program, https://allofus.nih.gov) will likely cause an even faster growth of the amounts of variation data in the future; the large amounts of data are already straining data storage capacities and analysis pipelines (McPherson, 2014).

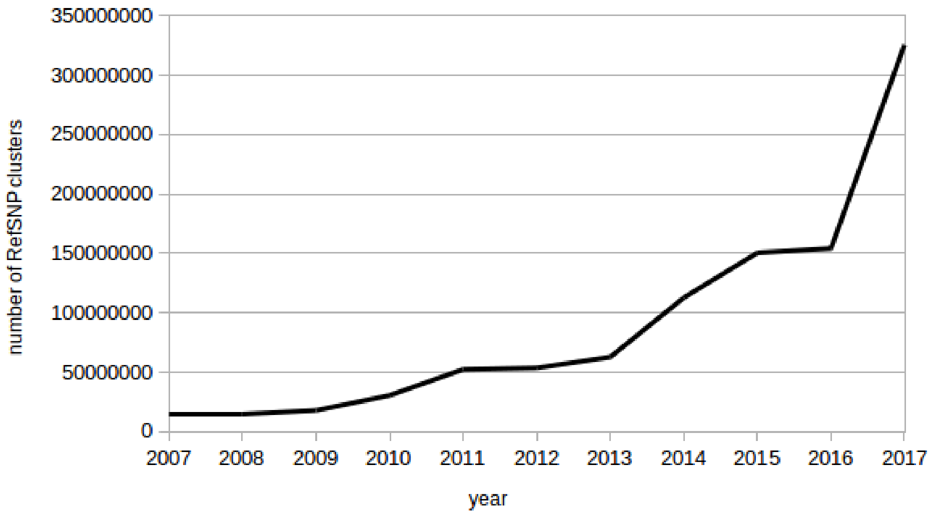


Figure 1: Growth of human variation data in dbSNP in the last 10 years

Quality of variant databases is a very important aspect to consider, because the information in these databases can be used in health decision-making, research and clinical practice. Quality evaluation criteria for variation databases were developed by a Human Variome Project (HVP) workgroup, and are divided into four major areas, each having several components (Vihinen, et al., 2016). These areas are data quality, technical quality, accessibility and timeliness.

Gene-variant databases or LSDBs focus either on a single gene or on a group of genes related to certain diseases and are generally considered as the most reliable source of variation as these are typically curated by experts in the genes and diseases (Vihinen, et al., 2016).

The formats and platforms used for LSDBs and central databases is often very different, making comparison of variants complicated. Many LSDBs are using the Leiden Open Variation Database (LOVD) management software (Fokkema, et al., 2011) which is developed to provide a flexible and freely available tool for gene-centered collection and display of DNA variations. Other frequently used database management software includes MUTbase (Riikonen and Vihinen, 1999) and Universal Mutation Database (UMD) platform (Beroud, et al., 2005). The LOVD software (version 3) also provides patient-centered data storage and storage of NGS data. LOVD is open source, its underlying relational database management system being MySQL (www.lovd.nl). LOVD also offers to host databases on their servers, freeing database managers of making backups and updating their software.

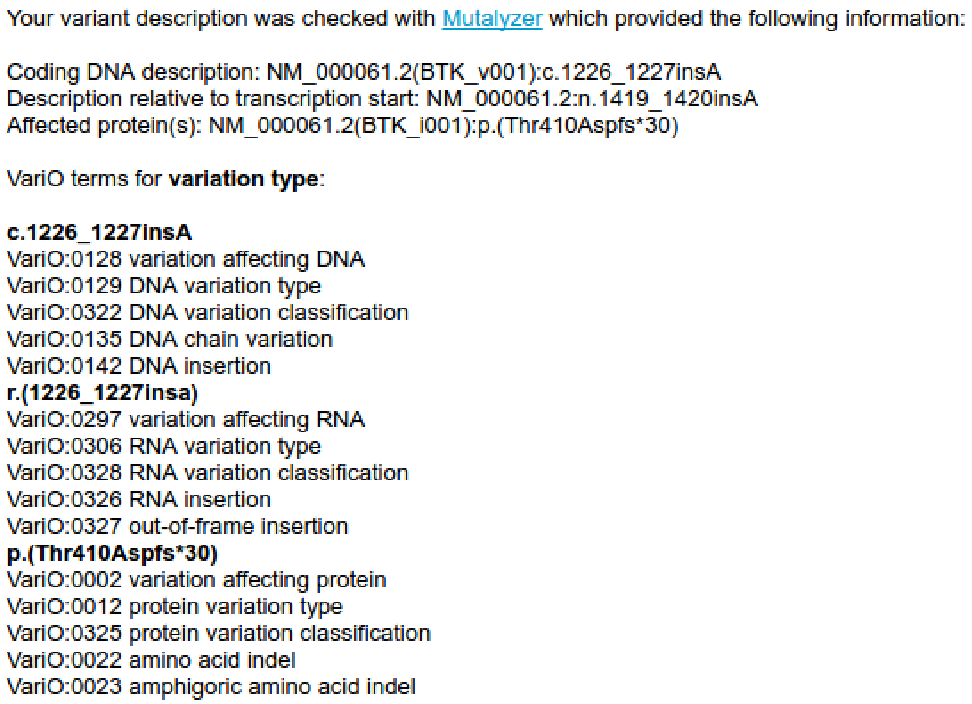
Standardization is one component of data quality. A systematic representation of information facilitates data integration, comparison of data, automated searching within and across databases, and the development of dedicated software tools. One of the recommendations for LSDBs is the use of standardized nomenclature (Cotton, et al., 2008). The HUGO Gene Nomenclature Committee (HGNC) provides systematic gene names and symbols (Yates, et al., 2017). Standardized reference sequences are generated by the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI) in the Locus Reference Genomic (LRG) sequence format (Dalgleish, et al., 2010). If for a certain gene no LRG is available, a request for one can be made, or one can use a NCBI reference sequence (RefSeq) (O'Leary, et al., 2016). Systematics for the description of sequence variants can be found in the Human Genome Variation Society (HGVS) nomenclature (den Dunnen and Antonarakis, 2000; den Dunnen, et al., 2016). General recommendations are provided, as well as specific recommendations on the DNA, RNA and protein level. The HGVS recommendations are designed to be accurate, unambiguous, stable, meaningful, but also flexible to correct inconsistencies and to extend the nomenclature for cases not previously covered. Three organizations, the HGVS, the HVP and the Human Genome Organization (HUGO), established the Sequence Variant Description Working Group (SVD-WG). The SVD-WG takes care of the nomenclature website (http://varnomen.hgvs.org/), answers questions, handles incoming requests to change or extend the recommendations, prepares, where necessary, proposals for community consultation, publishes new versions of the standards and assigns HGVS nomenclature version numbers. The HGVS nomenclature has been widely adopted and has become an internationally accepted standard. All major variant databases, including dbSNP, support the HGVS nomenclature, and organizations such as the American College of Medical Genetics and Genomics (ACMG) are recommending its use for the interpretation of sequence variants (Richards, et al., 2015). Mutalyzer is a tool that has been developed to check if an HGVS description is correct (https://mutalyzer.nl) (Wildeman, et al., 2008). Next to this name checking function, the tool has several other options, such as a syntax checker, a position converter, a name generator and a description extractor. The hgvs Python package has similar functionalities: parsing, representing, formatting, and mapping variants between genome, transcript, and protein sequences (Hart, et al., 2015).

The abovementioned standardization examples concern variant descriptions. Annotation, i.e. adding explanatory or commentary notes, is often used to enrich data. Genome annotation or DNA annotation is the process of adding layers of analysis to and interpretation of a sequence, and is done at the nucleotide level, protein level and process level (Stein, 2001). An option to come to a standardized annotation of variant descriptions is the use of an ontology, a controlled vocabulary conceptualizing a knowledge domain by defining the central terms and their relationships. Annotation of variation using an ontology will facilitate data integration and data mining, pattern recognition and statistics, and the development of analysis and prediction tools (Vihinen, 2014a). Some well-known and widely used ontologies in live sciences are the Gene Ontology (GO) and the Sequence Ontology (SO). The GO is a controlled vocabulary for describing the roles of genes and gene products (Ashburner, et al., 2000). The SO (Eilbeck, et al., 2005) is for describing features and properties of biological sequences.

Variation databases and annotation

The recently developed Variation Ontology (VariO; http://variationontology.org) is a specific ontology for describing and annotating types, effects, and mechanisms of variations (Vihinen, 2014a). The ontology has four major levels, type, function, structure and property, in addition to the three molecular levels DNA, RNA and protein. VariO terms can be further modified by attribute terms.

Variation type terms describe the origin and classification of variation. Examples of variation type terms at all three molecular levels, DNA, RNA and protein, are given in Fig. 2.



Your variant description was checked with Mutalyzer which provided the following information:

Coding DNA description: NM_000061.2(BTK_v001):c.1226_1227insA
Description relative to transcription start: NM_000061.2:n.1419_1420insA
Affected protein(s): NM_000061.2(BTK_i001):p.(Thr410Aspfs*30)

VariO terms for **variation type**:

c.1226_1227insA
VariO:0128 variation affecting DNA
VariO:0129 DNA variation type
VariO:0322 DNA variation classification
VariO:0135 DNA chain variation
VariO:0142 DNA insertion
r.(1226_1227insa)
VariO:0297 variation affecting RNA
VariO:0306 RNA variation type
VariO:0328 RNA variation classification
VariO:0326 RNA insertion
VariO:0327 out-of-frame insertion
p.(Thr410Aspfs*30)
VariO:0002 variation affecting protein
VariO:0012 protein variation type
VariO:0325 protein variation classification
VariO:0022 amino acid indel
VariO:0023 amphigoric amino acid indel

Figure 2: VariOtator output for variation type. The variant description NM_000061.2:c.1226_1227insA was used as input.

The full lineage up to but not including the top-level term, ‘VariO:0001 variation’, is provided. This example was generated with the VariOtator tool, using the coding DNA variant description ‘NM_000061.2:c.1226_1227insA’ as input.

VariO function terms describe the functions affected by the variation, e.g. ‘VariO:0399 effect on translation’. VariO structure terms are for describing the affected structural features, e.g. ‘VariO:0085 effect on alpha helix’. Property terms can be used for diverse features, an example is ‘VariO:0301 effect on RNA stability’.

Since ontologies can be complicated, we developed the user-friendly, easy-to-use online application VariOtator, to promote the use of VariO and to safeguard the consistency of annotation with VariO. VariOtator assigns automatically VariO terms to variant type descriptions and suggests terms for the other three VariO levels, function, structure and property.

IDbases: BTKbase

BTKbase (Vihinen, et al., 1995; Väliaho, et al., 2006) is an example of a locus-specific database (LSDB) which is manually curated by experts on the gene and disease. It is one of the 131 IDbases, LSDBs for immunodeficiency-causing variations. IDbases contain in addition to gene variation, also information about clinical presentation (Piirilä, et al., 2006). The IDbases were implemented under the MUTbase system (Riikonen and Vihinen, 1999) but are gradually being moved to the LOVD system (Fokkema, et al., 2011).

Bruton tyrosine kinase (*BTK*) variations lead to X-linked agammaglobulinemia (XLA, MIM# 300755), a hereditary primary immunodeficiency (PID) (Tsukada, et al., 1993; Vetrie, et al., 1993). XLA is characterized by failure to produce mature B lymphocytes and is associated with a failure of Ig heavy chain rearrangement. This block in B cell differentiation results in severely decreased numbers of B lymphocytes and an almost complete lack of plasma cells and very low or missing immunoglobulin levels of all isotypes. The disorder resides in the *BTK* gene, a key regulator in B-cell development (Rawlings and Witte, 1994). The *BTK* gene (HGNC approved name Bruton tyrosine kinase; reference sequence LRG_128) has 19 exons, with exon 1 completely, and exons 2 and 19 partially outside the coding region. The BTK protein (UniProt ID: Q06187; recommended name tyrosine-protein kinase BTK) belongs to the Tec family of related cytoplasmic protein kinases, and has the following domain organization in common: the N-terminus pleckstrin homology (PH) domain, the Tec homology (TH) domain, the Src homology 3 (SH3) domain, the Src homology 2 (SH2) domain, and the catalytic tyrosine kinase (TK) domain.

BTKbase (https://structure.bmc.lu.se/idbase/BTKbase/) adheres to a number of standards including HGNC gene name, LRG reference sequence, HGVS variation nomenclature, and follows the recommendations for LSDBs (Vihinen, et al., 2012) and their curation (Celli, et al., 2012). In addition to the variant descriptions, the records contain literature citations and annotation in detail at DNA, RNA, and protein levels, including annotation with VariO terms. Also, the most important clinical parameters and laboratory findings are included, if available.

An overview of the variation statistics in BTKbase is provided. These include the distribution of variants to *BTK* gene regions and BTK protein domains, the distribution of variation types based on their effects on DNA or RNA level, and the distribution of amino acid substitutions (AASs).

Benchmark datasets: VariSNP and representativeness

A major problem with variant databases is the interpretation of the variants, i.e. what is the consequence of the variation on the phenotype. Large numbers of identified variations in sequencing projects are novel and knowledge about disease association is absent (Niroula and Vihinen, 2016). Experimental analysis of the variations would be too costly and time-consuming, and in practice impossible due to the large numbers of newly identified variants. Therefore, many computational tools have been developed to predict the pathogenicity of variants, and are based on different principles. Evolutionary conservation is among the most useful data items for predictions. Evolution based tools include PANTHER (Thomas and Kejariwal, 2004), PROVEAN (Choi, et al., 2012) and SIFT (Ng and Henikoff, 2001). Many methods utilize machine learning (ML) algorithms. PON-P2 (Niroula, et al., 2015) is a ML-based method using features such as evolutionary sequence conservation, amino acid properties, and functional annotations. Another category are the meta-predictors, methods that use the predictions of other methods to make their own decisions. Some examples are PON-P (Olatubosun, et al., 2012), Condel (Gonzalez-Perez and Lopez-Bigas, 2011) and PredictSNP (Bendl, et al., 2014). For the development and assessment of predictor performance, approved and widely accepted benchmark datasets are needed (Nair and Vihinen, 2013).

Benchmark datasets are standard representative datasets with known outcome and they are essential for assessment of predictor performance (Vihinen, 2012). Benchmark datasets can also be used for training and testing of new predictors when based on ML methods. The VariBench database holds the first systematic benchmark datasets for variation effects (Nair and Vihinen, 2013). The criteria considered for inclusion of data and datasets included relevance, representativeness, non-redundancy, experimentally verified cases, positive and negative cases, scalability, and reusability. Since the release of VariBench, many new variants had been discovered and the need for newer and larger datasets was apparent. Easy updating of the datasets was also a requirement because of the accelerating speed of variation detection. For this, we developed VariSNP, a database with neutral variant datasets. dbSNP was used as the source of data for the new benchmark datasets. It is considered the largest variation database (over 324 million reference SNPs, April 2017) and is a public domain archive for a broad collection of simple genetic variations (Sherry, et al., 2001).

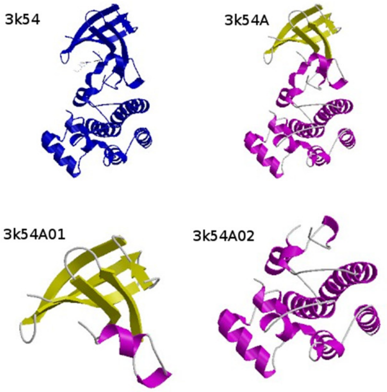
Since our aim was to generate benchmark datasets for benign variants, and dbSNP contains both disease-related and non-disease related variants, subsets of dbSNP were created by filtering out variants found in ClinVar, UniProt and PhenCode datasets, and which were annotated as either pathogenic or disease-causing. These three databases are considered to be among the most comprehensive resources for disease-related variants.

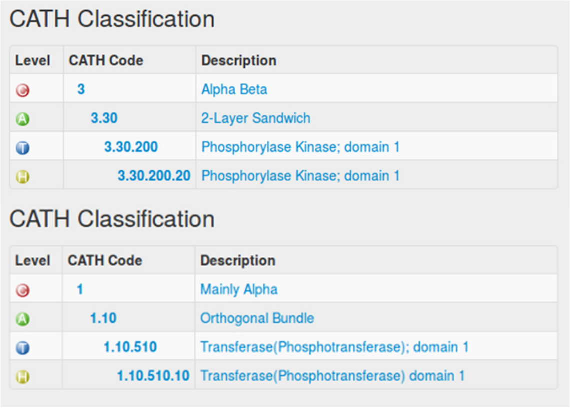
ClinVar contains reports of relationships among human variations and medically relevant phenotypes with supporting evidence (Landrum, et al., 2014). The Swiss-Prot section of UniProt (UniProtKB/Swiss-Prot) contains manually annotated records with information on protein sequences extracted from literature and curator-evaluated computational analysis (UniProt Consortium, 2017). Manual curation includes a thorough review of available information on sequence variants (mostly single amino acid substitutions) and associated genetic disease information. PhenCode (phenotypes for ENCODE) complements human phenotype and clinical data in various LSDBs with data on genome sequences, evolutionary history, and function from the ENCODE project and other resources in the University of California, Santa Cruz (UCSC) Genome Browser (Giardine, et al., 2007).

One of the criteria and requirements of benchmark datasets is their representativeness, how well do the data in the dataset provide a good example of existing cases, the population. For variant pathogenicity prediction, this means that the data represent the space of variations and their effects. Pathogenicity/tolerance prediction methods are often based on ML methods which require training and testing datasets consisting of known examples, and the tool will not attain its complete performance if these examples do not represent the variations space.

To evaluate the representativeness of datasets, we studied some features of the data, including coverage and distributions of the human proteome space, the CATH classification, the relation of proteins to Pfam families, the Enzyme Commission (EC) classification of proteins, and the allocation of GO terms to proteins.

CATH (Class, Architecture, Topology, Homology) is a classification of protein structures, providing information on the evolutionary relationships of protein domains (Sillitoe, et al., 2015). Protein structures are obtained from the Protein Data Bank (PDB) and split into the consecutive polypeptide chains, where applicable. Using both automated methods and manual curation, protein domains are identified within these chains. The domains are then classified within the CATH hierarchy: according to their secondary structure content domains are assigned to one of the four Class levels, all alpha, all beta, mixture of alpha and beta, little secondary structure. At the Architecture level, secondary structure arrangement information is used for classification. The Topology level relates to information about the connection and arrangement of the secondary structure elements; assignments to the Homologous superfamily level are made if there is good evidence of evolutionary relationship, i.e. they are homologous (http://www.cathdb.info/wiki). Examples of CATH classifications are given in Fig. 3.





CATH Classification

Level	CATH Code	Description
C	3	Alpha Beta
A	3.30	2-Layer Sandwich
T	3.30.200	Phosphorylase Kinase; domain 1
H	3.30.200.20	Phosphorylase Kinase; domain 1

CATH Classification

Level	CATH Code	Description
C	1	Mainly Alpha
A	1.10	Orthogonal Bundle
T	1.10.510	Transferase(Phosphotransferase); domain 1
H	1.10.510.10	Transferase(Phosphotransferase) domain 1

Figure 3: CATH classification of two domains in the PDB structure 3k54. 3k54: PDB structure, 3k54A: chain A, 3k54A01: domain 01, 3k54A02: domain 02, top CATH classification: 3k54A01, bottom CATH classification: 3k54A02.

The PDB structure 3k54 is the crystal structure for human BTK kinase domain (Marcotte, et al., 2010), and has one chain, 3k54A, and two structure domains, 3k54A01 and 3k54A02. The CATH classification of the two domains is provided from the 1st (Class) level to the 4th level (Homology or superfamily).

The Pfam database (http://pfam.xfam.org) is a collection of protein domain families. Protein domains are functional regions of which proteins in general have one or more. Each Pfam entry is represented by multiple sequence alignments and hidden Markov models (Finn, et al., 2016). Pfam entries are classified into one of six categories, depending on the length and nature of the sequence regions included in the entry. The six categories are domain, family, repeat, motif, coiled-coil and disordered.

EC numbers relate to the classification scheme for enzymes, which is based on the chemical reactions they catalyze (International Union of Biochemistry and Molecular Biology. Nomenclature Committee and Webb, 1992). The 6 main classes are EC 1: Oxidoreductase, EC 2: Transferases, EC 3: Hydrolases, EC 4: Lyases, EC 5: Isomerases and EC 6: Ligases. Each of these main classes have up to three levels of subclasses, so a full EC classification is made up of the 4 numbers referring to the 4 levels of the classification. The database is accessible at http://www.chem.qmul.ac.uk/iubmb/enzyme/.

The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects: biological process, cellular component and molecular function in a species-independent manner (http://www.geneontology.org). Biological process refers to a biological objective to which the gene or gene product contributes. Molecular function is defined as the biochemical activity (including specific binding to ligands or structures) of a gene product. Cellular component refers to the place in the cell where a gene product is active.

Protein groups and sensitivity to variation

Variation databases contain only known variants. One fundamental question related to sequence variants is how many of possible variants are harmful or benign. Despite numerous sequencing studies, proportions of harmful and harmless substitutions are not known for proteins and protein groups. Related to this is whether there are differences in variant frequencies between proteins, protein groups, protein structural classes, chromosomes, or amino acid substitution types. Functional analysis is possible using experimental methods such as massively parallel mutagenesis. This technique has been used for certain genes/proteins (Haller, et al., 2016; Starita, et al., 2015), but methods like this have not been/are not being employed on a large scale.

There is extensive variation in the mutation rate between and within human genes associated with Mendelian disease (Smith, et al., 2016), differences in proportions of harmful and benign variants are to be expected between gene and disease classes. Genes related to cardiomyopathy were found to have very low rates of genetic variation (Pan, et al., 2012). Other studies have investigated the numbers of pathogenic variants in certain proteins (Niroula and Vihinen, 2015; Väliaho, et al., 2015), but no systematic studies have been performed. Sufficiently large experimental datasets are not available, therefore we investigated the proportions of harmful and benign variants in nine protein groups by predicting the outcome/effect of all possible single amino acid substitutions (SAASs) in proteins belonging to these groups.

Aims of the study

The main aims of this study were to develop tools for improving variation annotation, specifically the standardization of annotation, and for supporting the development of methods for variant interpretation.

More specific aims were:

Development of an automated tool for annotating variant descriptions with Variation Ontology terms (paper I).

Provide an overview of variation in BTK (paper II).

Generate variation benchmark datasets from dbSNP (paper III).

Examine differences in sensitivity to variation in disease and non-disease related protein groups (paper IV).

Investigate the representativeness of variation benchmark datasets (paper V).

Materials and methods

Variation data

VariOtator (Paper I)

VariO was downloaded from its website, http://variationontology.org. The ontology is available in three file formats of which we used the Web Ontology Language (OWL) format. OWL is an ontology language for the Semantic Web with defined meaning. It is designed for use by applications that need to process the content of information instead of just presenting information to humans (https://www.w3.org/TR/owl2-overview/).

BTKbase (Paper II)

Statistics on the three molecular levels, DNA, RNA and protein, were determined in BTKbase (Väliaho, et al., 2006). These included the distribution of variants and variant types in the BTK domains and types of nucleotide and amino acid substitutions.

Database of variation benchmark datasets VariSNP (Paper III)

Variation data were collected from dbSNP (Sherry, et al., 2001). For filtering out pathogenic/disease-causing variants, variation data were collected from the PhenCode (Giardine, et al., 2007), the ClinVar (Landrum, et al., 2014) and the UniProtKB/Swiss-Prot (UniProt Consortium, 2017) databases. These three databases are considered to be among the most comprehensive resources for disease-related variants. Since our aim was to generate benchmark datasets for benign variants, and dbSNP contains both disease-related and non-disease related variants, subsets of dbSNP were created by filtering out variants found in ClinVar, UniProt and PhenCode datasets, and which were annotated as either pathogenic or disease-causing. ClinVar contains reports of relationships among human variations and medically relevant phenotypes with supporting evidence (Landrum, et al., 2014). The Swiss-Prot section of UniProt (UniProtKB/Swiss-Prot) contains manually annotated records with information on protein sequences extracted from literature and curator-evaluated computational analysis (UniProt Consortium, 2017). Manual curation includes a thorough review of available information on sequence variants (mostly single amino acid substitutions) and associated genetic disease information. PhenCode (phenotypes for ENCODE) complements human phenotype and clinical data in various LSDBs with data on genome sequences, evolutionary history, and function from the ENCODE project and other resources in the University of California, Santa Cruz (UCSC) Genome Browser (Giardine, et al., 2007).

Sensitivity of protein categories (Paper IV)

We investigated all possible SAASs in nine groups of proteins with a highly reliable prediction method, PON-P2 (Niroula, et al., 2015) to see if there are differences in sensitivity for variations and to reveal the proportions of harmful and benign SAASs. Although there are several prediction methods for variation consequences, numerous assessments have indicated that PON-P2 has superior performance among related tools (Bendl, et al., 2014; Niroula, et al., 2015; Riera, et al., 2016). PON-P2 is also fast and has a low error rate. The groups of proteins included those involved in diseases as well as housekeeping and non-disease genes and proteins.

