UniProt Genomic Mapping for Deciphering Functional Effects of Missense Variants

Peter B. McGarvey[1,4, §], Andrew Nightingale[3,4], Jie Luo [3,4], Hongzhan Huang[2,4], Maria J. Martin[3,4], Cathy Wu[2,4], and the UniProt Consortium[4]

1. Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC, USA.
2. Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA
3. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
4. SIB Swiss Institute of Bioinformatics (SIB), Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland; Protein Information Resource (PIR), Washington, DC and Newark, DE, USA; European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

[§]Corresponding author

Email addresses:

PBM:	pbm9@georgetown.edu
AN:	anight@ebi.ac.uk
J.L:	jluo@ebi.ac.uk
MJM:	martin@ebi.ac.uk
H.H:	huang@dbi.udel.edu
CW:	wuc@dbi.udel.edu

Keywords: UniProt database; missense variants; genome mapping

Abstract

Understanding the association of genetic variation with its functional consequences in proteins is essential for the interpretation of genomic data and identifying causal variants in diseases. Integration of protein function knowledge with genome annotation can assist in rapidly comprehending genetic variation within complex biological processes. Here, we describe mapping UniProtKB human sequences and positional annotations such as active sites, binding sites, and variants to the human genome (GRCh38) and the release of a public genome track hub for genome browsers. To demonstrate the power of combining protein annotations with genome annotations, we present specific biological examples in disease-related genes and proteins; illustrating how the combination of protein and genome annotations helps the functional interpretation of variants. Computational comparisons of UniProtKB annotations and protein variants with ClinVar clinically annotated SNP data show that 32% of UniProtKB variants collate with 8% of ClinVar SNPs. The majority of co-located UniProtKB disease-associated variants (86%) map to 'pathogenic' ClinVar SNPs. UniProt and ClinVar are collaborating to provide a unified clinical variant annotation for genomic, protein and clinical researchers. The genomic track hubs are downloadable from the UniProt FTP site and discoverable as public track hubs at the UCSC and Ensembl genome browsers.

Introduction

Genomic variants may cause deleterious effects through many mechanisms, including aberrant gene transcription and splicing, disruption of protein translation, and altered protein structure and function. Understanding the potential effects of non-synonymous (missense) single nucleotide polymorphisms (SNPs) on protein function is key for clinical interpretation (MacArthur et al., 2014; Richards et al., 2015), but this information is not readily available. Standard next-generation sequencing (NGS) annotation tools identify simple issues like translation stops and frameshift mutations but connecting missense variants to more subtle effects like the potential disruption of enzymatic active sites, protein-binding sites, or post-translational modification sites within proteins is not easily achieved. UniProtKB represents decades of effort in capturing protein function, catalytic activity, cofactors, pathway information, subcellular location, protein-protein interactions, patterns of expression, and phenotypes or diseases associated with protein-altering variants using literature-based and semi-automated expert curation (Famiglietti et al., 2014). Included in this information are positional annotations such as enzyme active sites; modified residues; binding domains; protein isoforms; glycosylation sites; protein variations and more. Tools, such as PolyPhen-2 (Adzhubei, Jordan, & Sunyaev, 2013), already use a small subset of the protein annotations from UniProtKB, such as active sites, to give a predicted effect of missense variants. These variant effect prediction algorithms are now routinely used to aid in the interpretation of variants, such as in Ensembl's Variant Effect Predictor (McLaren et al., 2016) and CADD (Kircher et al., 2014). There is a continuing effort to improve the accuracy of such tools (Ioannidis et al., 2016; Shihab et al., 2013) but the accuracy of these algorithms has to be questioned when using their predictions to interpret the effect of a variant (Mahmood et al., 2017). Therefore, aligning curated protein functional information with genomic annotation and making it seamlessly available to the genomic, proteomic and clinical communities should greatly inform studies on the functional of variants. This is possible within genome browsers (Kent et al., 2002; Yates et al., 2016) that provide an interactive graphical representation of genomic data. They utilize standard data file formats, enabling the import and integration of multiple independent studies, as well as an individual's users own data through community track hubs, (Raney et al., 2014). Thus, tracks hubs provide a useful means to share and integrate UniProtKB's protein annotation data for the genome research community to visually interpret. We illustrate the utility of representing protein functional annotations at the genomic level via track hubs and demonstrate how, by combining with genome annotations, this information can be used to interpret the effect of missense variants in disease-related genes and proteins using specific biological examples.

As discussed by Mahmood et al., accurate effect assignment of variants is essential for a correct clinical evaluation. UniProtKB disease associated variants come primarily from the literature. We use OMIM (OMIM, 2018) as a source for disease names and descriptions and as a means of knowing what literature we maybe missing. There are many other sources of clinically associated variants. ClinVar (Landrum et al., 2018) is a open database for the deposition of variants called via clinical genome screens; where the scientist adding variants is responsible for assigning a clinical significance class to individual variants following the ACMG clinical significance recommendation (Richards et al., 2015). As ClinVar's set of variants closely reflects the type of variants curated in UniProt we have evaluated UniProtKB variants against equivalent collocated ClinVar variants. We have analysed where there is discordance to establish how both resources can improve their clinical significance association and provide a unified effect for variation to aid clinical research.

Methods

Mapping UniProtKB protein sequences to their genes and genomic coordinates is achieved with a four phase Ensembl import and mapping pipeline. The mapping is currently only conducted for the UniProt human reference proteome with the GRC reference sequence provided by Ensembl. We summarize the approach here, additional details, figures and references are provided in the Supplemental Methods document.

Phase One: Mapping Ensembl identifiers and translations to UniProtKB protein sequences.

UniProt imports Ensembl translated sequences and associated identifiers, including gene symbols and the HGNC identifier. An Ensembl translation is mapped to a UniProtKB sequence only if the Ensembl translated sequence is 100% identical to the UniProtKB sequence with no insertions or deletions. When an Ensembl translation does not match an existing UniProtKB canonical sequence or an isoform in a UniProtKB/Swiss-Prot entry, the Ensembl translation is added as a new UniProtKB/TrEMBL entry. The 100% identity threshold to an Ensembl sequence is strict and means that a relatively small number of important, well annotated proteins in UniProtKB/Swiss-Prot are not currently mapped to the genome (~3%). We are developing processes to map some of these and will make them available in future releases, others like some specific immune system alleles may not be mapped.

Phase Two: Calculation of UniProt genomic coordinates

Given the UniProt to Ensembl mapping, UniProt imports the genomic coordinates of every gene and the exons within a gene. Included are the 3' and 5' UTR offsets in the translation and exon splice phasing. With this collated coordinate data, UniProt calculates the portion of the protein sequence in each exon and defines the genomic coordinate for the amino acids at the beginning and end of each exon. This set of peptide fragments with exon identifiers and coordinates is stored as the basis for protein to genomic mappings in UniProt.

Phase Three: Converting UniProt position annotations to their genomic coordinates: UniProtKB sequence position annotations or “features” have either a single amino acid location or amino acid range within the UniProtKB canonical protein sequence. Using the exon coordinates