Protein sequences were collected for nine categories of proteins, representing those involved in diseases as well as housekeeping and non-disease proteins. The disease related groups were proteins from the so-called actionable genes, cancer genes, cardiomyopathy related genes, developmental disorder genes, genes for early infantile epileptic encephalopathy, PID genes, and neurodegenerative disease-related genes. Non-disease related proteins were from a selection of housekeeping genes and from a random selection of genes from the HGNC database which were not disease-related.

The actionable genes group consisted of 56 genes from the ACMG recommendations (Green, et al., 2013) for which findings should be reported, since therapies to treat individuals with variants in these genes exist. The cancer group consisted of 166 genes, selection criteria ‘somatic’ and ‘missense’, downloaded in April 2016 from the Cancer Gene Census repository (http://cancer.sanger.ac.uk/census). The 46 cardiomyopathy genes were from (Pan, et al., 2012). The 53 genes in the developmental set were selected (criterion: “Developmental”) from the dataset published by (Goh, et al., 2007). The 37 epilepsy-related genes were taken in April 2016 from the Online Mendelian Inheritance in Man (OMIM) database (Amberger, et al., 2015), using phenotypic series PS308350 (Epileptic encephalopathy, early infantile) as selection criterion. The 200 housekeeping genes were a random selection from the dataset published by (Eisenberg and Levanon, 2013). The 283 genes in the PID set were taken from the ImmunoDeficiency Resource (Samarghitean, et al., 2007), from IDbases (Piirilä, et al., 2006), from a classification by the International Union of Immunological Societies (IUIS) expert committee for PIDs (Picard, et al., 2015) and from a recent review (Vihinen, 2015a). The 126 neurodegenerative genes came from the Neurodegenerative Disease Variation database (http://bioinf.suda.edu.cn/NDDvarbase/LOVDv.3.0/genes). The non-disease related genes were a random selection of 200 genes from the HGNC database (Yates, et al., 2017) which were not disease-related (no OMIM id) and which were not in the housekeeping set.

Gene-specific indices were obtained from the publications (Aggarwala and Voight, 2016; Itan, et al., 2015; Lek, et al., 2016; Petrovski, et al., 2013; Samocha, et al., 2014).

Representativeness of variation benchmark datasets (Paper V)

The VariBench (Nair and Vihinen, 2013) and VariSNP (paper III) databases contain variation benchmark datasets which are used for training and testing of prediction methods. The representativeness of datasets from these databases and some other datasets was studied. As no related research on this topic was found, we chose some features thought to capture the representativeness.

Variation benchmark datasets were collected from the VariSNP database, from the VariBench database, from filtered versions of five benchmark datasets for pathogenicity prediction (Grimm, et al., 2015), from PolyPhen-2 HumVar training datasets (Adzhubei, et al., 2010) and from SwissVar (Mottaz, et al., 2010). Protein structure data were downloaded from the PDB in Europe (https://www.ebi.ac.uk/pdbe/) or the Research Collaboratory for Structural Bioinformatics (RCSB) PDB (http://www.rcsb.org/pdb), protein sequence data from the UniProt database. Cross-mapping files (UniProt-PDB) were obtained from the EBI. CATH data were downloaded from the CATH website, Pfam data from the Pfam database. EC-UniProt ID cross references came from the UniProt Retrieve/ID mapping service (http://www.uniprot.org/uploadlists/), GO terms cross-references were obtained using the EBI QuickGO service (http://www.ebi.ac.uk/QuickGO/GAnnotation). Number of genes per chromosome and Ensembl-UniProt cross-references were obtained using the Ensembl Biomart service (http://www.ensembl.org/biomart/martview/).

The distribution of variants over the human chromosomes was examined, using the number of genes per chromosome as a weighting factor.

Background information of the studied features is needed for comparing. This was easy to obtain for some of the investigated properties, such as the distribution of genes over the human chromosomes and the distribution of GO terms in the human proteome. For other features this was less obvious and we had to consider the present understanding of the representative event space, e.g. in the case of protein folds we used structures in the PDB (Berman, et al., 2000) as background data.

Background data for CATH superfamilies were taken as follows: representative protein chains were obtained from RCSB PDB by using a file with protein chain clusters with 95% identity. The first chain from each (12,583) cluster was taken as a representative and the frequencies of CATH superfamilies were determined for each domain in that chain. For the human proteome, 4 classes, 30 architectures, 508 topologies and 907 superfamilies were determined.

For the Pfam background data, the frequencies of Pfam domains (5,734) in the file with UniProt ID-Pfam ID cross-references (17,340) were determined.

The EC background data consisted of 4,220 human proteins with one or more EC numbers at level 4 (the full numbers). At the first level these were 4,692 proteins, at the second level 4,605 and at the third level 4,479. Sometimes a classification of a protein does not include all levels, that explains the differences in these numbers.

Mapping of the 20,201 proteins to GO resulted in 19,137 UniProt sequences with one or more GO identifiers. The frequencies of the unique GO terms were calculated and served as background.

Methods

Most data were downloaded as tab-delimited files, except PhenCode data which were only available as MySQL tables. These were stored in a local MySQL database. All variant selection, filtering and mapping steps were performed with Python (2.7) scripts. Filtering was done by comparing variant descriptions following the HGVS (den Dunnen, et al., 2016) recommendations. HGVS variant descriptions were checked with the Mutalyzer variant nomenclature checker (Wildeman, et al., 2008). For statistical analyses, the stats package from the Python SciPy library was used. From this package, specifically the implementations of the chi square test (scipy.stats.chisquare), of the binomial test (scipy.stats.binomial) and of the Kolmogorov-Smirnov 2-sample test (scipy.stats.ks_2samp) were used. Predictions of the consequences of amino acid substitutions were obtained using the PON-P2 prediction method (Niroula, et al., 2015). Webpages were developed using PHP and JavaScript, web services with the Python RDFLib library and soaplib package. To retrieve information from the ontology, the SPARQL query language was used for searching the OWL version of VariO.

Summary of results

The quality of variant data in databases and benchmark datasets is of utmost importance. We developed some tools and analyzed a variant database and variant benchmark datasets with regard to quality aspects.

As stated before, standardization is an important component of quality. Standardization of annotation can be achieved with the use of an ontology. VariO is an ontology specifically for variation. To assure the consistent use of VariO, we developed a tool, VariOtator, for automatically annotating variant descriptions with VariO (paper I). Variation databases are for storing, annotating and making variation data publicly available. As an example of an LSDB, the variation database BTKbase is presented in paper II. The interpretation of the effects of variants is often lacking and experimental analysis is not feasible at a large scale, so computational prediction methods are being used. Benchmark datasets are needed for the development and performance assessment of such methods. A benchmark database for neutral variants, VariSNP, was generated by selecting variants from dbSNP and filtering for deleterious variants (paper III). Variation databases contain only variants which have been identified in e.g. sequencing projects. The proportions of benign and pathogenic variants in proteins and protein groups are not known. Predictions of harmful and harmless substitutions of all possible single amino acid substitutions were made for proteins in nine disease and non-disease groups (paper IV). Representativeness of a dataset, i.e. how well does a dataset represent the space, is one quality aspect of benchmark datasets. The representativeness of existing benchmark datasets was investigated (paper V).

VariOtator, an online tool for variation annotation (paper I)

The user-friendly VariOtator tool assigns VariO terms automatically for variation type annotations. Variant descriptions following the HGVS nomenclature are accepted and a VariO annotation is generated, either with or without the full ontology lineage. When the user provides a full coding DNA description, i.e. with a reference sequence, the variant description is first checked with the Mutalyzer name checker (https://mutalyzer.nl). The RNA description is predicted from the DNA description, the protein prediction is provided by Mutalyzer. The variation types at all three molecular levels are then determined using these descriptions, e.g. 'ins' is an insertion. These variation types are looked up in the OWL version of VariO and either just the top VariO term or the term including its full lineage in the ontology is returned, depending on what the user had chosen. An output example of the Graphical User Interface (GUI) version of VariOtator (http://variationontology.org/VariOtator.php) for variant type is given in Fig. 2. The HGVS variant description (at coding DNA level) used as input was NM_000061.2:c.1226_1227insA, an insertion of an adenine between positions 1226 and 1227 in the reference sequence (NCBI RefSeq) NM_000061, version 2. The variation type annotation consists of a part on the DNA level, a part on RNA level and a part on the protein level. Note that the last two parts are based on predicted variant descriptions, hence the brackets in the HGVS RNA and protein descriptions.

This fully automated part is available in several implementations; the GUI and three batch versions, one with a web interface that requires a tab-delimited file as input, and two stand-alone versions (Linux and Windows) requiring an LOVD download file as input. There is also a web service available for programmatic access.

For the VariO sublevels function, structure and property, suggested annotation is provided based on provided details. This part could not be automated, the user chooses the relevant term in every step. More than one term can be chosen, and attribute terms can be added to specify previously chosen terms. Terms of the Evidence Ontology (Chibucos, et al., 2014) can be added to VariO structure and property terms to describe the experimental methods and evidence on which the annotation is based on.

BTKbase, an example of a curated variant database (paper II)

The distribution of variations and variation types is presented in Table 5.1 (paper II). Although variants appear in all gene regions and BTK domains, this distribution is not even. Some domains contain more than the expected number of variants, some less, compared to a random distribution. Likewise, some exons contain more than the expected number of variants. Missense variations causing AASs was the largest group. No AAS were found in the SH3 domain. Arginine was found to be the most substituted amino acid, this also leads to the enrichment of tryptophan. Proline is the amino acid to which most other amino acids were substituted to. On nucleotide level, G>A and C>T substitutions were the largest categories of substitutions. Compared to the other IDbases, BTKbase shows a very similar distribution of variation types (Piirilä, et al., 2006).

Variation benchmarks: VariSNP (paper III)

A total of 13 datasets for neutral variations could be generated, following the functional categorization used in dbSNP (Table 1, paper III). These datasets are single nucleotide variants, non-coding transcript variants, downstream variants 500B, frameshift variants, cds-indel variants splice acceptor variants, splice donor variants, stop gained variants, stop lost variants, synonymous codon variants, upstream variants 2KB, utr-3-prime variants, and intronic variants. The dataset with the intronic variants is by far the largest set, at the time of publication of paper III over 5 million variants (more than 90% of all variants), at present (update 2017-02-16) already almost 26 million variants. This illustrates the enormous growth of variant data in dbSNP. Future updates are expected to generate much bigger datasets.

The distribution of variants in the single nucleotide variants dataset was studied to some detail. As in BTKbase, the G>A substitutions were the largest class, followed by C>T substitutions. Both types are transitions, and the higher rate of these were found to be typical for human genes (Stephens, et al., 2001). Of the AASs, arginine is the most frequently substituted amino acid, and arginine to glutamine substitutions are the most frequent. This overrepresentation of substituted arginines can be explained by the high mutability of codons containing CpG dinucleotides (Ollila, et al., 1996); four out of six codons coding for arginine have a CpG dinucleotide in the first and second position.

Information in the single nucleotide variants dataset contains information from three sources: dbSNP, the Mutalyzer Name Checker, and from the VariOtator tool, which generated the VariO annotation. The other datasets do not contain all this information because for instance intronic variant descriptions cannot be checked with the Mutalyzer Name Checker.

These datasets are being updated on a regular basis.

Protein groups and sensitivity to variation (paper IV)

Nine datasets with in total 1066 genes and corresponding proteins were collected, of which 996 genes were unique, since there was some overlap between some groups (paper IV). All 19 possible SAASs for each position in all proteins were generated. For these 13 540 914 SAASs, predictions were made with the PON-P2 predictor, which classifies variants into three categories: pathogenic, benign, and unknown (Niroula, et al., 2015).

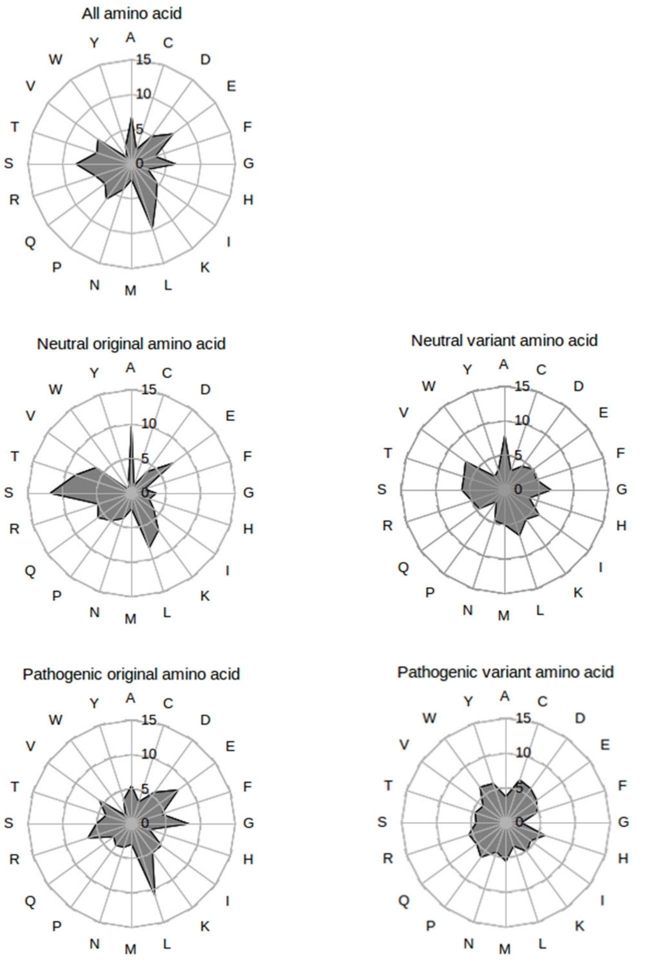
The nine groups varied in size, the group related to epilepsy being the smallest with 37 predicted proteins and the group related to PIDs being the largest with 263 proteins. There is some overlap of a few groups, which was the largest for the cancer and actionable groups. This is expected since the actionable group contains very well studied genes and proteins from the other groups.

Due to requiring more than one nucleotide substitution in a codon, most studied SAASs are unlikely to occur in nature. Only 150 out of 380 possible SAASs can originate from single-nucleotide substitutions.

Large differences in the proportions of harmful, benign and unknown variants were found between the groups. The largest proportion of harmful variants was found in the cancer group (ratio harmful/neutral 4.70), the lowest ratios were found in the non-disease group (0.06) and the housekeeping group (0.90). However, some proteins in the housekeeping group have high percentages of pathogenic predictions: the ubiquitin-conjugating enzyme E2 B from this group had the highest percentage of all proteins (90.3%). This illustrates the large differences in the proportions of harmful and benign variants within the groups. Differences in the percentages of unknown variants were also large, both between and within the protein groups. These were in the range from 99% for histone H3.3 in the cancer group to almost 0 for several proteins in the non-disease group.

Protein domains which appeared frequently in the protein groups showed a broad spectrum of variant frequencies. The most frequently found domain, the immunoglobulin I-set domain, had the lowest proportions of neutral and pathogenic predictions together. Major protein structural class-specific differences were not detected.

All studied groups showed distinctive patterns of AAS types in the original and variant amino acids. The distributions for the original and variant AASs in the actionable dataset are presented in Fig. 4. This is the same figure as Figure 4 in paper IV, except that the titles of the plots have been corrected; an erratum has been sent to the journal and will be published soon.



All amino acid

Neutral original amino acid

Neutral variant amino acid

Pathogenic original amino acid

Pathogenic variant amino acid

Figure 4: Distributions of amino acid substitutions in the actionable dataset. Amino acid substitutions among the original and variant amino acids for those predicted to be neutral and pathogenic.

The most frequent neutral AASs are substitutions from serine (11.7%) and to alanine (7.9%). The least frequent neutral AASs are the substitutions from tryptophan (0.5%) and to proline (2.1%). Substitutions from leucine (10.9%) and to cysteine (6.5%) are the most frequent pathogenic predictions, whereas the least frequent pathogenic predictions are from tryptophan (1.5%) and to G (2.3%).

Comparison of genic intolerance scores (gene-specific indices) showed only marginal correlations with our predictions.

The distribution of predictions over the chromosomes showed that the differences for pathogenic variants are somewhat smaller than for the neutral predictions, and that the differences between the chromosomes are smaller than between the protein groups.

Differences between proteins regarding proportions of harmful, benign and unknown predictions are large. In some proteins, no AASs, in others almost all variants were predicted to be harmful.

Representativeness of benchmark datasets (paper V)

The distributions of variants over the chromosomes and proteins varies greatly between the benchmark datasets. The distributions of the variants over the whole human genome, so all chromosomes, are biased in all datasets. The differences between the chromosomes are great. The lowest number of chromosomes with an unbiased variant distribution in a dataset is 2, the highest 13 chromosomes. The distributions for the X chromosome are in all datasets biased, and for chromosome 19 in all but one dataset.

Mapping protein variant descriptions to a PDB structure, which is essential for being able to perform analyses for most other features, was only possible for fractions of the data (paper V). Mapping rates vary from 7.8-54%, with ratios in the pathogenic datasets always being higher than in the neutral counterparts.

Similar to the situation for mapping to PDB structures, mapping to CATH domains and superfamilies varies greatly, ranging from 29.5% to 69.9% for domains, and from 26.8% to 68.1% for the CATH classification (superfamilies). In our reference proteome, there are 4 classes, 30 architectures, 508 topologies (folds) and 907 homologies (superfamilies). The maximum numbers in the datasets are 4, 30, 419 and 700, respectively.

On the Class level, all datasets are unbiased. This is also the case for most datasets on the Architecture level, whereas on the Topology and Homology levels all datasets are biased.

Mapping to Pfam families could be done for 86% of the proteins in the human proteome, which mapped to 5,734 Pfam families. The distributions of variants over Pfam domains are biased for all datasets. The fractions of variants within a Pfam domain range from ~37% to ~80%, and these percentages are invariably lower for the neutral datasets compared to their pathogenic counterparts. The distributions to the Pfam families are also variable, 14 datasets mapped to more than 1,000 Pfam families, two datasets to more than 2,000.

Mapping to an EC classification could be done for 21% of the proteins in our reference dataset, to in total 1,292 EC classes (4th level, so full numbers). Since not all proteins are enzymes, the low percentage is expected. On the first level of the EC classification, no dataset has a different distribution from the reference set, whereas on the second level, 5 out of the 24 sets are biased. On the third level, 20 sets are biased, on the fourth level all sets are biased.

Mapping to GO resulted in 17,637 unique GO terms for 95% of the proteins in the reference set. The distribution of variants over GO terms is not biased for all sets on the aspect level, but on the term level, all sets are biased.

Discussion

As part of our efforts to improve the quality of variation data, we developed the VariOtator tool for assisting in the use of VariO, the ontology for standardized annotation of variation. We described the LSDB BTKbase, which was also used as a case study for assessing database quality. We generated the benchmark database VariSNP, which can be used for the development and performance assessment of prediction methods. We looked at differences in sensitivity to variation in different protein and disease groups. The representativeness of benchmark datasets was our last point of attention.

VariOtator

The use of VariOtator will assist in the use of VariO and promote the consistency of VariO annotation. Since the VariO type annotation is fully automated, variation type annotation can be systematically added to databases, thus enhancing their standardization. VariOtator is being used to add VariO type annotation to BTKbase (https://databases.lovd.nl/shared/genes/BTK) and other IDbases (Piirilä, et al., 2006). The suggestions of VariO terms generated by VariOtator for the other VariO levels, function, structure, and property, aid users to choose VariO terms consistently.

Instructions for annotators are available (Vihinen, 2014b), and examples of VariO annotation can be found in (Vihinen, 2015b) or at http://www.variationontology.org/Examples.shtml. Guidelines for curating gene variant databases or LSDBs have been published (Celli, et al., 2012). Adding variant annotations by means of an automated tool such as VariOtator could expand these guidelines.

One of the central variation databases, UniProtKB/Swiss-Prot, is working on improving the representation of functional characterization data by combining ontologies such as VariO and GO. The characteristics of a normal protein can be specified by GO terms, the effect(s) of a variant on it by VariO terms (Famiglietti, et al., 2014). The use of (a batch version of) VariOtator can be of great help in this, especially for retrospective annotation of variants already in the database.

BTKbase

The distribution of variants was found to be uneven over the gene regions and corresponding BTK domains. The PH and SH2 domains contain about the expected number of variants at DNA level, the TH and SH3 domains contain less than expected and the kinase (TK) domain contain more than the expected number of variants. The uneven distribution is also the case for the variation types at protein level, e.g. no AASs were detected in the SH3 domain (paper II). This is in accordance with the findings that there is substantial variation in the mutation rate between and with human genes associated with Mendelian disease (Smith, et al., 2016). The other IDbases, of which BTKbase is one, show very similar distributions of variation types (Piirilä, et al., 2006).

The quality of BTKbase was assessed using the recently developed database quality evaluation criteria (Vihinen, et al., 2016).

BTKbase was taken as an example of a manually curated LSDB. BTKbase was recently moved to the LOVD database management system and VariO type annotation was added to each variant description. The use of VariO annotation in the database will greatly simplify the generation of database statistics. BTKbase was also used as a case study for assessing database quality (Vihinen, et al., 2016).

VariSNP

The datasets from the VariSNP database have been used in research projects to construct datasets for developing and assessing the performance of prediction methods. The predictors in question were PredictSNP2, a meta-predictor using the outcome of 5 other methods to come to a consensus score (Bendl, et al., 2016) and ENTPRISE, an ML approach using protein sequence and structure features (Zhou, et al., 2016). The VariSNP single nucleotides dataset was also used our sensitivity project (paper IV). The VariSNP database has been cited/quoted a few times as well (Niroula and Vihinen, 2016; Vihinen, 2015b).

Updates of the datasets have been made available on a regular basis, however updating these datasets will become problematic because of the enormous growth of dbSNP, the source of the VariSNP database. dbSNP is increasingly using a ‘Clinical Significance’ annotation, making it easier for users to download those data they are interested in. It remains to be seen, however, how useful and reliable that annotation is. This annotation is provided by the submitters and not interpreted by NCBI, and available for only a fraction of the variants in dbSNP. Updating the datasets will also be needed in the future. The smaller groups such as the neutral stop-loss and stop-gained datasets are of special interest since these are not available elsewhere.

Sensitivity of protein groups

There are large differences in the proportions of harmful and neutral AASs in proteins belonging to the different disease groups we studied. Differences in mutation rates between and within human genes are well known (Hodgkinson and Eyre-Walker, 2011; Smith, et al., 2016). Variation in mutation rate can be found at different scales, from the nucleotide level up to the chromosomal level. Variation at the single nucleotide level was found across bacterial and eukaryotic species: C and G nucleotides are approx. twice as mutable as A and T nucleotides (Hershberg and Petrov, 2010). Variation can be due to context, the effect of having certain nucleotides surrounding a specific site, or non-context related variation called cryptic variation (Hodgkinson, et al., 2009). The 'CpG effect' (Hodgkinson and Eyre-Walker, 2011; Ollila, et al., 1996) is an example of context-dependent effect: the rate of mutations at CG dinucleotides is ~10-fold greater than at other sites. Variation in mutations rate at larger-scale levels is also found, within and between chromosomes. At genomic level, the greatest differences can be found between the autosomes and the sex chromosomes. The Y chromosome has a mutation rate of at least 50% higher than that of the autosomes, which in turn have a ~30% higher mutation rate compared to the X chromosome. It is suggested (Smith, et al., 2016) that a reason for certain genes being associated with disease may be the presence of hypermutable sites.

This result is important to realize, e.g. when creating benchmark datasets or in the interpretation of variant effects. Prediction methods based only on evolutionary information have been shown to perform less well than methods using other features as well.

Representativeness

Variant benchmark datasets can be used for the development of prediction methods as well as for the assessment of the performance of such methods. Requirements for benchmark datasets include relevance, representativeness, non-redundancy, experimentally verified cases of both positive and negative cases, scalability and reusability (Nair and Vihinen, 2013). The representativeness of a dataset can be determined by examining features of the data thought to capture the representativeness. We examined to which extent some properties are present in variation benchmark datasets and found huge differences between the studied datasets.

The coverage of the studied features varied greatly between the datasets (Table 7, paper V), and is the best for the largest dataset. The coverage of the proteome space is in general quite low. The smallest size of a dataset would be 20,201 variants to cover each protein at least once. Some of the datasets are much bigger or close to this number, however the distribution of variant to the proteins is very unbalanced. In the studied benchmark datasets, some proteins are represented by more than 2,200 variants, many others by only one variant. Also, the uneven distribution of variants to the proteins is the main cause for the differences in the chromosomal distributions.

These differences can be explained by the huge differences in the number of known variants for certain genes and encoded proteins. Certain diseases and genes/proteins are very well studied and many variants are known. For other proteins, only one variant can have been found incidentally. The strongly biased distributions on chromosome X for the pathogenic datasets can be explained by the fact that X-linked single-gene disorders have full penetrance. We did not find the same for the Y chromosome, which is expected. This can be explained by the very low numbers of variants on this chromosome.

Coverage of the first levels of the CATH and EC classifications is in general complete, except for the smallest datasets. These levels are the least specific, with CATH having only 4 classes and EC 6, so it is easy to get complete coverage. The coverage of the highest levels of CATH and EC (4th level) varies greatly: the range for CATH is 1-77%, for EC 1-99%. The intermediate levels have intermediate coverages, so the less specific the classification, the better the coverage. The coverage of the CATH and EC levels in the datasets indicate that many protein types are present, albeit far from complete. The coverages of Pfam families and GO terms also vary greatly. There seems to be some correlation between the number of variants in a dataset and the coverage of Pfam and GO.

Depending on the level, the distributions in the datasets of the CATH and EC classifications are unbiased, partly biased or biased for all datasets. For CATH, this can be partly explained by the reduction of the coverage from the 2nd to the 3rd level by about a half, but in the EC classifications the reduction of the coverage from the 2nd to the 3rd level is much smaller. So probably other factors play a role. The distributions of Pfam and GO at terms level are always biased, which cannot be explained by a low coverage of these features, e.g. in the largest dataset, GO terms are covered by more than 98%.

Conclusions

Precision medicine is entering clinical practice nowadays. The tailored diagnosis and treatment is dependent on genome sequencing and data analysis. NGS technologies are becoming increasingly used in research and common in the clinical sector. Large sequencing projects are producing huge amounts of variation data and it is expected that this will increase at an accelerating rate in the near future. Not only will this growth put strains on data storage, data analysis will also become computationally challenging. Knowledge of the underlying mechanisms of disease is essential, but the effects of newly discovered variants are often unknown. Since experimental verification of variation effects is not feasible on a large scale, computational methods have and are being developed to predict these.

Storage of variant data in database is not only becoming demanding with regard to storage capacity, the quality of the data and databases is also of utmost importance. One aspect of database quality is standardization, making easy searching and comparison of data possible. The use of ontologies is one way of achieving standardized variation annotation. The Variation Ontology VariO was specifically developed for describing the effects, consequence and mechanisms of DNA, RNA and protein variations. In addition to these 3 molecular levels, VariO has four major levels: variation type, variation function, variation structure and variation property. We developed the online VariOtator tool to assist in the consistent use of VariO. For variation type annotation, the tool is fully automated, for the other three levels, annotation is generated based on details provided by the user.

A group of diseases which has also benefited from NGS are the PID. Variation data related to one of these, XLA, are deposited in BTKbase, an LSBD containing variants in BTK, the gene involved in BTK. Database statistics at the three molecular levels DNA, RNA and protein show a wide spectrum of variant and variation types, and differences in the distribution of variants over the BTK protein domains. BTKbase was used as a case study for evaluating the quality of databases using defined quality assessment criteria. BTKbase is being updated depending on new submissions and variation data being published. At the time of the study, BTKbase contained public data on 1375 variants in patients, of which 742 unique, at present (August 2017) these figures are 1684, and 851, respectively.

For the development and performance assessment of variant prediction tools, variant benchmark datasets are needed. We developed VariSNP, a database of neutral variants selected from dbSNP. To our knowledge, the neutral SAASs set in VariSNP was the largest and qualitatively the best neutral variants dataset available. The database is updated on a regular basis and is being used in several projects.

It is well known that groups proteins and diseases have a different sensitivity to sequence variation. To examine this for some interesting disease groups and a few non-disease groups, we predicted the outcome of all simulated SAASs in all proteins involved in these groups and found large differences within and between these groups in the ratios of harmful, benign and unknown variants. This information can be helpful in the interpretation of variant effects and in the development of computational methods/predictors.

Variant benchmark datasets are essential for the development and assessment of computational methods, especially those using ML methods. One aspect of benchmark dataset quality is its representativeness. We studied the representativeness of some benchmark datasets, using features of the data which we think to capture the representativeness and found that all datasets are more or less unrepresentative of the protein universe. Since most datasets are rather limited in size and often unbalanced for variants in certain proteins, especially in the pathogenic datasets, this is not surprising. Coverage of the features we studied was variable for the different datasets, for some features in some datasets (in general, the larger ones) however, almost complete. It is expected that using the more representative datasets in the development of ML-based prediction methods will improve their performance. Obtaining highly representative benchmark variant datasets will remain a challenging task.

Acknowledgements

I would like to thank my supervisor Mauno Vihinen. You gave me the opportunity to do this PhD study in the first place. You were always available for questions and discussions about what to study. You came with new ideas all the time, and corrected our manuscripts endlessly. Without you this thesis would not have been there.

Without my colleagues and roommates Abhishek Niroula and Gabriel Teku at BMC, working hours would have been a lot less pleasant. Thank you for your company.

There is also life next to research. I would to thank all my friends from Bike Kitchen Lund and Lund Social Group for making my social life in Lund nice. Repairing bicycles and having a beer is a good way of getting your mind somewhere else.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7(4):248-249.

Aggarwala V, Voight BF. 2016. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet 48(4):349-355.

Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P and others. 2017. Ensembl 2017. Nucleic Acids Res 45(D1):D635-D642.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res 43(Database issue):D789-D798.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT and others. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25-29.

Bendl J, Musil M, Stourac J, Zendulka J, Damborsky J, Brezovsky J. 2016. PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions. PLoS Comput Biol 12(5):e1004962.

Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J. 2014. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol 10(1):e1003440.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28(1):235-242.

Beroud C, Hamroun D, Collod-Beroud G, Boileau C, Soussi T, Claustres M. 2005. UMD (Universal Mutation Database): 2005 update. Hum Mutat 26(3):184-191.

Carrasco-Ramiro F, Peiro-Pastor R, Aguado B. 2017. Human genomics projects and precision medicine. Gene Ther. DOI: 10.1038/gt.2017.77

Celli J, Dalgleish R, Vihinen M, Taschner PE, den Dunnen JT. 2012. Curating gene variant databases (LSDBs): toward a universal standard. Hum Mutat 33(2):291-297.

Chen L, Liu P, Evans TC, Jr., Ettwiller LM. 2017. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science 355(6326):752-756.

Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M. 2014. Standardized description of scientific evidence using the Evidence Ontology (ECO). Database (Oxford) 2014:bau075.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS One 7(10):e46688.

Consortium IHGS. 2004. Finishing the euchromatic sequence of the human genome. Nature 431(7011):931-945.

Cotton RG, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AF, Carrera P, Cox DW, Gottlieb B, Greenblatt MS and others. 2008. Recommendations for locus-specific databases and their curation. Hum Mutat 29(1):2-5.

Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW and others. 2010. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med 2(4):24.

den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15(1):7-12.

den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE. 2016. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat 37(6):564-569.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6(5):R44.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. Trends Genet 29(10):569-574.

Famiglietti ML, Estreicher A, Gos A, Bolleman J, Gehant S, Breuza L, Bridge A, Poux S, Redaschi N, Bougueleret L and others. 2014. Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. Hum Mutat 35(8):927-935.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A and others. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1):D279-D285.

Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. 2011. LOVD v.2.0: The next generation in gene variant databases. Hum Mutat 32(5):557-563.

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA and others. 2015. A global reference for human genetic variation. Nature 526(7571):68-74.

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H and others. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat 28(6):554-562.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. 2007. The human disease network. Proc Natl Acad Sci U S A 104(21):8685-8690.

Gonzalez-Perez A, Lopez-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88(4):440-449.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333-351.

Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE and others. 2013. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med 15(7):565-574.

Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW and others. 2015. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat 36(5):513-523.

Haller G, Alvarado D, McCall K, Mitra RD, Dobbs MB, Gurnett CA. 2016. Massively parallel single-nucleotide mutagenesis using reversibly terminated inosine. Nat Methods 13(11):923-924.

Hart RK, Rico R, Hare E, Garcia J, Westbrook J, Fusaro VA. 2015. A Python package for parsing, validating, mapping and formatting sequence variants using HGVS nomenclature. Bioinformatics 31(2):268-270.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6(9):e1001115.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12(11):756-766.

Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. PLoS Biol 7(2):e1000027.

International Union of Biochemistry and Molecular Biology. Nomenclature Committee, Webb EC. 1992. Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press.

Itan Y, Shang L, Boisson B, Patin E, Bolze A, Moncada-Velez M, Scott E, Ciancanelli MJ, Lafaille FG, Markle JG and others. 2015. The human gene damage index as a gene-level approach to prioritizing exome variants. Proc Natl Acad Sci U S A 112(44):13615-13620.

Lander ES. 2011. Initial impact of the sequencing of the human genome. Nature 470(7333):187-197.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W and others. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860-921.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42(Database issue):D980-D985.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB and others. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536(7616):285-291.

MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA and others. 2014. Guidelines for investigating causality of sequence variants in human disease. Nature 508(7497):469-476.

Marcotte DJ, Liu YT, Arduini RM, Hession CA, Miatkowski K, Wildes CP, Cullen PF, Hong V, Hopkins BT, Mertsching E and others. 2010. Structures of human Bruton's tyrosine kinase in active and inactive conformations suggest a mechanism of activation for TEC family kinases. Protein Sci 19(3):429-439.

McPherson JD. 2014. A defining decade in DNA sequencing. Nat Methods 11(10):1003-1005.

Mottaz A, David FP, Veuthey AL, Yip YL. 2010. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics 26(6):851-852.

Nair PS, Vihinen M. 2013. VariBench: a benchmark database for variations. Hum Mutat 34(1):42-49.

National Center for Biotechnology Information. 2014. The NCBI handbook [Internet]. The Database of Short Genetic Variation (dbSNP). The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information.

Nekrutenko A, Taylor J. 2012. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet 13(9):667-672.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res 11(5):863-874.

Niroula A, Urolagin S, Vihinen M. 2015. PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS One 10(2):e0117380.

Niroula A, Vihinen M. 2015. Classification of amino acid substitutions in mismatch repair proteins using PON-MMR2. Hum Mutat 36(12):1128-1134.

Niroula A, Vihinen M. 2016. Variation interpretation predictors: principles, types, performance, and choice. Hum Mutat 37(6):579-597.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D and others. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733-D745.

Olatubosun A, Valiaho J, Harkonen J, Thusberg J, Vihinen M. 2012. PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 33(8):1166-1174.

Ollila J, Lappalainen I, Vihinen M. 1996. Sequence specificity in CpG mutation hotspots. FEBS Lett 396(2-3):119-122.

Pan S, Caleshu CA, Dunn KE, Foti MJ, Moran MK, Soyinka O, Ashley EA. 2012. Cardiac structural and sarcomere genes associated with cardiomyopathy exhibit marked intolerance of genetic variation. Circ Cardiovasc Genet 5(6):602-610.

Peplow M. 2016. The 100,000 Genomes Project. BMJ 353:i1757.

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 9(8):e1003709.

Picard C, Al-Herz W, Bousfiha A, Casanova JL, Chatila T, Conley ME, Cunningham-Rundles C, Etzioni A, Holland SM, Klein C and others. 2015. Primary immunodeficiency diseases: an update on the classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency 2015. J Clin Immunol 35(8):696-726.

Piirilä H, Väliaho J, Vihinen M. 2006. Immunodeficiency mutation databases (IDbases). Hum Mutat 27(12):1200-1208.

Rawlings DJ, Witte ON. 1994. Bruton's tyrosine kinase is a key regulator in B-cell development. Immunol Rev 138:105-119.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E and others. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17(5):405-424.

Riera C, Padilla N, de la Cruz X. 2016. The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. Hum Mutat 37(10):1013-1024.

Riikonen P, Vihinen M. 1999. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 15(10):852-859.

Samarghitean C, Valiaho J, Vihinen M. 2007. IDR knowledge base for primary immunodeficiencies. Immunome Res 3:6.

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnstrom K, Mallick S, Kirby A and others. 2014. A framework for the interpretation of de novo mutation in human disease. Nat Genet 46(9):944-950.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94(3):441-448.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463-5467.

Schatz MC, Langmead B. 2013. The DNA Data Deluge: Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze. IEEE Spectr 50(7):26-33.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308-311.

Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG and others. 2015. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43(Database issue):D376-D381.

Smith T, Ho G, Christodoulou J, Price EA, Onadim Z, Gauthier-Villars M, Dehainault C, Houdayer C, Parfait B, van Minkelen R and others. 2016. Extensive variation in the mutation rate between and within human genes associated with Mendelian disease. Hum Mutat 37(5):488-494.

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S. 2015. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics 200(2):413-422.

Stein L. 2001. Genome annotation: from sequence to biology. Nat Rev Genet 2(7):493-503.

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH and others. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. Science 293(5529):489-493.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big Data: Astronomical or Genomical? PLoS Biol 13(7):e1002195.

Telenti A, Pierce LC, Biggs WH, di Iulio J, Wong EH, Fabani MM, Kirkness EF, Moustafa A, Shah N, Xie C and others. 2016. Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci U S A 113(42):11901-11906.

Thomas PD, Kejariwal A. 2004. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A 101(43):15398-15403.

Tsukada S, Saffran DC, Rawlings DJ, Parolini O, Allen RC, Klisak I, Sparkes RS, Kubagawa H, Mohandas T, Quan S and others. 1993. Deficient expression of a B cell cytoplasmic tyrosine kinase in human X-linked agammaglobulinemia. Cell 72(2):279-290.

UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res 45(D1):D158-D169.

Vetrie D, Vořechovský I, Sideras P, Holland J, Davies A, Flinter F, Hammarström L, Kinnon C, Levinsky R, Bobrow M and others. 1993. The gene involved in X-linked agammaglobulinemia is a member of the src family of protein-tyrosine kinases. Nature 361(6409):226-233.

Vihinen M. 2012. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics 13 Suppl 4:S2.

Vihinen M. 2014a. Variation Ontology for annotation of variation effects and mechanisms. Genome Res 24(2):356-364.

Vihinen M. 2014b. Variation ontology: annotator guide. J Biomed Semantics 5(1):9.

Vihinen M. 2015a. Immunodeficiency, Primary: Affecting the Adaptive Immune System. eLS. p 1-6.

Vihinen M. 2015b. Types and effects of protein variations. Hum Genet 134(4):405-421.

Vihinen M, Cooper MD, de Saint Basile G, Fischer A, Good RA, Hendriks RW, Kinnon C, Kwan SP, Litman GW, Notarangelo LD and others. 1995. BTKbase: a database of XLA-causing mutations. International Study Group. Immunol Today 16(10):460-465.

Vihinen M, den Dunnen JT, Dalgleish R, Cotton RG. 2012. Guidelines for establishing locus specific databases. Hum Mutat 33(2):298-305.

Vihinen M, Hancock JM, Maglott DR, Landrum MJ, Schaafsma GC, Taschner P. 2016. Human Variome Project Quality Assessment Criteria for Variation Databases. Hum Mutat 37(6):549-558.

Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29(1):6-13.

Väliaho J, Faisal I, Ortutay C, Smith CI, Vihinen M. 2015. Characterization of all possible single-nucleotide change caused amino acid substitutions in the kinase domain of Bruton tyrosine kinase. Hum Mutat 36(6):638-647.

Väliaho J, Smith CIE, Vihinen M. 2006. BTKbase: the mutation database for X-linked agammaglobulinemia. Hum Mutat 27(12):1209-1217.

Yang X, Chockalingam SP, Aluru S. 2013. A survey of error-correction methods for next-generation sequencing. Brief Bioinform 14(1):56-66.

Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. 2017. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res 45(D1):D619-D625.

Zhou H, Gao M, Skolnick J. 2016. ENTPRISE: An Algorithm for Predicting Human Disease-Associated Amino Acid Substitutions from Sequence Entropy and Predicted Protein Structures. PLoS One 11(3):e0150965.

Paper I

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

VariOtator, a Software Tool for Variation Annotation with the Variation Ontology

Gerard C. P. Schaafsma and Mauno Vihinen*

Protein Structure and Bioinformatics, Department of Experimental Medical Science, Lund University, BMC B13, Lund SE-221 84, Sweden

Communicated by A. Jamie Cuticchia

Received 13 November 2015; accepted revised manuscript 6 January 2016.
Published online 15 January 2016 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22954

ABSTRACT: The Variation Ontology (VariO) is used for describing and annotating types, effects, consequences, and mechanisms of variations. To facilitate easy and consistent annotations, the online application VariOtator was developed. For variation type annotations, VariOtator is fully automated, accepting variant descriptions in Human Genome Variation Society (HGVS) format, and generating VariO terms, either with or without full lineage, that is, all parent terms. When a coding DNA variant description with a reference sequence is provided, VariOtator checks the description first with Mutalyzer and then generates the predicted RNA and protein descriptions with their respective VariO annotations. For the other sublevels, function, structure, and property, annotations cannot be automated, and VariOtator generates annotation based on provided details. For VariO terms relating to structure and property, one can use attribute terms as modifiers and evidence code terms for annotating experimental evidence. There is an online batch version, and stand-alone batch versions to be used with a Leiden Open Variation Database (LOVD) download file. A SOAP Web service allows client programs to access VariOtator programmatically. Thus, systematic variation effect and type annotations can be efficiently generated to allow easy use and integration of variations and their consequences.
Hum Mutat 37:344–349, 2016. © 2016 Wiley Periodicals, Inc.

KEY WORDS: annotation; ontology; Variation Ontology; bioinformatics; software; database; LSDB; variation; LOVD; mutation

Introduction

Information on (genetic) variation is being collected in a wide range of databases. Central databases comprise, for example, UniProtKB [UniProt Consortium, 2015], ClinVar [Landrum et al., 2014], and Ensembl Variation [Cunningham et al., 2015]. Other types of variation databases include locus-specific databases (LSDBs), of which those using the Leiden Open Variation Database (LOVD) management software [Fokkema et al., 2011] form the majority. LSDBs are generally considered as the most reliable source of variation information as these resources are typically curated by experts in the genes and diseases.

A systematic representation of information facilitates data integration, comparison of data, automated searching within and across databases, and the development of dedicated software tools. Published recommendations for LSDBs [Cotton et al., 2008] include the use of a standardized nomenclature. Systematic gene names and symbols are implemented and approved by the HUGO Gene Nomenclature Committee (HGNC) [Gray et al., 2013]. Standardized reference sequences in the Locus Reference Genomic (LRG) sequence format [Dalgleish et al., 2010] are being created and curated at the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI). The LRG records contain stable fixed reference DNA sequences along with all relevant transcript and protein sequences essential to the description of gene variants, and an exon numbering system [MacArthur et al., 2014]. Efforts are being made to standardize variant descriptions, such as the use of the Human Genome Variation Society (HGVS) nomenclature [den Dunnen and Antonarakis, 2000]. Guidelines for establishing [Vihinen et al., 2012] and curating [Celli et al., 2012] LSDBs highlight the importance of systematics in gene variant databases.

Gene variant databases would benefit from a standardized annotation of variant descriptions, so that automated searches and analyses within and across databases would become possible and/or much easier. One way to create a standardized annotation is to use an ontology, a controlled vocabulary conceptualizing a knowledge domain by defining the central terms and their relationships. The use of consistent terminology is essential to guarantee that the information, the message, is correctly understood [Vihinen, 2015a]. The Gene Ontology (GO) [Ashburner et al., 2000] and the Sequence Ontology (SO) [Eilbeck et al., 2005] are widely used for describing gene products in terms of their associated biological processes, cellular components, and molecular functions (GO) or for describing features pertinent to sequence annotation (SO). These ontologies have a very broad scope. The Variation Ontology (VariO) [Vihinen, 2014a] was developed as a specific ontology for describing and annotating types, effects, and mechanisms of variations. Note that VariO does not contain clinical terms apart from pathogenicity association.

VariO can be used for describing variations and their consequences only, that is, it cannot be utilized for annotation of normal or wild-type situations. It should be possible to describe any type of variation and effect with VariO. The ontology is work in progress and new terms will be added whenever necessary, for example, to facilitate annotations based on novel technologies. VariO annotations

VariOtator is available at http://variationontology.org/VariOtator.php.

*Correspondence to: Mauno Vihinen, Protein Structure and Bioinformatics, Department of Experimental Medical Science, Lund University, BMC B13, Lund SE-221 84, Sweden. E-mail: mauno.vihinen@med.lu.se

Contract grant sponsors: Vetenskapsrådet (Swedish Research Council), Faculty of Medicine, Lund University; Barncancerfonden.

are made by combining terms. VariO is organized in three main levels: DNA, RNA, and protein, each of which has four sublevels: variation type, function, structure, and property. Each of these sublevels has then more detailed terms. Variation type terms describe the origin and classification of variations, including such terms as "VariO:0136 DNA substitution" and "VariO:0147 epigenetic DNA variation." General functions affected by variation can be described with function terms. Structure terms are for describing affected DNA, RNA, and protein structural features, and vary substantially between the three levels. Property terms are used for diverse characteristics, such as conservation of DNA variation site or effect on protein abundance. In addition to these four sublevels, attribute terms can be used as modifiers of the structure and property terms, for instance, to describe effects on quantity or affected interactions due to the variation. Since the function terms are general, modifying these with attribute terms does not add much information and thus attribute terms are not used with these. Specific descriptions can be made with property terms.

Guidelines for how the annotation is made and how to use VariO in different situations have been published [Vihinen, 2014a, 2015b). The flowchart for steps in VariO annotation has been described [Vihinen, 2014b]. Briefly, the database curator collects all relevant information about the variant and its effects, mechanisms, and consequences. This may include data from laboratories, databases, and literature. The precise position of the variant in the three reference sequences for DNA, RNA, and protein (where relevant) will be obtained and annotated with variation type annotations. For function annotations, the user has to choose the appropriate terms at the relevant levels DNA, RNA, and/or protein. Structural changes due to variation can be explained in detail. Depending on the annotation level, the property annotation items can vary. Both the structure and property annotations can be modulated by using attributes to make the annotations more detailed. Further, Evidence Ontology (ECO) terms can be added to provide users with the possibility to evaluate the strength of evidence behind annotations. ECO describes the methods used to obtain the annotated results [Chibucos et al., 2014].

To facilitate easy annotation and to enhance the consistency in the use of the VariO terms for annotating variants, a user-friendly online application called VariOtator was developed. For variation type annotations the tool is fully automated, it accepts variant descriptions according to the HGVS nomenclature and generates VariO annotation, either with or without the full ontology lineage. When the user provides a full coding DNA variant description, that is, with the reference sequence, the description is first checked with the Mutalyzer Name Checker (https://mutalyzer.nl) tool [Wildeman et al., 2008]. For the other sublevels, function, structure, and property, annotations cannot be automated. VariOtator generates annotation at these levels based on provided details.

Implementation

Several implementations of the VariOtator tool are available at http://variationontology.org/ (Fig. 1). There is a Graphical User Interface (GUI) at http://variationontology.org/VariOtator.php for interactive submissions. For variation type annotations, there are three batch versions at http://variationontology.org/VariOtatorBatch.php, one with a Web interface that requires as input a tab-delimited file and two stand-alone versions that use as input an LOVD download file. These stand-alone versions (Linux and Windows) can be downloaded and installed locally. The Web service is available at http://variationontology.org/VariOService/?wsdl.

The core of the VariOtator is a Python (2.7) script with the RDFLib package (4.2.0), in combination with vario.owl and eco.owl, the Web Ontology Language (OWL 2) versions of VariO and ECO, respectively (see Fig. 1). The most recent version of vario.owl can be downloaded from http://variationontology.org/download.shtml, (at the moment version 1.04), the latest eco.owl file (release 2015-01-12) was downloaded from the Evidence Ontology Website (http://evidenceontology.googlecode.com/svn/trunk/eco.owl).

The VariOtator Web interface (http://variationontology.org/VariOtator.php) was developed using PHP5 and JavaScript (Fig. 1). Checking variant descriptions with Mutalyzer is done through their SOAP Web service (https://mutalyzer.nl/webservices). The VariOtator SOAP Web service makes use of the python-soaplib package (0.8.1), the client example uses the suds.client package (0.4).

The stand-alone batch versions (LOVD_VariOtator) for use with LOVD download files were developed with the cx-Freeze package (4.3.1) (http://cx-freeze.sourceforge.net). The Linux version is a ready-to-use package, the Windows version is available as a Windows Installer package (msi file). All the software is freely available and released and distributed under the terms of the GNU Affero General Public License version 3 (GPLv3).

Features

Web Interface

The user interface is organized according to the levels in VariO. First, the user chooses between annotation for variation type, function, structure, and property, and whether full lineage is desired or not (Fig. 2). If full lineage is chosen, all VariO terms down to the root term ("VariO:0001 variation") in the ontology are provided. The resulting VariO annotations are shown on screen, and can be downloaded as a text file. After each annotation step, the user can choose either to restart or to continue, in which case the new annotations are added to the previous one(s). The annotation can also be downloaded in a text file.

Variation Type Annotation

When choosing for variation type, the user can enter a variant description in HGVS format (Fig. 2). If a reference sequence is provided with a coding DNA description, the variant description will be checked with Mutalyzer and predicted RNA and protein descriptions and their annotations are provided, as well. VariOtator accepts the variation details in several formats including coding DNA, RNA, and protein descriptions, with or without the reference sequence. Both one- and three-letter amino acid codes are accepted. Examples of input are LRG_1:c.72G>A, NG_008680.1(PAX2):c.412A>G, NM_003990.3:c.412A>G, chr15:g.40702997G>A, r.(21a>u), and p.Trp26Cys. An example of VariOtator output with full lineage for variation type annotation can be found in Figure 3.

Since the terms "VariO:0313 transition," "VariO:0314 pyrimidine transition," "VariO:0315 purine transition," and "VariO:0316 transversion" can be used for annotation of both DNA and RNA substitutions, ancestor terms are included in the annotation also when full lineage is not chosen. So, if the resulting VariO terms are either "VariO:0315 purine transition" or "VariO:0314 pyrimidine transition," the ancestor terms "VariO:0136 DNA substitution" and "VariO:0313 transition" are added in the case of DNA or "VariO:0312 RNA substitution" and "VariO:0313 transition" are added in the case of RNA. Similarly, when the final VariO term is

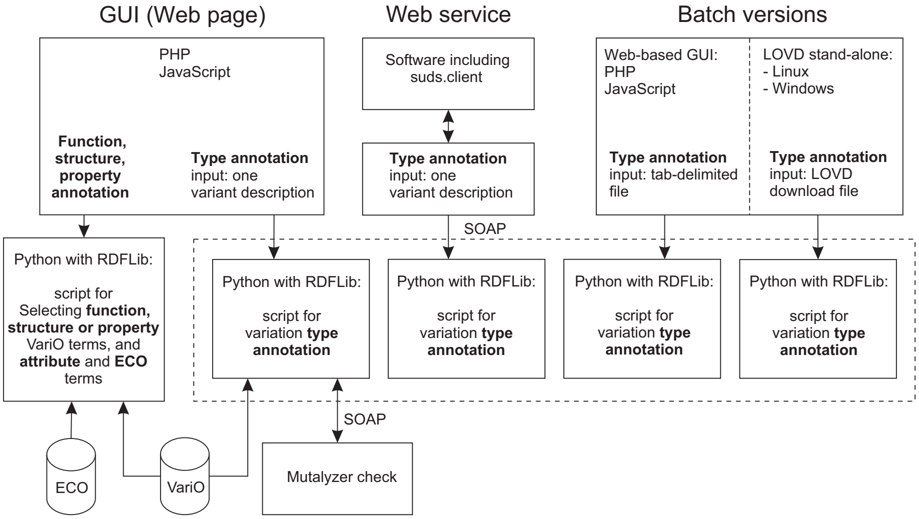

GUI (Web page)
Web service
Batch versions
PHP
JavaScript
Function, structure, property annotation
Type annotation input: one variant description
Software including suds.client
Type annotation input: one variant description
Web-based GUI: PHP JavaScript
LOVD stand-alone: - Linux - Windows
Type annotation input: tab-delimited file
Type annotation input: LOVD download file
SOAP
Python with RDFLib: script for Selecting function, structure or property VariO terms, and attribute and ECO terms
Python with RDFLib: script for variation type annotation
Python with RDFLib: script for variation type annotation
Python with RDFLib: script for variation type annotation
Python with RDFLib: script for variation type annotation
SOAP
ECO
VariO
Mutalyzer check

Figure 1. Overview of the VariOtator implementations. The scripts in the dotted box all have the same functionality, but are optimized for different purposes (mainly regarding to I/O). ECO: Evidence Ontology, format eco.owl; VariO: Variation Ontology, format vario.owl; GUI: Graphical User Interface; LOVD: Leiden Open Variation Database; SOAP: computer network messaging protocol; RDFLib: Python library for working with the Resource Description Framework (RDF).

"VariO:0316 transversion," either "VariO:0136 DNA substitution" or "VariO:0312 RNA substitution" is added.

Annotation of Variation Affecting Function

The first step for annotation of variations affecting function is to choose the molecular level (DNA, RNA, or protein) (Fig. 2). Then, an overview of VariO terms on the specific level is displayed. The relevant term is chosen by clicking it after which the annotation is generated. If necessary, more than one term can be chosen. An example of function annotation can be found in Figure 3.

Annotation of Variation Affecting Structure

As with functional annotation, the user first chooses the molecular levels (Fig. 2). If the selected term has sublevels, they are shown. This way the user can make very detailed annotations. Once the structure terms are chosen, it becomes possible to pick an attribute term, to modify and specify the annotation. For example, the quantity of the structure terms can be specified by using quantity change attributes including those for increased, decreased, missing, and not changed. In the next phase, an ECO term can be added to annotate the (experimental) evidence and method based on which the annotation is made. An example of structure annotation can be found in Figure 3.

Annotation of Variation Affecting Property

Variation properties are annotated similar to structural variations, including the use of attribute terms and ECO annotations. Both structure and property terms are specific for the molecular levels. The protein level allows for the largest number of choices due to the very wide spectrum of effects. An example of property annotation with attribute and ECO terms can be found in Figure 3.

Batch Versions

As variation type annotations can be automated and there are numerous databases containing very large numbers of variants, effective tools are needed for their annotation. For this purpose, we developed batch versions (http://variationontology.org/VariOtatorBatch.php) (see Fig. 1). The Web-based batch version requires a tab-delimited file with variant descriptions (coding DNA and optionally protein variant descriptions) as input. The results are provided in a tab-delimited file containing the original variant descriptions and predicted RNA variant descriptions, and the VariO terms for the variation type annotations at the DNA, RNA, and optionally at protein levels. The VariO terms are given including the full lineage up to but not including the root term "VariO:0001 variation." In contrast to the Web interface, the terms are not checked with Mutalyzer. There is a batch version for that purpose.

For annotating databases using the LOVD management system, stand-alone batch versions are available, both for Linux and Windows. These can be downloaded from http://variationontology.org/VariOtatorBatch.php. These versions are for variation type annotation only, and the variant descriptions are not checked with Mutalyzer. The user has to add columns for VariO annotations at the three molecular levels, DNA, RNA, and protein, to the LOVD database prior to using the VariOtator tool, as the tool uses an LOVD download file as input and adds the VariO terms to the relevant columns in that file. As how to add columns to the database,

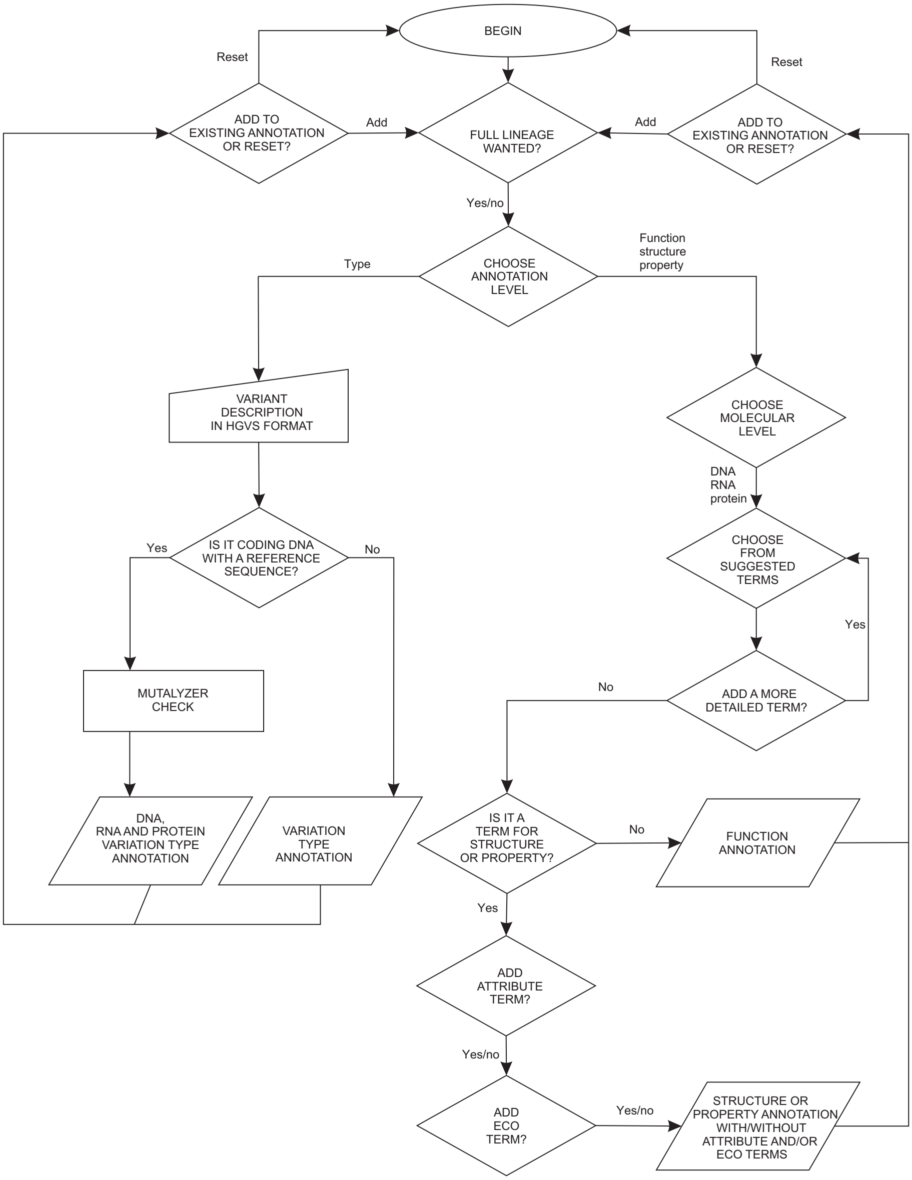
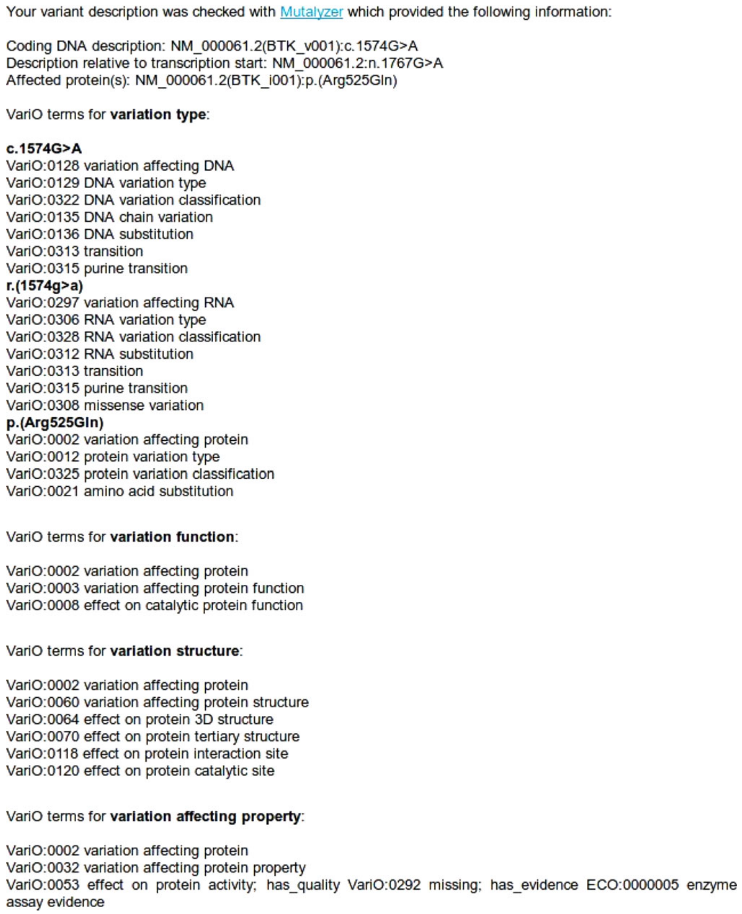

BEGIN
Reset
Reset
ADD TO EXISTING ANNOTATION OR RESET?
Add
FULL LINEAGE WANTED?
Add
ADD TO EXISTING ANNOTATION OR RESET?
Yes/no
CHOOSE ANNOTATION LEVEL
Type
Function structure property
VARIANT DESCRIPTION IN HGVS FORMAT
CHOOSE MOLECULAR LEVEL
DNA RNA protein
Yes
IS IT CODING DNA WITH A REFERENCE SEQUENCE?
No
CHOOSE FROM SUGGESTED TERMS
Yes
MUTALYZER CHECK
No
ADD A MORE DETAILED TERM?
DNA, RNA AND PROTEIN VARIATION TYPE ANNOTATION
VARIATION TYPE ANNOTATION
IS IT A TERM FOR STRUCTURE OR PROPERTY?
No
FUNCTION ANNOTATION
Yes
ADD ATTRIBUTE TERM?
Yes/no
ADD ECO TERM?
Yes/no
STRUCTURE OR PROPERTY ANNOTATION WITH/WITHOUT ATTRIBUTE AND/OR ECO TERMS

Figure 2. Flowchart of the annotation process with VariOtator.



Your variant description was checked with Mutalyzer which provided the following information:

Coding DNA description: NM_000061.2(BTK_v001):c.1574G>A
Description relative to transcription start: NM_000061.2:n.1767G>A
Affected protein(s): NM_000061.2(BTK_i001):p.(Arg525Gln)

VariO terms for **variation type**:

c.1574G>A
VariO:0128 variation affecting DNA
VariO:0129 DNA variation type
VariO:0322 DNA variation classification
VariO:0135 DNA chain variation
VariO:0136 DNA substitution
VariO:0313 transition
VariO:0315 purine transition
r.(1574g>a)
VariO:0297 variation affecting RNA
VariO:0306 RNA variation type
VariO:0328 RNA variation classification
VariO:0312 RNA substitution
VariO:0313 transition
VariO:0315 purine transition
VariO:0308 missense variation
p.(Arg525Gln)
VariO:0002 variation affecting protein
VariO:0012 protein variation type
VariO:0325 protein variation classification
VariO:0021 amino acid substitution

VariO terms for **variation function**:

VariO:0002 variation affecting protein
VariO:0003 variation affecting protein function
VariO:0008 effect on catalytic protein function

VariO terms for **variation structure**:

VariO:0002 variation affecting protein
VariO:0060 variation affecting protein structure
VariO:0064 effect on protein 3D structure
VariO:0070 effect on protein tertiary structure
VariO:0118 effect on protein interaction site
VariO:0120 effect on protein catalytic site

VariO terms for **variation affecting property**:

VariO:0002 variation affecting protein
VariO:0032 variation affecting protein property
VariO:0053 effect on protein activity; has_quality VariO:0292 missing; has_evidence ECO:0000005 enzyme assay evidence

Figure 3. VariOtator Web interface output. Variation type annotation with full lineage using NM_000061.2:c.1574G>A as input, and variation function, structure, and property annotation on protein level, with full lineage and attribute and ECO terms.

we refer to the LOVD documentation and VariOtator help files. Annotations are automatically added to the database when uploading the LOVD download file. If for some reason the annotation cannot be made an error log is provided, so that the users can solve the problematic cases.

Web Service

For programmatic access to the VariO annotation type tool, a SOAP Web service was developed, with which VariO annotation generation can be fully integrated into other software. An example client script for how to use this Web service can be found at http://variationontology.org/VariO-client-suds.py. It is a Python script that takes one variant description at a time, and generates the VariO annotations for that description, including the full lineage up to the root term (not included). A Web Service Definition Language (WSDL) description is available at http://variationontology.org/VariOService/?wsdl for easy generation of client programs in many languages. Again, the variant descriptions are not checked with Mutalyzer, this can be done with their Web services, if desired. Because of this, predicted RNA and protein descriptions are not provided when entering a coding DNA variant description, as is the case with the web interface.

Discussion

The VariOtator tool was developed to guarantee the consistency of annotations with VariO terms and to help with the task of variation annotation. By using the fully automated scripts, variation type annotation systematics in databases can be significantly improved. The specific batch versions for use with LOVD databases allows addition of variation type annotations to be included to the resources in the most widely used LSDB environment. For the three other annotation levels, function, structure, and property, the tool provides an easy-to-use and user-friendly interface. A guide with detailed instructions for annotators is available [Vihinen, 2014b] and examples of annotation with VariO can be found in [Vihinen, 2015b] or at http://www.variationontology.org/Examples.shtml.

An example of the use of VariO in an LSDB can be found at http://databases.lovd.nl/shared/genes/BTK. BTKbase [Väliaho et al., 2006; Schaafsma and Vihinen, 2015] is a database for Bruton agammaglobulinemia tyrosine kinase (BTK) variants causing X-linked agammaglobulinemia, a rare primary immunodeficiency [Tsukada et al., 1993; Vetrie et al., 1993]. BTKbase has been previously maintained with MUTbase software [Riikonen and Vihinen, 1999; Väliaho et al., 2006]. As part of the conversion to the LOVD database management system, some novel features such as VariO annotations were included. In the LOVD installation, the VariO annotations can be found in the columns "VariO Annotation DNA level," "VariO Annotation RNA level," and "VariO Annotation protein level." With the generated batch tools, variation type annotations can be automatically added to all the other LOVD-based LSDBs. It is up to the database curators to make these annotations as VariOtator is not integrated with LOVD. Once the annotations are added, it will become possible to make new kinds of analyses over known variation space. The other types of annotations are so variable and complex that they cannot be automated. VariO annotations are currently being added to the remaining 130 IDbases [Piirila et al., 2006] and are already available in NDDVD, NeuroDegenerative Diseases Variation Database (http://bioinf.suda.edu.cn/NDDvarbase/LOVDv.3.0/genes) containing information for 126 genes in 49 diseases [Yang et al., in preparation].

Acknowledgment

Disclosure statement: The authors declare no conflict of interest.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, et al. 2000. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.

Celli J, Dalgleish R, Vihinen M, Taschner PE, den Dunnen JT. 2012. Curating gene variant databases (LSDBs): toward a universal standard. Hum Mutat 33:291–297.

Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M. 2014. Standardized description of scientific evidence using the Evidence Ontology (ECO). Database (Oxford) 2014:bau075.

Cotton RG, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AF, Carrera P, Cox DW, Gottlieb B, Greenblatt MS, Hilbert P, Lehväslaiho H, et al. 2008. Recommendations for locus-specific databases and their curation. Hum Mutat 29:2–5.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, et al. 2015. Ensembl 2015. Nucleic Acids Res 43:D662–D669.

Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW, Beroud C, Dobson G, et al. 2010. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med 2:24.

den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15:7–12.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6:R44.

Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. 2011. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32:557–563.

Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. 2013. Genenames.org: the HGNC resources in 2013. Nucleic Acids Res 41:D545–D552.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42:D980–D985.

MacArthur JA, Morales J, Tully RE, Astashyn A, Gil L, Bruford EA, Larsson P, Flicek P, Dalgleish R, Maglott DR, Cunningham F. 2014. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. Nucleic Acids Res 42:D873–D878.

Piirilä H, Väliaho J, Vihinen M. 2006. Immunodeficiency mutation databases (IDbases). Hum Mutat 27:1200–1208.

Riikonen P, Vihinen M. 1999. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 15:852–859.

Schaafsma GCP, Vihinen M. 2015. Genetic variation in Bruton tyrosine kinase. In: Plebani A, Lougaris V, editors. *Agammaglobulinemia*. Switzerland: Springer International Publishing. p 75–85.

Tsukada S, Saffran DC, Rawlings DJ, Parolini O, Allen RC, Klisak I, Sparkes RS, Kubagawa H, Mohandas T, Quan S, Belmont JW, Cooper MD, et al. 1993. Deficient expression of a B cell cytoplasmic tyrosine kinase in human X-linked agammaglobulinemia. Cell 72:279–290.

UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res 43:D204–D212.

Väliaho J, Smith CIE, Vihinen M. 2006. BTKbase: the mutation database for X-linked agammaglobulinemia. Hum Mutat 27:1209–1217.

Vetrie D, Vořechovský I, Sideras P, Holland J, Davies A, Flinter F, Hammarström L, Kinnon C, Levinsky R, Bobrow M, Smith CIE, Bentley DR. 1993. The gene involved in X-linked agammaglobulinemia is a member of the src family of protein-tyrosine kinases. Nature 361:226–233.

Vihinen M. 2014a. Variation Ontology for annotation of variation effects and mechanisms. Genome Res 24:356–364.

Vihinen M. 2014b. Variation Ontology: annotator guide. J Biomed Semantics 5:9.

Vihinen M. 2015a. Muddled genetic terms miss and mess the message. Trends Genet 31:423–425.

Vihinen M. 2015b. Types and effects of protein variations. Hum Genet 134:405–421.

Vihinen M, den Dunnen JT, Dalgleish R, Cotton RG. 2012. Guidelines for establishing locus specific databases. Hum Mutat 33:298–305.

Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29:6–13.

Paper II

Genetic Variation in Bruton Tyrosine Kinase

5

Gerard C.P. Schaafsma and Mauno Vihinen

5.1 Introduction

Bruton agammaglobulinemia tyrosine kinase (*BTK*) variations lead to X-linked agammaglobulinemia (XLA, MIM# 300300), a hereditary primary immunodeficiency [26, 29]. XLA is caused by a block in B cell differentiation resulting in severely decreased numbers of B lymphocytes and an almost complete lack of plasma cells and very low or missing immunoglobulin levels of all isotypes. The patients have increased susceptibility to mainly bacterial infections because of virtually absent humoral immune responses. The frequency of XLA has been estimated to be 1:200,000 live births. The disease is considered to have full penetrance. Female carriers are healthy but display nonrandom X-chromosome inactivation in their B cells. Only a few female patients have been identified.

The *BTK* gene (LRG_128, reference sequence used U78027.1) contains 19 exons (Fig. 5.1) and codes for a protein of 77 kDa. Exon 1 is outside the coding region. BTK is expressed in all hematopoietic lineages except for T lymphocytes and plasma cells [23]. BTK belongs to the Tec family of related cytoplasmic protein tyrosine kinases (PTKs) formed by BMX (BMX non-receptor tyrosine kinase), ITK (IL2-inducible T cell kinase), TEC (tec protein tyrosine kinase), and TXK (TXK tyrosine kinase). Except for TXK, they have the same domain organization, from the N-terminus pleckstrin homology (PH) domain, Tec homology (TH) domain, Src homology 3 (SH3) domain, SH2 domain, and catalytic tyrosine kinase (TK) domain.

The three-dimensional structure has been determined for the PH domain and the first half of the TH domain [7], the SH3 domain [4], the SH2 domain [6], and the kinase domain [12]. For the full-length BTK, there is a low-resolution structure in an extended conformation [13]. BTK interacts with several partners [15].

G.C.P. Schaafsma • M. Vihinen (✉)
Department of Experimental Medical Science, Lund University,
Lund SE-22184, Sweden
e-mail: mauno.vihinen@med.lu.se

A. Plebani, V. Lougaris (eds.), *Agammaglobulinemia*,
Rare Diseases of the Immune System 4, DOI 10.1007/978-3-319-22714-6_5

Variations in BTK account for about 80 % of agammaglobulinemia cases. Several other genes can lead to a failure of B cell development and agammaglobulinemia [1]. These genes encode components of the pre-B cell receptor or proteins that are activated by cross-linking of the pre-B cell receptor. Defects in these genes lead to a block in B cell differentiation at the pro-B to pre-B cell transition. Other forms of agammaglobulinemia appear with growth hormone deficiency or as autosomal recessive diseases. Some autosomal recessive agammaglobulinemias have been identified involving pre-B cell receptor (pre-BCR) or BCR component genes μ-heavy chain (*IGHM*), λ5/14.1 (*IGLL1*, immunoglobulin lambda-like polypeptide 1), Igα (*CD79A*), and Igβ (*CD79B*). Variations in the B cell linker protein (*BLNK*), which is essential for Igμ signal transduction, and *PIK3R1* (phosphoinositide-3-kinase, regulatory subunit 1 (alpha)) for phosphoinositide 3-kinase regulator are downstream of BCR.

5.1.1 BTKbase

BTKbase is the first immunodeficiency variation database (IDbase) founded in 1994 [32]. Subsequently more than 130 immunodeficiency variation databases (IDbases) have been released [19]. BTKbase contains public variation entries for 1362 patients from 1198 unrelated families (total number of variants in these unrelated families is 1209) showing 742 unique molecular events.

BTKbase aims at collecting all published variations. Data are either directly submitted or derived from more than 100 publications. The database format has been previously published [31, 34]. The data are presented as individual entries, each carrying a unique patient identification number (PIN) and accession number, systematic names according to the Human Genome Variation Society (HGVS) variation nomenclature, a short verbal description of the variation, submission information (submission and update dates, version numbers, and submitter details), literature citations, and annotation in detail at DNA, RNA, and protein levels. In addition, the most important clinical parameters and laboratory findings are included, provided they are available.

IDbases, including BTKbase, follow a number of standards including the use of HUGO Gene Nomenclature Committee (HGNC) gene names (www.genenames.org), HGVS variation nomenclature [3], and IDRefSeqs (reference sequences for primary immunodeficiency genes and proteins). Currently, IDbases are in the process of changing to Locus Reference Genomic (LRG) reference sequences, which are already available for some 100 immunodeficiency genes (www.lrg-sequence.org). BTKbase follows the recommendations for locus-specific variation databases (LSDBs) [33] and their curation [2].

BTKbase is freely available at http://structure.bmc.lu.se/idbase/BTKbase/. The website contains information related to XLA and *BTK*. The bioinformatics pages include several tables for statistics of *BTK* variations. The variation distributions are shown along sequences in illustrative ways. The submission page provides variation checking facilities and electronic submission services. The variation browser allows visual means for browsing variations along the protein sequence. The reference information for variation publications and related protein structures are included in their own sections.

5.2 Analysis of BTK Variations

XLA arises as a block in B cell development. BTKbase contains information in many entries for the immunological status of patients. These properties have been extensively discussed in a previous publication [28]. The majority of the reported patients have significantly reduced numbers of B cells and Ig levels. A large portion of patients with X-linked diseases have de novo variations.

5.2.1 Variation Statistics

Extensive statistical analyses of variations at the three molecular levels, DNA, RNA, and protein, were performed. Since data per unique families are considered the most representative regarding, e.g., mutational effects and prevalence, the discussion about variation statistics mainly relates to these.

Variations appear throughout the BTK domains as well as in exons and introns (Fig. 5.1, Table 5.1); however, the distribution is not even. Some exons contain more variations than expected. The PH and SH2 domains contain approximately the expected number of variations, whereas there are less than expected in the TH and SH3 domains and more than expected in the kinase domain (Table 5.2). The TH domain has two structural elements [31, 35], an N-terminal BTK motif and a C-terminal proline-rich region which contains two proline-rich regions capable of intra- and intermolecular interactions [4, 17]. The reason for under-representativeness of the TH domain may be that it likely has a partially intrinsically disordered structure in the C-terminal half of the domain, and therefore, variations do not have a major effect. On the other hand, XLA-causing variants do appear in the Zn^{2+}-binding BTK motif.

We have recently investigated the putative effects of all possible amino acid substitutions due to single nucleotide changes in the BTK TK domain [27]. Altogether 67 % of the 1495 substitutions were predicted to be harmful. Although this number

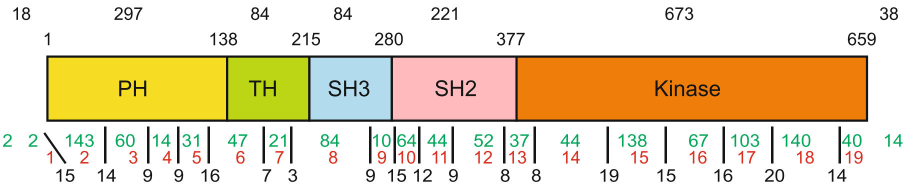


Fig. 5.1 Distribution of all variations to *BTK* gene regions and BTK protein domains. Variations in exons are indicated by green numbers in exons which are numbered in red. Variations in introns are in black below the domain chart. Domain borders are above the chart and numbers of variations in the domains above them

Table 5.1 Distribution of variation and variation types in BTK domains for all cases, independent families, and unique variations

Domain	Upstream			PH (414)			TH (231)			SH3 (194)			SH2 (292)			TK (846)			Others			Total			% of total		
Type	All	Fam	Uniq	All	Fam	Uniq	All	Fam	Uniq	All	Fam	Uniq	All	Fam	Uniq	All	Fam	Uniq	All	Fam	Uniq	All	Fam	Uniq	All	Fam	Uniq
Missense	0	0	0	119	100	48	11	10	10	0	0	0	112	93	42	328	290	147	0	0	0	570	493	247	41.5	40.8	33.3
Nonsense	0	0	0	44	40	19	16	14	9	46	39	11	23	22	12	97	85	43	0	0	0	226	200	94	16.4	16.5	12.7
Deletion; in-frame	1	1	1	11	6	6	2	2	2	0	0	0	5	5	4	21	18	11	0	0	0	40	32	24	2.9	2.6	3.2
Deletion; frameshift	0	0	0	3	2	1	0	0	0	0	0	0	1	1	1	2	1	1	4	3	3	10	7	6	0.7	0.6	0.8
Gross deletion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	14	14	14	14	14	1.0	1.2	1.9
Insertion; in-frame	0	0	0	3	3	2	0	0	0	0	0	0	0	0	0	3	3	3	0	0	0	6	6	5	0.4	0.5	0.7
Insertion; frameshift	0	0	0	3	2	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	4	3	1	0.3	0.2	0.1
Indel	0	0	0	3	3	3	3	3	3	2	2	2	1	1	1	6	6	5	0	0	0	15	15	14	1.1	1.2	1.9
Intron	11	10	6	27	23	14	9	8	5	6	5	3	29	28	20	65	59	35	6	6	6	153	139	89	11.1	11.5	12.0
Intron; in-frame	0	0	0	3	3	2	0	0	0	1	1	1	6	4	3	0	0	0	1	1	1	11	9	7	0.8	0.7	0.9
Intron; out-of-frame	0	0	0	7	6	5	1	1	1	1	1	1	6	3	3	15	12	10	2	2	2	32	25	22	2.3	2.1	3.0
Frameshift	0	0	0	63	56	41	37	31	26	23	22	18	37	34	30	86	81	66	2	1	1	248	225	182	18.0	18.6	24.5
Multiple variant	0	0	0	7	6	4	5	5	5	2	2	2	0	0	0	6	4	2	0	0	0	20	17	13	1.5	1.4	1.8
Complex	2	1	1	3	2	2	0	0	0	2	2	2	1	1	1	2	2	2	1	1	1	11	9	9	0.8	0.7	1.2
Unknown	4	4	4	1	1	1	0	0	0	0	0	0	0	0	0	2	2	2	8	8	8	15	15	15	1.1	1.2	2.0
Total	18	16	12	297	253	149	84	74	61	84	75	40	221	192	117	633	563	327	38	36	36	1375	1209	742	100	100	100

Between brackets length of the domain in terms of cDNA nucleotides, *all* all analyzed alleles, *fam* unrelated families, *uniq* unique molecular events in DNA sequence

Table 5.2 Spectrum of variants in the structural BTK domains

Domain	Length	Length/total length	Normalized expected	Observed	χ^2	P^a
PH	414	0.209	242	253	0.47384	0.49122
TH	231	0.117	135	74	27.69467	$<10^{-6}$***
SH3	194	0.098	114	75	13.07900	0.00030***
SH2	292	0.148	171	192	2.60845	0.10630
TK	846	0.428	495	563	9.31070	0.00228**
Total	1977	1	1157	2257		

[a] Significance levels: ** $p<0.01$ *** $p<0.001$

seems very high, it is considered to be realistic because the kinase domain contains so many conserved regions and has several functions. The situation is likely very different in the SH3 and TH domains.

The variants are classified in Table 5.1 based on their effects on DNA or RNA level. The largest group of the variants is amino acid substitution causing missense variations (41 % of independent families). The SH3 domain is the only one where amino acid substitutions do not occur. Although SH3 domains are abundant in the human proteome, no disease-causing amino acid substitutions have been reported in any of them.

Nonsense variations account for 17 % of all variations, frameshift variations 19 %, and intronic variations 14 %. The proportions of deletions (4 %) and insertions (0.7 %) are very low and different from those reported in previous publications [10, 28] where proportions of 20 % (deletions) and 7 % (insertions) were given. These differences are due to the way the variants were counted, e.g., a variant with a DNA name "deletion" and an RNA name "frameshift" has been considered here as a frameshift variation. In the future we will avoid this kind of issues by adopting variation naming according to the Variation Ontology [30].

The distribution of variation types is very similar compared to the other IDbases [19]. The ratio of missense/nonsense variations, 2.5, is slightly higher in BTKbase compared to IDbases (1.5). Multiple variants in *BTK* have been identified in 17 families, complex variations in 9 families, and miscellaneous cases in 15 families.

There are altogether 341 unique amino acid substitutions. The theoretical maximum is 4151: thus, until now we have 8.2 % of the total variation; however, just a fraction of them is harmful and thus identifiable from XLA patients. In the case of nonsense variations, a larger portion has been seen in patients. There are 94 (28 %) of all the possible ($n=297$) variants in the BTKbase. According to χ^2 statistics, there is highly significant overrepresentation ($p<0.0001$).

When we are looking at the changes at amino acid level, it is apparent that arginine, as previously indicated, harbors the largest number of variants (Table 5.3). However, the most common outcome at protein level is protein truncation due to incorporation of a stop codon to the coding region. Altogether 29.5 % of single nucleotide changes lead to protein truncation.

Table 5.3 Amino acid substitutions indicated in percentages

									Hydrophilic													
	Hydrophobic								Acidic		Basic			Polar				Special				
->	A	F	I	L	M	V	W	Y	D	E	H	K	R	N	Q	S	T	C	G	P	X	Total
A	0					1			0.6	0.4						0	0		0	0.8	0	2.8
F		0	0	0.4		0.1		0.3								1		0.1			0.1	2.1
I		0.1	0	0	0.1	0						0	0	1	0	0.1	0.6				0	2
L		1.3	0.3	0	0	0.3	0.1				0		0.6		0.1	0.6				3.8	0.6	7.6
M			1.1	0.1	0	0.4						0.4	0.1				2				0	4.2
V	0.4	0.7	0	0	0	0.1			0.4	0.1									0.3	0	0.1	2.2
W				0.1									0.8			0.3		0.3	0		4.1	5.6
Y		0						0	0.7		0.6			0.6		0.8		1.3			4.8	8.7
D	0					0.4		0.1	0	0.1	0.1			0.3					0.3		0	1.4
E	0					0			0.7	0		0.3			0				0.4		1.8	3.2
H				0				0.1	0.1		0		0.4	0	0.1					0.3	0	1.1
K			0.1		0					1.1		0	0.4	0.3	0		0				1.4	3.4
R			0	0.3	0		6.3				4.1	0.6	0.1		4.5	1	0.3	2.5	2	1.1	8.3	31
N			0					0.1	0		0	0.1		0		0	0				0.1	0.4
Q			0	0						0	0.3	0	0.1		0					0.1	6.2	6.7
S	0	0.7	0.1	0.1			0	0.4					0.1	0		0	0	0	0	0.7	1	3.2
T	0.1		0.4		0							0	0	0		0	0			1.1	0	1.7
C		0.7					0.3	1.7					0.4			0.3		0	0.4		0.7	4.5
G	0.1					0.1	0.1		1.1	2.1			1.5			0		0	0.1		0.3	5.6
P	0.3			0.6		0					0		0.3		0	0.7	0.7			0	0	2.5
Total	1	3.5	2.1	1.7	0.1	2.5	6.9	2.8	3.6	3.9	5	1.4	5	2.1	4.8	4.8	3.5	4.2	3.5	8	29.5	100

Arginine is by far the most frequently substituted amino acid (31 %). This has not only been observed in BTK before [28] but also in variant datasets extracted from dbSNP [21], and this overrepresentation of arginine is known to be due to the high mutability of the codons containing CpG dinucleotides. Arginine is coded by six codons, four of which have a CpG dinucleotide in the first and second codon position [18]. The overrepresentation of arginine as the most frequently substituted amino acid also leads to the enrichment of tryptophan as the residue other amino acids are substituted to; arginine was replaced by tryptophan in 6.3 % of all amino acid substitutions (Table 5.3).

Proline is the amino acid to which most amino acids have been substituted to (8 %) closely followed by tryptophan (6.9 %), histidine (5 %), arginine (5 %), glutamine (4.8 %), and serine (4.8 %).

The G>A and C>T substitutions form the largest classes of changes, ~24 % (Table 5.4). The types of base changes were investigated more closely. The changes from amino to keto base and vice versa are much more frequent than substitutions within these groups. There is clearly a higher frequency of transitions (purine to purine and pyrimidine to pyrimidine, 66 %) than transversions (34 %). The higher rate of transitions agrees with the higher rate (~70 %) of transitions found to be typical for human genes [25]. The strong to weak base substitutions are by far the biggest category, containing 60 % of the variations. This was also found in the VariSNP variant datasets [21].

Table 5.4 Nucleotide substitutions in unique families (%)

All

→	a	c	g	t	Total
a	0	3.5	8.3	2.8	14.6
c	5	0	2.1	23.5	30.7
g	23.7	5.5	0	7.8	36.9
t	3.6	10	4.2	0	17.8
Total	32.3	19	14.6	34.1	100

Summarized into amino and keto categories

→	Amino	Keto	Total
Amino	8.5	36.7	45.2
Keto	42.8	12	54.8
Total	51.3	48.7	100

Summarized into weak and strong categories

→	Weak	Strong	Total
Weak	6.4	26	32.4
Strong	60	7.6	67.6
Total	66.4	33.6	100

Summarized into purine and pyrimidine categories

→	Purines	Pyrimidines	Total
Purines	31.9	19.6	51.5
Pyrimidines	14.9	33.6	48.5
Total	46.9	53.1	100

5.2.2 Structural Consequences

BTK consists of five domains which, except for the SH3 domain, contain amino acid substitutions (Fig. 5.2). The effects and consequences of the variations vary widely. A recent study revealed that about two thirds of all kinase domain variations originating from a single nucleotide change likely lead to XLA [27]. This is not to say that two thirds of all possible amino acid changes were harmful since the majority of them do not originate from single base changes (because of the organization of the genetic code). Numerous variants affect functional sites, such as ligand- and substrate-binding regions at the domains. Stability-affecting changes are common. There are putative explanations available for the consequences of all the 1495 substitutions studied. These results are well in line with previous studies and predictions of BTK variants [5, 7–9, 11, 12, 14, 16, 20, 22, 24, 28, 32, 35–41].

Minor changes can be accommodated without major structural alterations. As has been seen in especially the PH domain, changes to electrostatics are common [16]. When the charge is reversed, added, or removed, the properties of the site are modified. If this happens on the protein surface of the binding site, then the interactions with partners are impaired or weakened.

Structural variations appear frequently in secondary structural elements. Although there are some variations at loops connecting these elements, the α- and

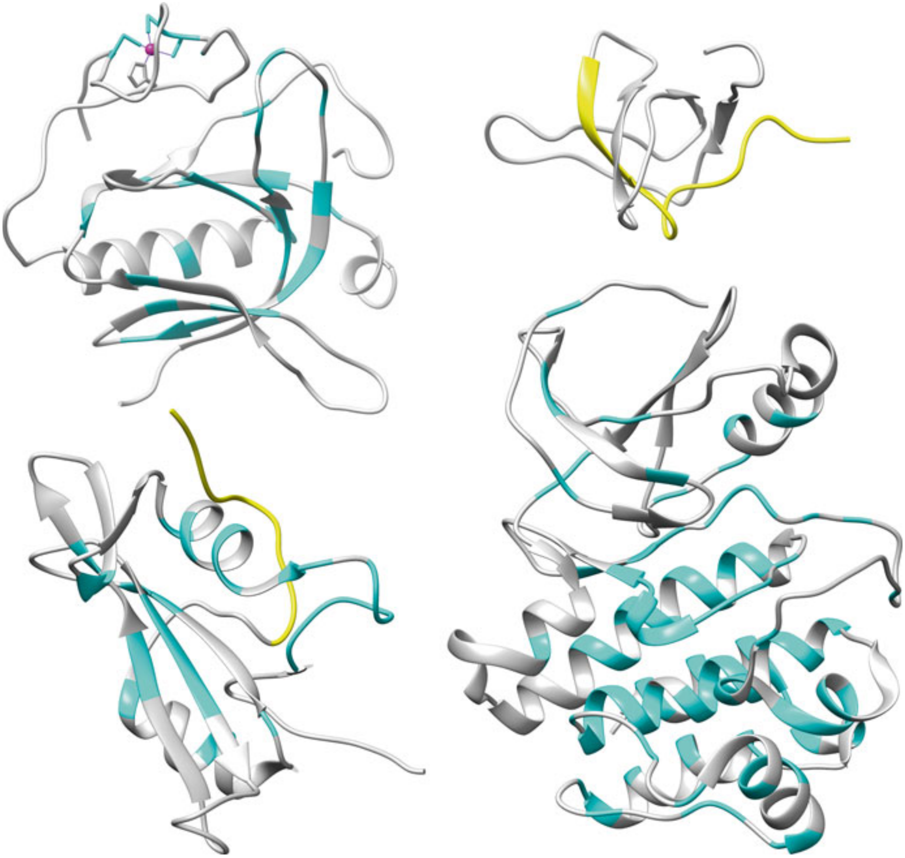


Fig. 5.2 Distribution of amino acid substitutions to BTK domains. Affected amino acids are shown in yellow. PH domain is on top left (PDB code 1BTK [7]). The first part of the TH domain including the BTK motif binding Zn^{2+} (magenta) is on the top of the domain. In the SH3 (1AWX [4]), top right, and SH2 domain (2GE9 [6]), bottom left, an in-frame deletion of 21 residues is indicated. The kinase domain (1K2P [12]) is at *bottom to the right*. Amino acid substitutions appear throughout all the domains except the SH3 domain where there are none

β-structures are more sensitive for substitutions. Structural variants are frequent on the protein core where there is no space for larger side chains due to tight packing. Further, introduction of charged or polar residues to the protein core, even if sterically possible, is usually harmful. Much more variation is allowed on the protein surface in areas not involved in intra- or intermolecular interactions. Some of these interactions are known; however, we do not even know the three-dimensional organization of the entire BTK. The domains are independently folding and connected by loops, which can be quite long. It is likely that the domain interactions are different in different folds of the entire protein. There is structural information for the entire BTK in elongated conformation [13]; however, this conformation is not likely the only one.

BTK variation information has been collected already for two decades into BTKbase, which has been a central resource for research and diagnosis. The database is constantly growing; however, the recent explosion in sequencing activities has not contributed much to the increased numbers in the database. That is presumably because many cases remain in laboratories and are never published or submitted to a database. It is in the interest of the entire community to share information about variations, especially in rare diseases such as XLA.

Acknowledgments Financial support from Lund University, Medical Faculty, and Swedish Research Council Vetenskapsrådet is gratefully acknowledged.

References

1. Berglöf A, Turunen JJ, Gissberg O, Bestas B, Blomberg KE, Smith CI (2013) Agammaglobulinemia: causative mutations and their implications for novel therapies. Expert Rev Clin Immunol 9:1205–1221
2. Celli J, Dalgleish R, Vihinen M, Taschner PE, den Dunnen JT (2012) Curating gene variant databases (LSDBs): toward a universal standard. Hum Mutat 33:291–297
3. den Dunnen JT, Antonarakis SE (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15:7–12
4. Hansson H, Mattsson PT, Allard P, Haapaniemi P, Vihinen M, Smith CIE, Härd T (1998) Solution structure of the SH3 domain from Bruton tyrosine kinase. Biochemistry 37:2912–2924
5. Holinski-Feder E, Weiss M, Brandau O, Jedele KB, Nore B, Bäckesjö CM, Vihinen M, Hubbard SR, Belohradsky BH, Smith CIE, Meindl A (1998) Mutation screening of the BTK gene in 56 families with X-linked agammaglobulinemia (XLA): 47 unique mutations without correlation to clinical course. Pediatrics 101:276–284
6. Huang KC, Cheng HT, Pai MT, Tzeng SR, Cheng JW (2006) Solution structure and phosphopeptide binding of the SH2 domain from the human Bruton tyrosine kinase. J Biomol NMR 36:73–78
7. Hyvönen M, Saraste M (1997) Structure of the PH domain and Btk motif from Bruton tyrosine kinase: molecular explanations for X-linked agammaglobulinaemia. Embo J 16:3396–3404
8. Jin H, Webster ADB, Vihinen M, Sideras P, Vořechovský I, Hammarström L, Bernatowska-Matuszkiewicz E, Smith CIE, Bobrow M, Vetrie D (1995) Identification of Btk mutations in 20 unrelated patients with X-linked agammaglobulinaemia (XLA). Hum Mol Genet 4:693–700
9. Korpi M, Väliaho J, Vihinen M (2000) Structure-function effects in primary immunodeficiencies. Scand J Immunol 52:226–232
10. Lindvall JM, Blomberg KE, Väliaho J, Vargas L, Heinonen JE, Berglöf A, Mohamed AJ, Nore BF, Vihinen M, Smith CIE (2005) Bruton tyrosine kinase: cell biology, sequence conservation, mutation spectrum, siRNA modifications, and expression profiling. Immunol Rev 203:200–215
11. Maniar HS, Vihinen M, Webster ADB, Nilsson L, Smith CIE (1995) Structural basis for X-linked agammaglobulinemia (XLA): mutations at interacting Btk residues R562, W563, and A582. Clin Immunol Immunopathol 76:S198–S202
12. Mao C, Zhou M, Uckun FM (2001) Crystal structure of Bruton tyrosine kinase domain suggests a novel pathway for activation and provides insights into the molecular basis of X-linked agammaglobulinemia. J Biol Chem 276:41435–41443
13. Márquez JA, Smith CIE, Petoukhov MV, Lo Surdo P, Mattsson PT, Knekt M, Westlund A, Scheffzek K, Saraste M, Svergun DI (2003) Conformation of full-length Bruton tyrosine kinase (Btk) from synchrotron X-ray solution scattering. Embo J 22:4616–4624

14. Mattsson PT, Lappalainen I, Bäckesjö CM, Brockmann E, Lauren S, Vihinen M, Smith CIE (2000) Six X-linked agammaglobulinemia-causing missense mutations in the Src homology 2 domain of Bruton tyrosine kinase: phosphotyrosine-binding and circular dichroism analysis. J Immunol 164:4170–4177
15. Mohamed AJ, Yu L, Bäckesjö CM, Vargas L, Faryal R, Aints A, Christensson B, Berglöf A, Vihinen M, Nore BF and others (2009) Bruton tyrosine kinase (Btk): function, regulation, and transformation with special emphasis on the PH domain. Immunol Rev 228:58–73
16. Okoh MP, Vihinen M (1999) Pleckstrin homology domains of tec family protein kinases. Biochem Biophys Res Commun 265:151–157
17. Okoh MP, Vihinen M (2002) Interaction between Btk TH and SH3 domain. Biopolymers 63:325–334
18. Ollila J, Lappalainen I, Vihinen M (1996) Sequence specificity in CpG mutation hotspots. FEBS Lett 396:119–122
19. Piirilä H, Väliaho J, Vihinen M (2006) Immunodeficiency mutation databases (IDbases). Hum Mutat 27:1200–1208
20. Saha BK, Curtis SK, Vogler LB, Vihinen M (1997) Molecular and structural characterization of five novel mutations in the Bruton tyrosine kinase gene from patients with X-linked agammaglobulinemia. Mol Med 3:477–485
21. Schaafsma GCP, Vihinen M (2015) VariSNP, a benchmark database for variations from dbSNP. Hum Mutat 36:161–166
22. Shen B, Vihinen M (2004) Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. Protein Eng Des Sel 17:267–276
23. Smith CIE, Baskin B, Humire-Greiff P, Zhou JN, Olsson PG, Maniar HS, Kjellén P, Lambris JD, Christensson B, Hammarström L, Bentley D, Vetrie D et al (1994) Expression of Bruton agammaglobulinemia tyrosine kinase gene, BTK, is selectively down-regulated in T lymphocytes and plasma cells. J Immun 152:557–565
24. Speletas M, Kanariou M, Kanakoudi-Tsakalidou F, Papadopoulou-Alataki E, Arvanitidis K, Pardali E, Constantopoulos A, Kartalis G, Vihinen M, Sideras P, Ritis K (2001) Analysis of Btk mutations in patients with X-linked agammaglobulinaemia (XLA) and determination of carrier status in normal female relatives: a nationwide study of Btk deficiency in Greece. Scand J Immunol 54:321–327
25. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH and others (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493
26. Tsukada S, Saffran DC, Rawlings DJ, Parolini O, Allen RC, Klisak I, Sparkes RS, Kubagawa H, Mohandas T, Quan S, Belmont JW, Cooper MD et al (1993) Deficient expression of a B cell cytoplasmic tyrosine kinase in human X-linked agammaglobulinemia. Cell 72:279–290
27. Väliaho J, Faisal I, Ortutay C, Smith CIE, Vihinen M (2015) Characterization of all possible single nucleotide change –caused amino acid substitutions in the kinase domain of Bruton tyrosine kinase. Hum Mutat 36:638–647
28. Väliaho J, Smith CIE, Vihinen M (2006) BTKbase: the mutation database for X-linked agammaglobulinemia. Hum Mutat 27:1209–1217
29. Vetrie D, Vořechovský I, Sideras P, Holland J, Davies A, Flinter F, Hammarström L, Kinnon C, Levinsky R, Bobrow M, Smith CIE, Bentley DR (1993) The gene involved in X-linked agammaglobulinaemia is a member of the src family of protein-tyrosine kinases. Nature 361:226–233
30. Vihinen M (2014) Variation Ontology for annotation of variation effects and mechanisms. Genome Res 24:356–364
31. Vihinen M, Belohradsky BH, Haire RN, Holinski-Feder E, Kwan SP, Lappalainen I, Lehväslaiho H, Lester T, Meindl A, Ochs HD, Ollila J, Vořechovský I et al (1997) BTKbase, mutation database for X-linked agammaglobulinemia (XLA). Nucleic Acids Res 25:166–171
32. Vihinen M, Cooper MD, de Saint Basile G, Fischer A, Good RA, Hendriks RW, Kinnon C, Kwan SP, Litman GW, Notarangelo LD, Ochs HD, Rosen FS et al (1995) BTKbase: a database of XLA-causing mutations. Immunol Today 16:460–465

33. Vihinen M, den Dunnen JT, Dalgleish R, Cotton RG (2012) Guidelines for establishing locus specific databases. Hum Mutat 33:298–305
34. Vihinen M, Lehväslaiho H, Cotton RDH (1999) Immunodeficiency mutation databases. In: Ochs HD, Smith CIE, Puck JM (eds) Primary immunodeficiency diseases. A molecular and genetic approach. Oxford University Press, New York/Oxford, pp 443–447
35. Vihinen M, Nilsson L, Smith CIE (1994) Structural basis of SH2 domain mutations in X-linked agammaglobulinemia. Biochem Biophys Res Commun 205:1270–1277
36. Vihinen M, Nilsson L, Smith CI (1994) Tec homology (TH) adjacent to the PH domain. FEBS Lett 350:263–265
37. Vihinen M, Nore BF, Mattsson PT, Backesjo CM, Nars M, Koutaniemi S, Watanabe C, Lester T, Jones A, Ochs HD, Smith CI (1997) Missense mutations affecting a conserved cysteine pair in the TH domain of Btk. FEBS Lett 413:205–210
38. Vihinen M, Vetrie D, Maniar HS, Ochs HD, Zhu Q, Vořechovský I, Webster ADB, Notarangelo LD, Nilsson L, Sowadski JM, Smith CIE (1994) Structural basis for chromosome X-linked agammaglobulinemia: a tyrosine kinase disease. Proc Natl Acad Sci U S A 91:12803–12807
39. Vihinen M, Zvelebil MJ, Zhu Q, Brooimans RA, Ochs HD, Zegers BJ, Nilsson L, Waterfield MD, Smith CIE (1995) Structural basis for pleckstrin homology domain mutations in X-linked agammaglobulinemia. Biochemistry 34:1475–1481
40. Vořechovský I, Vihinen M, de Saint Basile G, Honsova S, Hammarström L, Müller S, Nilsson L, Fischer A, Smith CIE (1995) DNA-based mutation analysis of Bruton tyrosine kinase gene in patients with X-linked agammaglobulinaemia. Hum Mol Genet 4:51–58
41. Zhu Q, Zhang M, Rawlings DJ, Vihinen M, Hagemann T, Saffran DC, Kwan SP, Nilsson L, Smith CIE, Witte ON, Chen S-H, Ochs HD (1994) Deletion within the Src homology domain 3 of Bruton tyrosine kinase resulting in X-linked agammaglobulinemia (XLA). J Exp Med 180:461–470

Paper III

DATABASES

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

VariSNP, A Benchmark Database for Variations From dbSNP

Gerard C.P. Schaafsma and Mauno Vihinen*

Protein Structure and Bioinformatics, Department of Experimental Medical Science, Lund University, Lund SE-221 84, Sweden

Communicated by Rachel Karchin

Received 1 September 2014; accepted revised manuscript 28 October 2014.
Published online 11 November 2014 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22727

ABSTRACT: For development and evaluation of methods for predicting the effects of variations, benchmark datasets are needed. Some previously developed datasets are available for this purpose, but newer and larger benchmark sets for benign variants have largely been missing. VariSNP datasets are selected from dbSNP. These subsets were filtered against disease-related variants in the ClinVar, UniProtKB/Swiss-Prot, and PhenCode databases, to identify neutral or nonpathogenic cases. All variant descriptions include mapping to reference sequences on chromosomal, genomic, coding DNA, and protein levels. The datasets will be updated with automated scripts on a regular basis and are freely available at http://structure.bmc.lu.se/VariSNP.
Hum Mutat 36:161–166, 2015. © 2014 Wiley Periodicals, Inc.

KEY WORDS: benchmark; dbSNP; genetic variation; mutation; variant effect analysis; variant effect prediction; variant position mapping

Introduction

The development of high-throughput sequencing technologies has caused the vast growth of genetic variation data. Interpretation of the effects of variations is often missing. Since experimental validation at a large scale is not feasible, computational methods for predicting the effects and consequences of variations have been developed. Approved and widely accepted benchmark datasets are needed to enable the systematic quantitative comparison of variant effect predictor performance. Benchmark datasets are standard representative datasets with known outcome and they are vital for measurement and judgment of predictor performance [Vihinen, 2012]. Criteria for variation benchmark sets have been discussed in detail [Nair and Vihinen, 2013]. Benchmark datasets can also be used for training and testing of novel predictors most often based on machine learning systems. The need for benchmarks has been expressed frequently [Thusberg et al., 2011; Johnston and Biesecker, 2013; Peterson et al., 2013].

The first systematic benchmark datasets for variation effects became available with the release of VariBench [Nair and Vihinen, 2013] and include datasets used in method performance assessment studies. For the data and datasets in VariBench, criteria for inclusion included relevance, representativeness, nonredundancy, experimentally verified cases, positive and negative cases, scalability, and reusability. While VariBench includes experimentally verified cases, the SNPdbe database [Schaefer et al., 2012] contains nonsynonymous single amino acid substitutions from >2,600 organisms also with predictions of computationally annotated functional impacts. Many new variants have been discovered since the release of VariBench. To keep up with the new developments, it became apparent that there is a need for newer and larger sets, which can also be easily updated. The Database for Short Genetic Variations (dbSNP) is a public domain archive for a broad collection of simple genetic variations [Sherry et al., 2001]. It is regarded as the largest variation database and was therefore used as the source of data for new benchmark datasets. Since dbSNP contains both disease-related variants (pathogenic, function affecting variants) and nondisease-related variants (variants not affecting function), neutral (nonpathogenic) subsets of the dbSNP database were generated by filtering out variants found in ClinVar, UniProtKB/Swiss-Prot, and PhenCode datasets, and which were annotated as either pathogenic or disease-causing in the source databases. These three resources are among the most comprehensive ones for disease-related variants. As dbSNP classifies variants into functional classes (Table 1), we were able to produce neutral benchmark datasets for all the populated classes, altogether 13 benchmarks.

ClinVar contains reports of relationships among medically important variants and phenotypes with supporting evidence [Landrum et al., 2014]. Submissions are accepted from different sources: clinical tests, research, and extracted from the literature. The content is highly structured and harmonized with controlled vocabularies and other data standards. Variations are mapped to mRNA, genomic, and protein reference sequences, according to the Human Genome Variation Society (HGVS) recommendations.

The Swiss-Prot section of the Universal Protein Knowledgebase (UniProtKB/Swiss-Prot) contains manually annotated records with information on protein sequences extracted from literature and curator-evaluated computational analysis [UniProt Consortium, 2014]. Manual curation includes a thorough review of available information on sequence variants (mostly single amino acid substitutions) and associated genetic disease information. There are almost 70,000 variations, 35% of which are associated with one of over 4,000 described genetic diseases [Famiglietti et al., 2014].

PhenCode (phenotypes for ENCODE) connects human phenotype and clinical data in various locus-specific databases (LSDBs) with data on genome sequences, evolutionary history, and function from the ENCODE project and other resources in the UCSC Genome Browser [Giardine et al., 2007]. It contains about 34,000

*Correspondence to: Mauno Vihinen, Protein Structure and Bioinformatics, Department of Experimental Medical Science, Lund University, BMC D10, Lund SE-221 84, Sweden. E-mail: mauno.vihinen@med.lu.se

Contract grant sponsors: Lund University; Vetenskapsrådet.

Table 1. NCBI Functional Classes

Functional class	Description	SO ID[a]
reference	Contig reference	–
missense	Alters codon to make an altered amino acid in protein product	0001583
cds-indel	Indel snp with length of multiple of 3 bp, not causing frameshift (inframe_variant SO: 0001650)	–
synonymous-codon	Sequence variant where there is no resulting change to the encoded amino acid	0001588
intron-variant	Transcript variant occurring within an intron	0001627
nc-transcript-variant	Transcript variant of a noncoding RNA gene	0001619
downstream-variant-500B	Sequence variant located within a half KB of the end of a gene	0001634
upstream-variant-2KB	Sequence variant located within 2 KB 5′ of a gene	0001636
stop-gained	Sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	0001587
stop-lost	Sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	0001578
frameshift-variant	Sequence variant that causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	0001589
utr-variant-3-prime	UTR variant of the 3′ UTR	0001624
upstream-variant-5KB	Sequence variant located within 5 KB 5′ of a gene	0001635
splice-acceptor-variant	Splice variant that changes the two base region at the 3′ end of an intron	0001574
splice-donor-variant	Splice variant that changes the two base pair region at the 5′ end of an intron	0001575

[a] SO ID: Sequence ontology ID; deprecated terms omitted. From ftp://ftp.ncbi.nih.gov/snp/specs/docsum_3.4.xsd

variants originating from LSDBs and about 57,000 variants from UniProtKB/Swiss-Prot.

All variant descriptions in VariSNP include mapping to reference sequences on chromosomal, genomic, coding DNA, and protein levels. Where possible, a mapping to PDB structure coordinates is included. The datasets are available on http://structure.bmc.lu.se/VariSNP, and will be updated on a regular basis.

Data Collection and Selection

The workflow for downloading, selecting, and filtering data from the different sources is depicted in Figure 1. The procedure is explained in more detail below.

dbSNP Data

Data were downloaded from the NCBI FTP-site (ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/XML), the latest update (August 5, 2013) was used. Each variant in dbSNP is assigned to one or more functional classes, which are described in Table 1. These classes are available in the docsum_3.4.xsd file at ftp://ftp.ncbi.nlm.nih.gov/snp/specs. Deprecated and not-implemented terms were omitted. A variation may belong to multiple functional classes. Multiplicity will result, for example, when a variation falls within an exon of one transcript and an intron of another for the same gene (http://www.ncbi.nlm.nih.gov/books/NBK21088/). Functional classifications can vary from genome build to build.

Selections of variations were made based on functional class and having values for "minor allele frequency" (MAF) and for "validation" (any type of validation, e.g., "by1000G" for variants in The 1000 Genomes Project). Disease-related variations were filtered out based on variants found in ClinVar, UniProtKB/Swiss-Prot, or PhenCode.

ClinVar Data

In ClinVar, reports about sets of assertions on the same variation/phenotype relationship are aggregated and given as a Reference ClinVar (RCV) accession. Because of this model, an allele appears in multiple RCV accessions whenever different phenotypes are reported for that allele. RCV records contain the information needed for our purpose.

The XML file (ClinVarFullRelease_2014–5.xml.gz) was downloaded from http://www.ncbi.nlm.nih.gov/clinvar/ (May 1, 2014). Selections were first made based on molecular consequence, variant type, and clinical significance. Values for molecular consequence describe the effects of sequence changes and are based on NCBI annotation, standardized by reference to identifiers from the Sequence Ontology [Eilbeck et al., 2005], see also Table 1. Values for variant type include terms such as single-nucleotide variant, deletion, and indel. Terms for clinical significance recommended by the American College of Medical Genetics and Genomics (ACMG [Richards et al., 2008]) include terms such as "Pathogenic," "Likely pathogenic," and "Benign". Variants classified as "Pathogenic" or "Likely pathogenic" were excluded from VariSNP.

UniProtKB/Swiss-Prot Data

Data were downloaded (May 30, 2014) in tab-delimited format through the SwissVar portal (http://swissvar.expasy.org/cgi-bin/swissvar/result?format=tab) and variants having a disease description were selected.

PhenCode Data

Data were downloaded from the PhenCode Website (http://phencode.bx.psu.edu/dist/phencode/database/), last update April 30, 2014, and stored in a local MySQL database. To be able to compare the variants with those in dbSNP, all variant descriptions were checked with the Mutalyzer Name Checker (http://mutalyzer.nl) and HGVS descriptions were added. Only variants with HGVS descriptions (72,544 out of 81,639) could be used for filtering.

Data Filtering, Checking, and Annotating

The HGVS descriptions of variants in the dbSNP selections were compared with HGVS descriptions in the ClinVar, UniProtKB/Swiss-Prot, and PhenCode selections. When a match was found in any of these three datasets, it was filtered out.

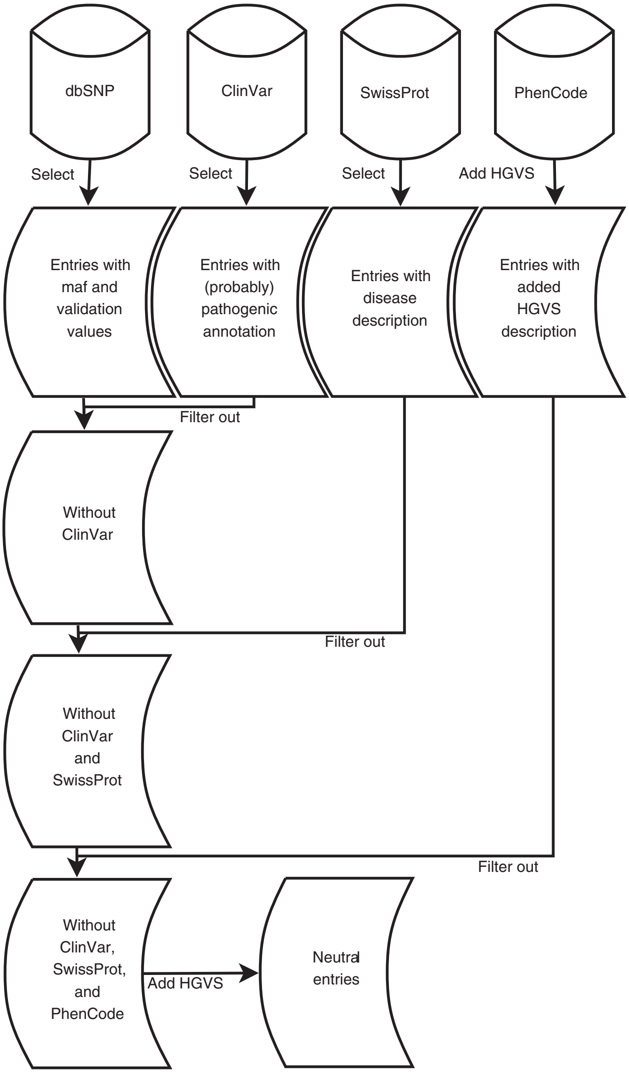

dbSNP
ClinVar
SwissProt
PhenCode
Select
Select
Select
Add HGVS
Entries with maf and validation values
Entries with (probably) pathogenic annotation
Entries with disease description
Entries with added HGVS description
Filter out
Without ClinVar
Filter out
Without ClinVar and SwissProt
Filter out
Without ClinVar, SwissProt, and PhenCode
Add HGVS
Neutral entries

Figure 1. Workflow schema of data selection and filtering.

After selection and filtering was performed, a file was produced with the HGVS variation descriptions. This file was then submitted to the batch version of the Mutalyzer Name Checker (http://mutalyzer.nl) for checking the HGVS descriptions. The results were combined with the filtered data set, using only the columns "Genesymbol," "Coding Reference," "Coding DNA Descr.," "Protein Reference," "Protein Descr.," and "Reference Sequence Start Descr."

An RNA prediction was generated from the coding DNA description and variation type annotations from the Variation Ontology (VariO, [Vihinen, 2014]) were added, using the batch version of the VariOtator tool (Schaafsma and Vihinen, in preparation).

Datasets Contents

neutral_snv

The neutral_snv dataset consists of single-nucleotide variations (SNVs; all entries in dbSNP with "missense" as functional class) with a MAF value and validation. The information in this dataset originates from three sources: dbSNP (dbSNP Build 138, genomeBuild 37.5, and groupLabel GRCh37.p10), Mutalyzer Name Checker, and VariOtator.

The dbSNP part comprises the following fields: dbSNP id (rsId), estimated average heterozygosity from allele frequencies, standard error of heterozygosity estimate, creation date, creation build, update date, update build, observed alleles, starting map location in contig coordinates, ending map location in contig coordinates, reference allele, orientation, MAF, minor allele, sample size, validation, HGVS names, allele origin, clinical significance, functional class, sequence identifier (gi), and accession.version number.

The Mutalyzer part has the following fields: Genesymbol, Reference Sequence Start Description, Coding Reference, Coding DNA description, Protein Reference, and (predicted) Protein Description. The RNA prediction can be found in the field predicted_RNA_variation. The VariO annotation consists of the appropriate VariO terms (full lineage included) for the coding DNA description and if applicable for the protein description (i.e., p. = as HGVS name does not describe a variant, so cannot have a VariO annotation).

Other datasets

Filtered datasets for the functional classes other than "missense" with more than zero variants were also generated. Each set was filtered in the same way as described above, except for checking the HGVS descriptions with Mutalyzer. Name checking was only done when possible (e.g., variants at an intronic position in combination with a coding DNA reference such as NM_000061.2 cannot be checked with Mutalyzer) and entries that could not be checked were not filtered out, as in the neutral_snv dataset.

Results and Discussion

The distribution of unfiltered and filtered variants in the functional classes is presented in Table 2.

The "intron-variant" class contains the largest number of unique variants, altogether about 91%. The average exon length (170 bp) is only about 3% of the average intron length (5,419 bp) [Sakharkar et al., 2004]. Apart from variants at or near splice sites or at exon–intron junctions, intronic variants are most often considered to be harmless, and there is no selection pressure against them. Thus, the "intron-variant" class being the largest is as expected. The "reference" functional class just refers to alleles on the contigs. The "missense" class contains over 102,000 variants, about 1.8% of the total number of variants with a MAF value and validation. The reason for the largest exclusion in this class is because amino acid substitutions are often disease-causing and because coding sequences are the most studied genome regions.

neutral_snv

Filtering of disease-causing variants left us with 80,346 entries. Checking these with the Mutalyzer Name Checker resulted in 78,951 variants with a proper HGVS description (Table 2), whereas the rest were discarded. For 1,394 variants, the HGVS description could not be checked. Reasons included "no gene specified," reference sequence not found, cases not having a correct nucleotide at the position, in-frame stop codon, and position out of range. Entries where the minor allele provided in dbSNP did not agree with the allele in the variant description were also discarded.

Table 2. Variants in dbSNP with a MAF Value and Validated, Classified According to Functional Class[a]

Functional class	Unfiltered	%	Filtered
cds-indel	306	<0.1	306
coding-sequence-variant	0	0	–
complex-change-in-transcript	0	0	–
downstream-variant-500B	47,795	0.83	47,711
downstream-variant-5KB	0	0	–
frameshift-variant	21	<0.1	20
incomplete-terminal-codon-variant	0	0	–
intron-variant	5,273,764	91.38	5,272,189
mature-miRNA-variant	0	0	–
missense	102,043	1.77	78,951
nc-transcript-variant	55,089	0.95	53,957
nmd-transcript-variant	0	0	–
nonsynonymous-codon	0	0	–
reference	193,943	3.36	–
splice-acceptor-variant	344	<0.1	338
splice-donor-variant	552	<0.1	543
splice-region-variant	0	0	–
stop-gained	1,369	<0.1	1,282
stop-lost	97	<0.1	97
synonymous-codon	91,904	1.59	91,462
upstream-variant-2KB	230,930	4.00	230,423
upstream-variant-5KB	0	0	–
utr-variant-3-prime	110,266	1.91	110,032
Unique variants	5,771,431		

[a]Note that a variant can appear in more than one functional class.

Table 3. Nucleotide Substitutions (%) in the neutral_snv Dataset

	Purines		Pyrimidines		
→	A	G	C	T	Total
A	0	13.7	3.1	2.5	19.3
G	29.5	0	5.3	4.3	39.1
C	4.8	5.8	0	21.1	31.7
T	1.9	2.0	6.0	0	9.9
Total	36.2	21.5	14.4	27.9	100

Table 4. Amino-Keto Substitutions (%) in the neutral_snv Dataset

	Amino	Keto	Total
Amino	7.9	43.1	51.0
Keto	42.7	6.3	49.0
Total	50.6	49.4	100.0

Table 5. Purine–Pyrimidine Substitutions (%) in the neutral_snv Dataset

	Purines	Pyrimidines	Total
Purines	43.2	15.2	58.4
Pyrimidines	14.5	27.1	41.6
Total	57.7	42.3	100.0

Table 3 shows an overview of the nucleotide substitutions in the neutral_snv dataset. The G>A (purine>purine) substitutions form the largest class of changes, almost 30%, followed by C>T transitions. Our results agree with the higher rate of transitions (~70%) compared with transversions found to be typical for human genes [Stephens et al., 2001].

The types of base changes were investigated more closely. Table 4 shows the substitutions from amino group containing nucleotides (A, C) to keto group containing nucleotides (G, T) and Table 5 the

Table 6. Weak–Strong Substitutions (%) in the neutral_snv Dataset

	Weak	Strong	Total
Weak	4.4	24.8	29.2
Strong	59.7	11.1	70.8
Total	64.1	35.9	100.0

purine (A, G) to pyrimidine (C, T) substitutions, and in Table 6 the weak base pair (A, T) to strong base pair (C, G) substitutions. The changes from amino to keto and vice versa are much more frequent than substitutions within these groups. There is clearly a higher frequency of transitions than transversions. The strong to weak base substitutions is by far the biggest category, containing almost 60% of the variations.

A comparison of caused amino acid substitutions is presented in Figure 2. The Arg>Gln substitutions are the most frequent, (4.56%), followed by Arg>His (4.09%), Ala>Thr (3.71%), Ala>Val (3.21%), Val>Ile (3.21%), Arg>Cys (3.02%), and Pro>Leu (3.03%) replacements.

Figure 3 shows the relative mutabilities of the amino acids to which the original residues are substituted (A), and of the original amino acids, how often amino acids have changed to a specific amino acid (B). The relative mutabilities were calculated using the following formula [Khan and Vihinen, 2007]:

$$\mathrm{Rm}(N) = \frac{(N_{\mathrm{obs}} \times N'_{\mathrm{exp}})}{(N'_{\mathrm{obs}} \times N_{\mathrm{exp}})}$$

where N' is the least mutated or variant residue type that was obtained by calculating the ratio between observed and expected value. N represents the number of original or variant residues for an amino acid type. For the original residues, tryptophan has the lowest ratio, whereas for the variant residues, proline has the lowest ratio. In the previous study [Khan and Vihinen, 2007], these were found to be lysine and alanine, respectively.

Arginine is the most frequently substituted amino acid, whereas methionine and histidine are the amino acids to which other amino acids are most often changed. The overrepresentation of arginines is known to be due to the high mutability of the codons containing CpG dinucleotides [Ollila et al., 1996], which can spontaneously mutate by deamination either to TG or CA dinucleotides. Arginine is coded by six codons, four of which have a CpG dinucleotide in the first and second codon position.

A different mutability score has been used in a study of 1000 Genomes variants by taking the total number of variants for a specific amino acid in the data and dividing by the frequency of occurrence for the specific amino acid in the genome [de Beer et al., 2013]. In the entire 1000 Genomes Project dataset, arginine had the highest mutability, whereas the more chemically complex amino acids, tryptophan, and phenylalanine, were the least mutable. Applying their method to our data shows exactly the same trend (Fig. 4).

Conclusions

To our knowledge, the neutral_snv dataset is the largest neutral variants dataset available. For the other functional classes, no

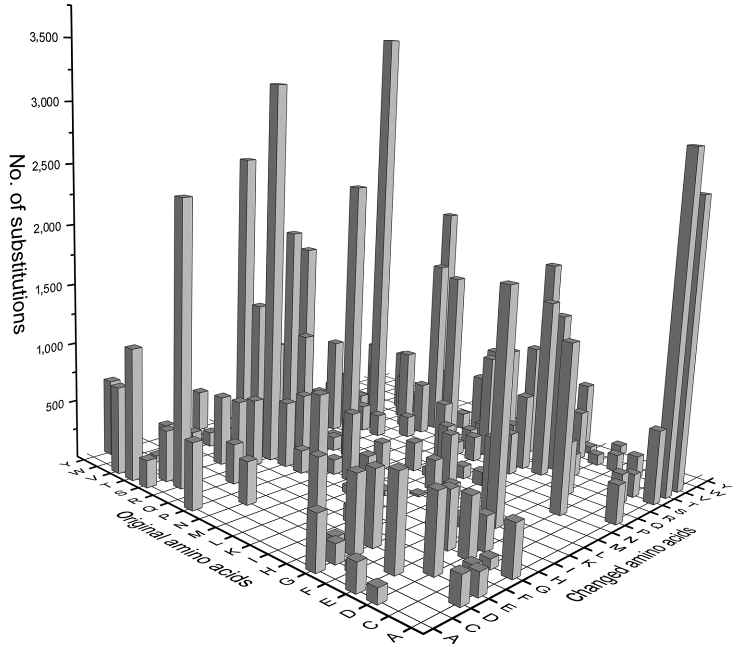


Figure 2. Lego plot of amino acid substitutions in the neutral_snv dataset; absolute numbers.

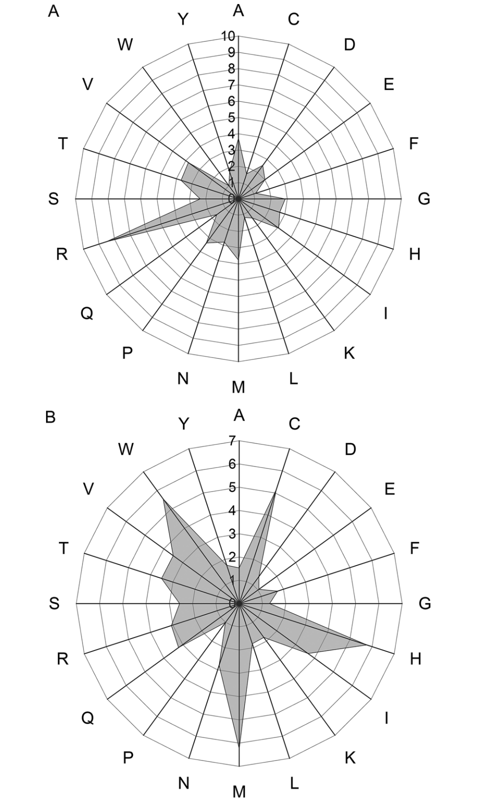


Figure 3. Relative mutabilities of mutated (A) and mutant (B) amino acids.

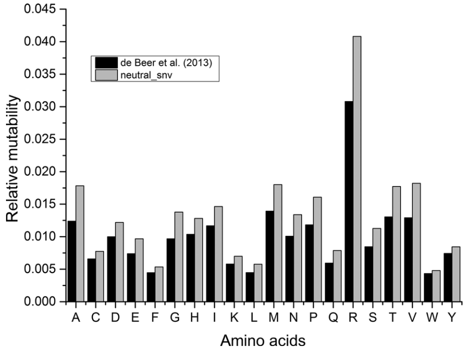


Figure 4. Relative mutability scores.

benchmark datasets have been previously available. A neutral dataset of 21,170 human nonsynonymous coding SNVs was used for method performance assessment [Thusberg et al., 2011]. Subsets of this neutral dataset, one of 17,393 cases [Olatubosun et al., 2012] and one of 14,848 cases (Niroula et al., in preparation), were used for prediction method development. Apart from being much larger, our dataset is also qualitatively better and improved, that is, due to more robust minor allele frequencies. Updates of the database will be done for each dbSNP release. All generated datasets, including those from previous updates, will be available at and can be downloaded from http://structure.bmc.lu.se/VariSNP.

Acknowledgment

Disclosure statement: The authors declare no conflict of interest.

References

de Beer TA, Laskowski RA, Parks SL, Sipos B, Goldman N, Thornton JM. 2013. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. PLoS Comput Biol 9:e1003382.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6:R44.

Famiglietti ML, Estreicher A, Gos A, Bolleman J, Géhant S, Breuza L, Bridge A, Poux S, Redaschi N, Bougueleret L, Xenarios I. 2014. Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. Hum Mutat 35:927–935.

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, Kern A, Kuhn R, et al. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat 28:554–562.

Johnston JJ, Biesecker LG. 2013. Databases of genomic variation and phenotypes: existing resources and future needs. Hum Mol Genet 22:R27–R31.

Khan S, Vihinen M. 2007. Spectrum of disease-causing mutations in protein secondary structures. BMC Struct Biol 7:56.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42:D980–D985.

Nair PS, Vihinen M. 2013. VariBench: a benchmark database for variations. Hum Mutat 34:42–49.

Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. 2012. PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 33:1166–1174.

Ollila J, Lappalainen I, Vihinen M. 1996. Sequence specificity in CpG mutation hotspots. FEBS Lett 396:119–122.

Peterson TA, Doughty E, Kann MG. 2013. Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol 425:4047–4063.

Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, Ward BE. 2008. ACMG recommendations for standards for interpreting and reporting of sequence variations: revisions 2007. Genet Med 10:294–300.

Sakharkar MK, Chow VT, Kangueane P. 2004. Distributions of exons and introns in the human genome. In Silico Biol 4:387–393.

Schaefer C, Meier A, Rost B, Bromberg Y. 2012. SNPdbe: constructing an nsSNP functional impacts database. Bioinformatics 28:601–602.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311.

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Duan J, Carr JL, et al. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493.

Thusberg J, Olatubosun A, Vihinen M. 2011. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32:358–368.

UniProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 42:D191–D198.

Vihinen M. 2012. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics 13(Suppl 4):S2.

Vihinen M. 2014. Variation ontology for annotation of variation effects and mechanisms. Genome Res 24:356–364.

Paper IV

Received: 22 February 2017 | Revised: 5 April 2017 | Accepted: 20 April 2017

DOI: 10.1002/humu.23236

RESEARCH ARTICLE

Large differences in proportions of harmful and benign amino acid substitutions between proteins and diseases

Gerard C. P. Schaafsma | Mauno Vihinen

Protein Structure and Bioinformatics, Department of Experimental Medical Science, Lund University, Lund, Sweden

Correspondence
Mauno Vihinen, Department of Experimental Medical Science, BMC B13, Lund University, SE-221 84 Lund, Sweden.
Email: mauno.vihinen@med.lu.se

Communicated by George P. Patrinos

Abstract

Genes and proteins are known to have differences in their sensitivity to alterations. Despite numerous sequencing studies, proportions of harmful and harmless substitutions are not known for proteins and groups of proteins. To address this question, we predicted the outcome for all possible single amino acid substitutions (AASs) in nine representative protein groups by using the PON-P2 method. The effects on 996 proteins were studied and vast differences were noticed. Proteins in the cancer group harbor the largest proportion of harmful variants (42.1%), whereas the non-disease group of proteins not known to have a disease association and not involved in the housekeeping functions had the lowest number of harmful variants (4.2%). Differences in the proportions of the harmful and benign variants are wide within each group, but they still show clear differences between the groups. Frequently appearing protein domains show a wide spectrum of variant frequencies, whereas no major protein structural class-specific differences were noticed. AAS types in the original and variant residues showed distinctive patterns, which are shared by all the protein groups. The observations are relevant for understanding genetic bases of diseases, variation interpretation, and for the development of methods for that purpose.

KEYWORDS
disease groups, pathogenicity, proteins, sensitivity, variation

1 | INTRODUCTION

Although modern next-generation sequencing technologies have identified huge variation datasets, which are publicly available in generic databases such as dbSNP (Sherry et al., 2001), ExAC (Lek et al., 2016), and UniProtKB/Swiss-Prot (The UniProt Consortium, 2017) and in locus-specific variation databases such as IDbases (Piirilä, Väliaho, & Vihinen, 2006) and the Duchenne Muscular Dystrophy database (Aartsma-Rus, van Deutekom, Fokkema, van Ommen, & den Dunnen, 2006), some fundamental questions still remain. One of these is how many of possible variants are harmful or benign. A further question is whether there are differences in variant frequencies between proteins, protein groups, protein structural classes, chromosomes, or amino acid substitution (AAS) types. All these questions are relevant for understanding and interpreting identified variations.

The mutation rate varies vastly depending on the chromosomal location (Smith et al., 2016). Thus, differences in proportions of harmful and benign variants are to be expected between gene and disease classes. Some disease groups, such as cardiomyopathy-related genes, have been claimed to be highly intolerant for variations (Pan et al., 2012). Recently, some studies have investigated the numbers of harmful variants in certain proteins including Bruton tyrosine kinase (BTK) (Väliaho, Faisal, Ortutay, Smith, & Vihinen, 2015) and mismatch repair system proteins (Niroula & Vihinen, 2015a). However, no systematic studies have been performed to test over large gene and protein groups. One reason has been that sufficiently large experimental datasets are missing.

Several prediction methods have been developed for variant tolerance/pathogenicity (Niroula & Vihinen, 2016), but they have largely variable performance (Bendl et al., 2014; Grimm et al., 2015; Riera, Padilla, & de la Cruz, 2016; Thusberg, Olatubosun, & Vihinen, 2011). To answer to the questions presented above, a highly reliable and fast prediction method is needed. Although experimental approaches are available for saturation mutagenesis (Haller et al., 2016) and references therein, there is a bottleneck in large-scale functional assays. We investigated all possible variants in nine groups of proteins with a highly reliable prediction method, PON-P2 (Niroula, Urolagin, & Vihinen, 2015). Large differences in proportions of harmful and benign variations were detected between the protein categories, types of variants, and their domain and chromosomal distributions.

2 | MATERIALS AND METHODS

Sequences were collected for nine categories of proteins, which represent those involved in diseases as well as housekeeping and nondisease genes and proteins. The actionable genes group (ACTION) includes 56 genes from the American College of Medical Genetics and Genomics recommendations (Green et al., 2013). Since there are therapies to treat individuals with variants in these genes, findings should be reported. Five hundred seventy-two cancer genes (CANCER) were taken from the Cancer Gene Census repository (http://cancer.sanger.ac.uk/census). From this list, a selection of 166 genes with the annotations "somatic" and "missense" was made. The cardiomyopathy set (CARDIO) contains 46 genes (Pan et al., 2012). The developmental disorder dataset (DEVEL): altogether 53 genes from Goh et al. (2007) and classified as "Developmental." The epileptic group (EPIL) of 37 genes for epileptic encephalopathy, early infantile (phenotypic series PS308350), are from the Online Mendelian Inheritance in Man (OMIM) database (Amberger, Bocchini, Schiettecatte, Scott, & Hamosh, 2015). The housekeeping dataset (HOUSE) contains 3,804 genes altogether (Eisenberg & Levanon, 2013). A random sample of 200 genes was taken for the analysis. Housekeeping genes are required for the maintenance of basal cellular functions essential for the existence of a cell, so they are expressed in most or all cells. Primary immunodeficiency (PID) genes include 283 entries from the ImmunoDeficiency Resource (Samarghitean, Väliaho, & Vihinen, 2007), IDbases (Piirilä et al., 2006), the most recent classification by the International Union of Immunological Societies (IUIS) expert committee for PIDs (Picard et al., 2015), and a recent review (Vihinen, 2015). The set of 126 neurodegenerative genes (NEURO) is from the Neurodegenerative Disease Variation database (Yang et al., submitted) at http://bioinf.suda.edu.cn/NDDvarbase/LOVDv.3.0/genes. The nondisease genes (NONDIS) dataset is a random selection of 200 genes from the HUGO Gene Nomenclature Committee (HGNC) database (Gray et al., 2013) without an OMIM id and not included in the HOUSE dataset. UniProt (The UniProt Consortium, 2017) protein accession numbers for these genes were obtained using the HGNC service, and if there was no UniProt accession available, the gene was discarded.

A Python (version 2.7.12) script was used to submit all 19 possible substitutions on all positions in the protein sequence to PON-P2 (http://structure.bmc.lu.se/PON-P2). The tool predicts the consequences of variants to be neutral, pathogenic, or unknown, and it provides predictions only for those variants for which reliable results can be obtained (default cut-off 0.95).

Statistical parameters were calculated using the stats package from the Python SciPy (release 0.18.1) library. Domains in the investigated proteins were obtained from the Pfam database (30.0) (Finn et al., 2016). Structural classifications for the domains were obtained by mapping Pfam families to the CATH database (version 4.1) (Sillitoe et al., 2015). The domains were assigned to one of the four CATH classes (mainly alpha, mainly beta, alpha beta, and few secondary structures). Protein structures were visualized with UCSF Chimera (version 1.11.2) (Pettersen et al., 2004). Protein sequence lengths were collected from the UniProtKB/Swiss-Prot database (The UniProt Consortium, 2017) and correlated with the ratios of harmful/total variants per protein. In a similar way, the number of paralogous sequences per protein, obtained from the Ensembl compara database (Herrero et al., 2016), were correlated with the ratios of harmful/total variants per protein.

The proportions of pathogenic and neutral variants in proteins were correlated to four recently introduced genic intolerance scores using Spearman's correlation coefficient. These scores include the gene damage index (Itan et al., 2015), residual variation intolerance score (Petrovski, Wang, Heinzen, Allen, & Goldstein, 2013), Aggarwala gene tolerance score (Aggarwala & Voight, 2016), and the probability of being loss-of-function intolerant score (Lek et al., 2016). The Samocha score (Samocha et al., 2014) was not studied because of a very low number of genes with index scores. As intolerance scores are not available for all the studied proteins, those without a score were not included in the comparison. Further correlation was made to the three categories of proteins obtained based on their network properties (Vinayagam et al., 2016).

3 | RESULTS AND DISCUSSION

The goal of this study was to investigate whether genes and proteins in different disease and functional groups have different sensitivity for variations and to reveal the proportions of harmful and benign AASs in protein categories. The method used to determine variation consequences was PON-P2 (Niroula et al., 2015). Numerous assessments have indicated it to have superior performance among related tools (Bendl et al., 2014; Niroula et al., 2015; Riera et al., 2016). PON-P2 is also fast, thus suitable for this kind of large-scale analysis. Further, as is shown below, our new data indicate that the ratio of benign variants predicted as harmful is very low, only 2.5%. The error rates of several other methods are even more than 10-fold higher (Bendl et al., 2014; Niroula et al., 2015; Niroula & Vihinen, 2016; Riera et al., 2016).

We collected nine datasets of altogether 1,066 genes and corresponding proteins (Table 1).

The total number of unique proteins is 996, because of overlaps between the groups (Supp. Table S1). Especially, the ACTION and CANCER sets overlap, with 17 proteins, and the CANCER and PID sets with 19 proteins. The sizes of the protein sets vary.

The ACTION set contains 56 proteins, whereas the CANCER and the PID groups contain 166 and 281 proteins, respectively. Overlap of the ACTION group is expected since this class contains some of the best studied genes and proteins from the other classes.

The predictions were made for all the possible 19 AASs at each position. In total, we made 13,540,914 unique predictions. The variants were classified into three categories: pathogenic (harmful), benign (neutral), and unknown significance (Niroula et al., 2015). In nature, the majority of the studied substitutions are very unlikely due to requiring more than one nucleotide substitution in a single codon. Out of the 380 possible AASs, only 150 can originate from single-nucleotide variations. Depending on the codon type, the number of single-nucleotide change-caused AASs (SNAVs) varies. We recently investigated all the SNAVs in the kinase domain of BTK (Väliaho et al., 2015) and in the

TABLE 1 Predicted outcome of variants in the nine datasets

	ACTION	CANCER	CARDIO	DEVEL	EPIL	HOUSE	NEURO	NONDIS	PID
Number of genes	56	166	46	53	37	193	126	175	281
Number of predicted proteins	56	159	45	53	37	187	120	146	263
Predicted proteins (%)	100.00	95.78	97.83	100.00	100.00	96.89	95.24	83.43	93.59
Number of amino acids	65,981	163,244	67,996	44,972	37,198	92,800	81,827	74,964	180,535
Number of possible variants in predicted proteins	1,253,639	3,101,636	1,291,924	854,468	706,762	1,763,200	1,554,713	1,424,316	3,430,165
Number of predicted variants	1,251,692	3,099,458	806,345	793,130	703,309	1,760,880	1,553,707	1,398,123	3,416,178
% predicted variants (of possible)	99.84	99.93	62.41	92.82	99.51	99.87	99.94	98.16	99.59
Number of variants predicted as 'neutral'	110687	277241	92985	98410	112301	460746	237948	1026537	825655
% neutral variants	8.84	8.95	11.53	12.41	15.97	26.17	15.32	73.42	24.17
Average number of neutral variants per protein	1,976.55	1,743.65	2,066.33	1,856.79	3,035.16	2,463.88	1,982.90	7,031.08	3,139.37
Median number of neutral variants per protein	580	222	780	642	671	1,185	801	5,208	1,438
Number of variants predicted as "pathogenic"	440,982	1,303,446	212,579	280,692	262,602	414,743	525,790	58,630	880,146
Pathogenic variants (%)	35.23	42.05	26.36	35.39	37.34	23.55	33.84	4.19	25.76
Average number of pathogenic variants per protein	7,874.68	8,197.77	4,723.98	5,296.08	7,097.35	2,217.88	4,381.58	401.58	3,346.56
Median number of pathogenic variants per protein	4,924.5	6,649	2,615	3,704	6,216	1,135	2,918.5	0	2,067
Number of variants predicted as "unknown"	700,023	1,518,771	500,781	414,028	328,406	885,391	789,969	312,956	1,710,377
Unknown variants (%)	55.93	49	62.11	52.2	46.69	50.28	50.84	22.39	50.07
Average number of unknown variants per protein	12,500.41	9,552.02	11,128.47	7,811.85	8,875.84	4,734.71	6,583.08	2,143.53	6,503.34
Median number of unknown variants per protein	7,500	6,305	4,408	5,680	6,754	3,356	4,549.5	166	4,859
Ratio of pathogenic and neutral variants	3.99	4.70	2.29	2.85	2.34	0.90	2.21	0.06	1.07

mismatch repair proteins (Niroula & Vihinen, 2015a) by using PON-BTK and PON-MMR2, respectively, and found large differences in the proportions of tolerated variants.

The investigated proteins show also in this study great variation in the proportions of harmful, benign, and unknown variants (Supp. Tables S2–S4). PON-P2 uses classification by reject option, which means that cases without strong evidence for being either harmful or harmless are grouped into the unknown class. The benefit is that the predicted harmful and tolerated variants are correct with very high likelihood (Niroula et al., 2015). Further, because of heterogeneity of the phenotype caused by variants (Vihinen, 2017), we expect the method to predict cases to the unknown class. The results thus present the lower boundaries for the harmful and benign classes. PON-P2 is a tolerance (pathogenicity) predictor and has been trained on known disease-causing and benign variants. Therefore, it cannot be used to make for example functional or structural predictions. Thus, although a certain variant may be, for example, structurally incompatible, it may not be harmful if the protein is not essential.

The most sensitive proteins for variations, that is, those with the largest proportion of pathogenic variants, are found in the CANCER group, although a few proteins from the HOUSE group also have very high percentages of pathogenic predictions. The protein with the highest percentage of pathogenic predictions (90.3%) is the ubiquitin-conjugating enzyme E2 B (P63146) that belongs to HOUSE group (Supp. Dataset S1). This is a very high proportion for harmful variations, but in line with the protein function. The protein is an essential component of protein degradation processes and postreplicative DNA damage repair. The sequence is entirely invariant with several species, indicating the high conservation required for function, which on the other hand leads to a high sensitivity for variants.

Proteins with the lowest ratio of pathogenic predictions (0%) are mainly from the HOUSE and NONDIS groups. Remarkable exceptions are the SET-binding protein (Q9Y6X0) from the HOUSE group and the inducible T-cell costimulator (Q9Y6W8) from the PID group. Proteins with the highest percentages for the neutral predictions (Supp. Dataset S2) belong mainly to the NONDIS group, and some to the HOUSE group, although there are also a few proteins from the disease groups PID and NEURO, which have more than 95% neutral predictions. The proteins with no neutral predictions all belong to a disease group (ACTION, CARDIO, CANCER, DEVEL, NEURO, or PID) except for the ubiquitin-conjugating enzyme E2 B, from the HOUSE group.

The ratio of unknown variants ranges from 99% for histone H3.3 in the CANCER set to close to 0 for several proteins mainly in the NONDIS category (Supp. Dataset S3). There are several reasons why variants are predicted to the unknown class, all of them related to the fact that the predictor cannot find sufficient support (at least 0.95) for classification into either the neutral or pathogenic classes. These include low sequence conservation and the presence of very different types of residues in corresponding positions in the multiple sequence alignment.

The distributions of the harmful and benign variants as well as those with unknown significance are shown in Figure 1 for the kinase domain of BTK (PDB id 3k54). Our previous study indicated that 67% of SNAVs are harmful (Väliaho et al., 2015). The corresponding number for all AASs is 72.7%. The distributions of predicted variant outcomes are clearly different at different positions due to the accessibility of

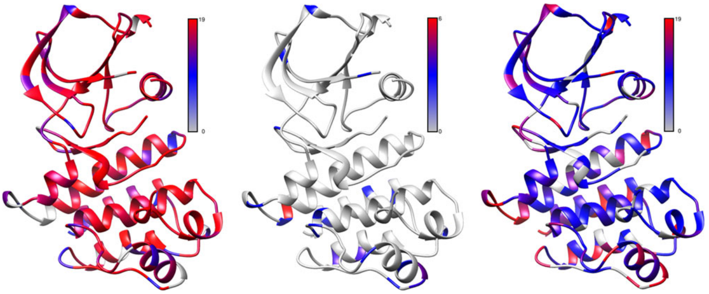


FIGURE 1 The distribution of predicted variants in BTK. The distribution of the predicted pathogenic (left) and neutral (middle) variants and those with unknown significance (right) in the tyrosine kinase domain of Bruton tyrosine kinase (PDB id 3k54). The distributions of the variant effects are highly structural context-dependent. The numbers of the variation types in each category are shown by the scales on the right side of the panel

the site, involvement in binding or catalysis, or location within secondary structural elements. Harmful variants are most frequent at the protein core and at the secondary structural elements. Neutral variants are mainly located to surface loops or toward the ends of α- and β-structures. These are also the locations for the largest numbers of unknown variants. The BTK kinase domain contains more harmful variants than many other domains or proteins due to importance of numerous sites for the catalytic activity, substrate binding, and multimodal regulation of the protein. Therefore, the structure contains numerous positions at which only a few, if any, substitutions are tolerated.

3.1 | Comparison of protein classes for variation sensitivity

PON-P2 has shown excellent performance, both for neutral and harmful variants (Bendl et al., 2014; Niroula et al., 2015; Riera et al., 2016). To find out the false prediction rate for neutral variants, we collected all AASs from VariSNP (release 2016-06-09) (Schaafsma & Vihinen, 2015a), a database for cases filtered from dbSNP (Sherry et al., 2001) to not contain disease variants. The selection consisted of 26,121 variants with a minor allele frequency of 0.01 or higher, and not present in the PON-P2 neutral training dataset. Of these, 17,667 (67.6%) variants could be predicted by PON-P2, 73.5% as neutral and 2.5% as pathogenic. The reason for having 24.0% of variants as unknown is at least partly because of phenotypic heterogeneity (Vihinen, 2017). Results indicate the ratio of false pathogenic predictions to be very low.

The investigated protein classes represent current knowledge, and it is likely that novel proteins will be added at least to disease classes in the future. These may slightly change the overall results but likely not very much as all the classes are already of substantial size and addition of a few proteins cannot largely alter the identified patterns. Although the selected classes represent numerous different functions and aspects of genes and proteins, they are only a sample of the entire genome and proteome. It remains to be seen whether substantially larger deviations from the investigated classes could be found. We do not expect major differences as the investigated classes well represent the studied classes including the extremes: cancer proteins and nondisease nonhousekeeping genes/proteins. Anyhow, it is important to know the sensitivity for variations in each gene/protein category as well as the deviations within the classes.

In the analyses of the protein categories, the percentage of predicted variants is high, over 92% (Table 1), even higher than for the PON-P2 test dataset (86%) (Niroula et al., 2015). Predictions were made for most of the proteins, typically for over 95% and in many datasets for all proteins. The only exception is the NONDIS dataset where predictions were made only for 146 out of 175 proteins (83%). If a protein sequence is unique for human, PON-P2 cannot make predictions as these would be unreliable, due to missing the selective pressure and sequence profile features that are based on multiple sequence alignments of orthologous sequences. Some sequences such as Q96L12 could not be mapped to an Ensembl reference sequence and were discarded. The percentage of predicted variants is very high, over 90%, except for the CARDIO set. This is due to one single protein, titin, which is the largest human protein with 34,350 residues in the longest isoform. Because of its repetitive nature (over 200 copies of type I and II domains), only 25.7% of the variants in this protein can be reliably predicted.

The percentages of variants classified as unknown are found to be around 50%–55%, the largest deviations being 62.1% for the CARDIO set and 22.4% for the NONDIS dataset (Table 1).

The ratios of neutral and pathogenic variants are clearly different for the protein groups. The largest number of neutral variants, 73.4%, appears in the NONDIS group. In this group, only 4.19% of the variants are predicted as harmful. These figures indicate the high tolerance of proteins in this group for variations.

In the other protein groups, the highest ratios of neutral variants appear in the HOUSE and PID proteins, 26.2% and 24.2%, respectively. The smallest ratios of neutral variants appear in the ACTION and CANCER groups, 8.8% and 9.0%, respectively. These groups are partly overlapping (Supp. Table S1). Of these, the CANCER group contains the largest frequency of harmful variants, 42.1%, in line with our previous study (Niroula & Vihinen, 2015b).

The HOUSE and PID categories have the lowest ratios of harmful variants, 23.6% and 25.8%, respectively. The CARDIO disease group,

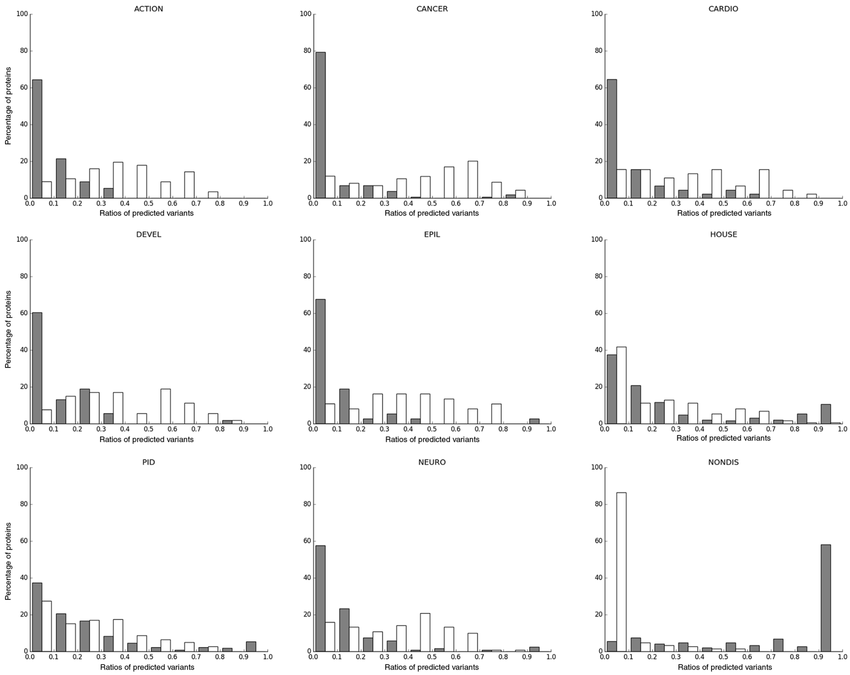


FIGURE 2 Ratios of predicted variation outcomes. Distribution of the proteins over the prediction ratios (neutral/total and pathogenic/total) for each dataset. The results for neutral variants are shown in gray bars and for pathogenic with white bars

claimed to be highly intolerant (Pan et al., 2012), has a higher degree of harmful variants (26.4%) and a lower frequency of neutral variants (11.5%) than these two groups. This class is not especially prone for harmful variants. There are other disease groups with significantly larger proportions of harmful variants.

When looking at the average and median numbers of variants per protein, we see huge differences. The median for pathogenic variants per protein is 0 for the NONDIS proteins (average 401.58) and 6649 for the CANCER class (average 8,197.77) (Table 1).

The range in the ratios for harmful versus neutral variants is also very wide, from 0.06 for the NONDIS group to 4.70 for the CANCER proteins. The only groups with a close to equal ratio are HOUSE and PID. The high ratio is expected for the CANCER class. The results indicate that even in the so-called cancer driver proteins, the majority of variants are not harmful. The percentage for unknowns is 49% for CANCER. By far, the smallest proportion of unknown variants is in the NONDIS group where only 22.4% of the predictions are of unknown significance.

The unknown class contains cases that the predictor cannot separate to either harmful or neutral. Many of these variants can have a variable phenotype depending on the other factors including severity, extent, and modulation in case of diseases (Vihinen, 2017). As the first application of the pathogenicity model, we recently introduced a method for predicting disease severity of variants (Niroula & Vihinen, 2017).

The distributions of proteins to bins according to ratios of predicted harmful and neutral variants are shown in Figure 2. The CANCER and NONDIS groups represent the two extremes. In CANCER, almost 80% of the proteins have less than 10% of variants predicted to be neutral, whereas in the NONDIS group, about 60% of the proteins contain >90% neutral variants. The HOUSE and PID groups have the most even distributions in the bins. The percentages of proteins in the bins with the largest ratios are small. The results indicate that the protein groups are heterogeneous regarding variant proportions; however, they show certain clear trends. Results are in line with those predicted by the ExAC project (Lek et al., 2016). They found that in total 3,230 genes are highly intolerant for function destroying variants. Thirty out of 42 (71%) proteins with a proportion of pathogenic variants >70% are among the ExAC data.

To see whether the protein size correlates with the sensitivity, we plotted sequence length versus ratio of pathogenicity for each protein class (Supp. Fig. S1) and studied the length distributions within the classes (Supp. Fig. S2). No correlation was noted. When we took into consideration the possibility that the existence of paralogous proteins could contribute to the proportions of pathogenic variations, no

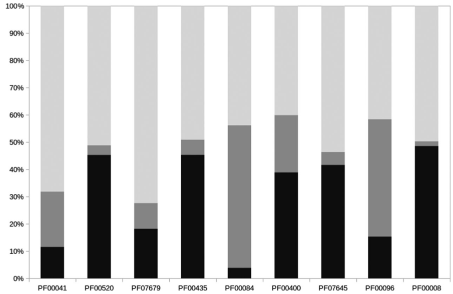


FIGURE 3 Analysis of variation distribution in protein domains. Distribution of neutral (gray), pathogenic (black), and unknown (light gray) predictions in the nine most frequent domains in the proteins. PF07679: Immunoglobulin I-set domain; PF00041: Fibronectin type III domain; PF00084: Sushi domain; PF00400: WD domain, G-beta repeat; PF00096: Zinc finger, C2H2 type; PF00008: EGF-like domain; PF07645: Calcium binding EGF domain; PF00520: Ion transport domain; PF00435: Spectrin repeat

statistically significant correlation was found in any class either (Supp. Fig. S3).

3.2 | Sensitivity of protein domains for variants

Next, we investigated the variants in the most frequent protein domains in the datasets. Altogether, nine domains appeared at least 40 times in the investigated proteins (Supp. Table S2). The immunoglobulin I-set domain (PF07679) is the most frequent one (189 times). Among the nine most frequent domains, it has the lowest proportion of neutral and pathogenic predictions together, 27.7% (Fig. 3). The immunoglobulin I-set domain is frequent in cell adhesion proteins, but appears also in many other types of proteins.

The domain with the highest sum of harmful and neutral variants (60.0%) is WD domain, G-beta repeat (PF00400), immediately followed by zinc finger domain, C2H2 type (PF00096, 58.0%), and Sushi repeat (PF00084, 56.2%). The Sushi repeat domain (SCR) has the highest proportion of neutral predictions, 52.3%. Sushi domains are involved in many recognition processes and so are also the I-set, WD40, zinc finger, and calcium-binding EGF-domain, and EGF-like domain.

The structural classifications for all domains in the investigated proteins were obtained by mapping to the CATH database (Supp. Table S3; Supp. Fig. S4). Mappings are available for 30.7% of the total length of all sequences. The mixed alpha beta class is the most common (45.6%) followed by the mainly alpha and mainly beta classes. Only seven domains belong to the smallest class of few secondary structures. The distributions for the predicted outcome of variants are quite similar for the three major categories. Variants with unknown prediction are clearly the smallest category in all the groups. The numbers in the proteins with few secondary structures are so small that the differences cannot be considered as statistically significant.

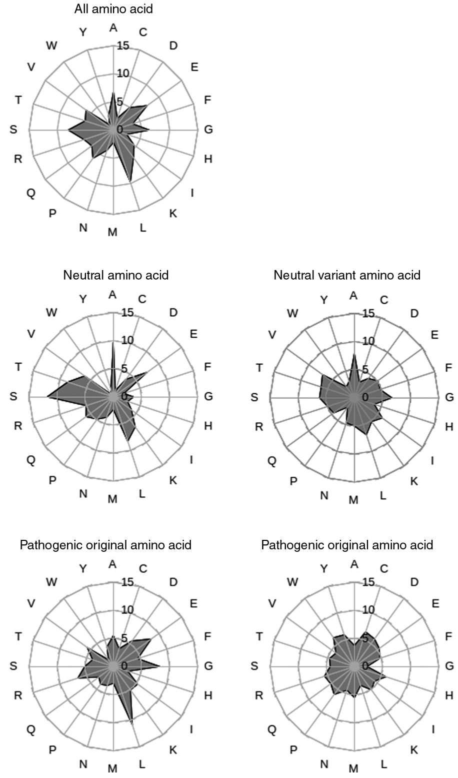


FIGURE 4 Distributions of amino acid substitutions. Amino acid substitutions among the original and variant amino acids for those predicted to be neutral and pathogenic; ACTION dataset

3.3 | Analysis of amino acids

To further investigate the types of variants among the neutral and harmful variants, amino acid distributions for the original and variant residues were studied. Results are shown in Figure 4 for the ACTION class. Results for the other groups are in Supp. Figure S5. The distributions are clearly different for the neutral and harmful variants as well as for the original residues. Results are very similar between the tested categories.

Substitutions from S are the most frequent ones among the neutral AASs (11.7%), followed by substitutions from A (9.8%), E (7.4%), and L (8.3%). The least frequent neutral predictions are for the substitutions from W (0.5%), C (1.2%), and Y (1.4%). Among the pathogenic predictions, the substitutions from L (10.9%) are the most frequent ones, followed by those from E (8.3%) and G (8.2%). The least frequent pathogenic substitutions are from W (1.5%).

Substitutions to A are the most frequent ones among the neutral AASs (7.9%), followed by substitutions to G (6.7%), L (6.9%), and V (7.2%). The least frequent neutral predictions are for the substitutions to C (3.0%), P (2.1%), W (2.5%), and Y (3.0%). Among the pathogenic predictions, the substitutions to C (6.5%) are the most frequent ones, followed by those to D (6.1%), P (6.1%), and W (6.4%). The least frequent pathogenic substitutions are to G (2.3%).

The figures for the variant residues show more even distributions, especially for the harmful predictions. Substitutions to G (2.3%) are the least frequent pathogenic predictions, whereas no clear maximum can be found. Glycine and alanine substitutions are largely tolerated outside functional sites as these smallest residues can be easily fitted into structures without steric clashes and structural alterations. For all the investigated datasets, A (7.9%), V (7.0%), G (6.7%), L (6.9%), S and T (both 6.2%), and I (6.1%) are the amino acids with the highest numbers of neutral predictions. The lowest numbers of neutral predictions are for P (2.1%) and W (2.5%). Proline has due to its special structure a rigid backbone conformation that breaks secondary structural elements. Tryptophan has the largest side chain and is thus difficult to accommodate to other positions than those on the protein surface.

The distributions of the predicted harmful and benign variants are qualitatively similar to those published previously, for the 1000 Genomes project (de Beer et al., 2013), those annotated in UniProt (Petukh, Kucukkal, & Alexov, 2015), and variants in BTK (Schaafsma & Vihinen, 2015b).

3.4 | Comparison to genic intolerance scores

Gene-specific indices have recently been introduced to describe intolerance for variants (Aggarwala & Voight, 2016; Itan et al., 2015; Lek et al., 2016; Petrovski et al., 2013; Samocha et al., 2014). Although interesting, they suffer from extremely small sample sizes, in the case of EVP6500-based (Tennessen et al., 2012) scores on average 5.8 variants per gene, and totally on 2,835 proteins with just one single variant. For the 1000 Genomes Project, the numbers are even less (1000 Genomes Project Consortium et al., 2015), and somewhat higher for the ExAC-based data (Lek et al., 2016). Random events are likely to have a great effect on these indices. Another problem is that they have not been experimentally validated unlike tolerance/pathogenicity predictors. The network-derived classes (Vinayagam et al., 2016) do not correlate with ratios of pathogenic or neutral variants in the entire protein set.

Results in Supp. Table S4 show only a marginal correlation with our predictions that can be considered highly reliable due to extensive benchmarking on numerous proteins (Niroula et al., 2015; Riera et al., 2016). The highest correlation coefficient is only 0.61, between the Aggarwala gene tolerance score and the ACTION set.

The essentiality of genes in the human genome has been investigated by the genome editing approach (Blomen et al., 2015; Wang et al., 2015). These gene sets were not investigated further here as they largely overlap with the housekeeping genes. The HOUSE set contains 130 essential genes mentioned by Blomen et al. (2015) and 140 essential genes mentioned by Wang et al. (2015).

Saturation mutagenesis combined with high-throughput functional assays would provide the gold standard data. However, such data are largely missing. Further, the functional assays should describe the cellular effect. The decreased/increased activity level required for phenotype varies widely. One of the extreme cases is adenosine deaminase where enzymatic activity less than 1% is indicative of the severe combined immunodeficiency (Arredondo-Vega, Santisteban, Daniels, Toutain, & Hershfield, 1998). Direct predictions with PON-P2 provide a much more reliable estimation of gene intolerance for substitutions than the derived indices.

3.5 | Chromosomal distribution

The distribution of the predicted tolerance effects over the chromosomes is shown in Supp. Table S5. No genes coding for the studied proteins are located on the Y chromosome, except for P15509, which is encoded by *CSF2RA* on the pseudoautosomal region 1. This protein is assigned also to the X chromosome. The percentages of neutral variants range from 12.3% for chromosome 13 to 43.8% for chromosome 18. The differences are smaller for the pathogenic variants, from 19.1% (chromosome 6) to 38.9% (chromosome 15). The percentages of variants predicted as unknown are from 35.4% (chromosome 18) to 57.8% (chromosome 13). Compared with the ranges of neutral and pathogenic predictions over the datasets (Table 1, neutral 8.8%–73.4%; pathogenic 4.2%–42.1%), the differences between the chromosomes are smaller than those between the protein groups. The range of the ratios for pathogenic and neutral variants is from 0.47 (chromosome 18) to 2.43 (chromosome 13), this range also being smaller than the range of ratios over the entire datasets.

In Supp. Table S6, the distribution of the studied proteins over the chromosomes is presented.

The largest number of variants originates from proteins coded by genes on chromosome 1, as expected since it is the largest chromosome and contains the largest number of coding genes. For the other chromosomes, there seems not to be a clear pattern. The EPIL set has the highest number of chromosomes without any protein coding any of the investigated proteins, but it is also the set with the lowest number of proteins, only 37.

3.6 | Ratios of harmful and benign variants

Our results indicate that the proportions of variants of harmful, benign, and those of unknown significance vary widely between proteins. Extreme examples (e.g., Q8N9V6, ankyrin repeat domain-containing protein 53; Q9P2E3, NFX1-type zinc finger-containing protein 1; Q5T870, proline-rich protein 9) include proteins in which no AASs are predicted to be harmful. On the other end of the spectrum, almost all variants are predicted to be harmful (e.g., P63146, ubiquitin-conjugating enzyme E2 B; P35222, catenin beta-1; P61586, transforming protein RhoA). It is not possible to give a single number for the variant effects as it depends on the protein and protein domain.

Previously, several attempts have been made to reveal the rate of harmful variants based on several approaches, especially the strength of positive and purifying selection. Fitness effects have been calculated utilizing evolutionary information for sequence relationships. Germline mutation rates have been estimated by using human genetic disease phenotype frequencies, between species nucleotide divergence at putatively neutral sites, and by sequencing genomes of relatives (Keightley, 2012).

The results from these studies indicate a wide spectrum and concentrate on genome and proteome wide estimates not taking the differences between genes/proteins into account. 20% of AAS have been presented to lead to loss of function, and up to 70% of low-frequency AASs have been claimed to be mildly deleterious (Kryukov, Pennacchio, & Sunyaev, 2007). Based on the recent data in the ExAC database, 99% of the listed variants are rare and most of them are not considered to be linked to any condition. In other studies, 15%, 29%–49%, and 48% of AASs were presented to be strongly deleterious (Boyko et al., 2008; Eyre-Walker, Woolfit, & Phelps, 2006; Subramanian, 2012).

The results presented here can be considered more reliable because of the following reasons. Our results are specific for each individual protein, not estimates over entire genomes/proteomes. The fitness-related methods are largely based on evolutionary information, which indeed is very important for variant tolerance predictors, such as PON-P2. However, evolutionary information alone is not sufficient for high-performance predictions. As assessments have shown, the mainly or only evolutionary information-based methods like SIFT (Ng & Henikoff, 2001), Condel (Gonzalez-Perez & Lopez-Bigas, 2011), PROVEAN (Choi, Sims, Murphy, Miller, & Chan, 2012), and MutationAssessor (Reva, Antipin, & Sander, 2007) are far behind the state-of-the-art tools (Bendl et al., 2014; Grimm et al., 2015; Niroula & Vihinen, 2016; Riera et al., 2016). The successful prediction of variant effects requires additional features to describe the variants and sites at which they occur. Just as with the genic intolerance methods, the tools that are based on allele frequencies still suffer from relatively small sizes of available datasets. In conclusion, the estimates of benign and harmful variants calculated in here are more accurate and realistic than those published before.

4 | CONCLUSIONS

Analysis of all possible AASs in nine protein groups indicated that there are big differences in the ratios of harmful, pathogenic, and unknown variants. This information can be utilized for and should be considered in the interpretation of variant effects. The published gene intolerance scores do not correlate with our observations (Supp. Table S4), which is not a surprise considering the small numbers of cases used to define those scores.

In the disease-associated protein groups, the proportion of predicted neutral variants ranges from around 38% for the proteins having a neutral/total prediction ratio between 0 and 0.1 in the PID dataset to about 80% in the CANCER dataset (Fig. 2). Interestingly, this observation does not coincide with the high pathogenic/total prediction ratio. The NONDIS set showed quite a different pattern: the highest number of proteins (80%) had a pathogenic/total prediction ratio of 0–0.1, where the highest number of proteins (~68%) had a neutral prediction ratio between 0.9 and 1.0. These proteins are likely not involved in essential functions and are therefore the most tolerant for variants. The distributions of the original and variant AASs are similar for all datasets presumably originating from the combination of PON-P2 features. Like protein classes, the frequently appearing protein domains in the dataset show a wide spectrum of variant frequencies. Protein structural classes do not show major differences for variants.

In summary, our observations indicate clear protein, protein domain, and AAS type differences due to variants. This information is relevant for variation interpretation and for the development of methods for that purpose.

ACKNOWLEDGMENT

We want to thank Vetenskapsrådet and the Swedish Childhood Cancer Foundation for support.

DISCLOSURE STATEMENT

The authors declare no conflict of interest.

REFERENCES

Aartsma-Rus, A., van Deutekom, J. C., Fokkema, I. F., van Ommen, G. J., & den Dunnen, J. T. (2006). Entries in the Leiden Duchenne muscular dystrophy mutation database: An overview of mutation types and paradoxical cases that confirm the reading-frame rule. *Muscle & Nerve*, *34*, 135–144.

Aggarwala, V., & Voight, B. F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, *48*, 349–355.

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43*, D789–D798.

Arredondo-Vega, F. X., Santisteban, I., Daniels, S., Toutain, S., & Hershfield, M. S. (1998). Adenosine deaminase deficiency: Genotype-phenotype correlations based on expressed activity of 29 mutant alleles. *The American Journal of Human Genetics*, *63*, 1049–1059.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., … Damborsky, J. (2014). PredictSNP: Robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Computational Biology*, *10*, e1003440.

Blomen, V. A., Majek, P., Jae, L. T., Bigenzahn, J. W., Nieuwenhuis, J., Staring, J., … Brummelkamp, T. R. (2015). Gene essentiality and synthetic lethality in haploid human cells. *Science*, *350*, 1092–1096.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., … Bustamante, C. D. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics*, *4*, e1000083.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, *7*, e46688.

de Beer, T. A., Laskowski, R. A., Parks, S. L., Sipos, B., Goldman, N., & Thornton, J. M. (2013). Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Computational Biology*, *9*, e1003382.

Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, *29*, 569–574.

Eyre-Walker, A., Woolfit, M., & Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, *173*, 891–900.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, *44*, D279–D285.

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabasi, A. L. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 8685–8690.

Gonzalez-Perez, A., & Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American Journal of Human Genetics*, *88*, 440–449.

Gray, K. A., Daugherty, L. C., Gordon, S. M., Seal, R. L., Wright, M. W., & Bruford, E. A. (2013). Genenames.org: The HGNC resources in 2013. *Nucleic Acids Research*, *41*, D545–D552.

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., … American College of Medical Genetics and Genomics. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine*, *15*, 565–574.

Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., … Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, *36*, 513–523.

Haller, G., Alvarado, D., McCall, K., Mitra, R. D., Dobbs, M. B., & Gurnett, C. A. (2016). Massively parallel single-nucleotide mutagenesis using reversibly terminated inosine. *Nature Methods*, *13*, 923–924.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., … Flicek, P. (2016). Ensembl comparative genomics resources. *Database (Oxford)*, *2016*, bav096.

Itan, Y., Shang, L., Boisson, B., Patin, E., Bolze, A., Moncada-Velez, M., … Casanova, J. L. (2015). The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 13615–13620.

Keightley, P. D. (2012). Rates and fitness consequences of new mutations in humans. *Genetics*, *190*, 295–304.

Kryukov, G. V., Pennacchio, L. A., & Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *The American Journal of Human Genetics*, *80*, 727–739.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*, 285–291.

Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research*, *11*, 863–874.

Niroula, A., Urolagin, S., & Vihinen, M. (2015). PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS One*, *10*, e0117380.

Niroula, A., & Vihinen, M. (2015a). Classification of amino acid substitutions in mismatch repair proteins using PON-MMR2. *Human Mutation*, *36*, 1128–1134.

Niroula, A., & Vihinen, M. (2015b). Harmful somatic amino acid substitutions affect key pathways in cancers. *BMC Medical Genomics*, *8*, 53.

Niroula, A., & Vihinen, M. (2016). Variation interpretation predictors: Principles, types, performance, and choice. *Human Mutation*, *37*, 579–597.

Niroula, A., & Vihinen, M. (2017). Predicting severity of disease-causing variants. *Human Mutation*, *38*, 357–364.

Pan, S., Caleshu, C. A., Dunn, K. E., Foti, M. J., Moran, M. K., Soyinka, O., & Ashley, E. A. (2012). Cardiac structural and sarcomere genes associated with cardiomyopathy exhibit marked intolerance of genetic variation. *Circulation: Cardiovascular Genetics*, *5*, 602–610.

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., & Goldstein, D. B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genetics*, *9*, e1003709.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*, 1605–1612.

Petukh, M., Kucukkal, T. G., & Alexov, E. (2015). On human disease-causing amino acid variants: Statistical study of sequence and structural patterns. *Human Mutation*, *36*, 524–534.

Picard, C., Al-Herz, W., Bousfiha, A., Casanova, J. L., Chatila, T., Conley, M. E., … Gaspar, H. B. (2015). Primary immunodeficiency diseases: An update on the classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency 2015. *Journal of Clinical Immunology*, *35*, 696–726.

Piirilä, H., Väliaho, J., & Vihinen, M. (2006). Immunodeficiency mutation databases (IDbases). *Human Mutation*, *27*, 1200–1208.

Reva, B., Antipin, Y., & Sander, C. (2007). Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biology*, *8*, R232.

Riera, C., Padilla, N., & de la Cruz, X. (2016). The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. *Human Mutation*, *37*, 1013–1024.

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., … Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, *46*, 944–950.

Samarghitean, C., Väliaho, J., & Vihinen, M. (2007) IDR knowledgebase for primary immunodeficiencies. *Immunome Research 3*:6.

Schaafsma, G. C. P., & Vihinen, M. (2015a). VariSNP, a benchmark database for variations from dbSNP. *Human Mutation*, *36*, 161–166.

Schaafsma, G. C. P., & Vihinen, M. (2015b). Genetic variation in Bruton tyrosine kinase. In A. Plebani & V. Lougaris (Eds.), *Agammaglobulinemia* (pp. 75–85). Switzerland: Springer International Publishing.

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*, 308–311.

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson N. L., … Orengo, C. A. (2015). CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, *43*, D376–D381.

Smith, T., Ho, G., Christodoulou, J., Price, E. A., Onadim, Z., Gauthier-Villars, M., … Eyre-Walker, A. (2016). Extensive variation in the mutation rate between and within human genes associated with Mendelian disease. *Human Mutation*, *37*, 488–494.

Subramanian, S. (2012). The abundance of deleterious polymorphisms in humans. *Genetics*, *190*, 1579–1583.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., … NHLBI Exome Sequencing Project. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, *337*, 64–69.

The UniProt Consortium. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*, D158–D169.

Thusberg, J., Olatubosun, A., & Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutation*, *32*, 358–368.

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., … Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science*, *350*, 1096–1101.

Väliaho, J., Faisal, I., Ortutay, C., Smith, C. I., & Vihinen, M. (2015). Characterization of all possible single-nucleotide change caused amino acid substitutions in the kinase domain of Bruton tyrosine kinase. *Human Mutation*, *36*, 638–647.

Vihinen, M. (2015). Immunodeficiency, primary: Affecting the adaptive immune system. In: *eLS* (pp. 1–6). Chichester, UK: John Wiley & Sons.

Vihinen, M. (2017). How to define pathogenicity, health, and disease? *Human Mutation*, *38*, 129–136.

Vinayagam, A., Gibson, T. E., Lee, H. J., Yilmazel, B., Roesel, C., Hu, Y., … Barabasi, A. L. (2016). Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy Sciences of the United States of America*, *113*, 4976–4981.

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Schaafsma GCP, Vihinen M. Large differences in proportions of harmful and benign amino acid substitutions between proteins and diseases. *Human Mutation*. 2017;38:839–848. https://doi.org/10.1002/humu.23236

Paper V

Printed by Media-Tryck, Lund 2017 NORDIC ECOLABEL, 3041 0903

LUND UNIVERSITY
Faculty of Medicine

Department of Experimental Medical Science

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2017:132
ISBN 978-91-7619-515-4
ISSN 1652-8220

9 789176 195154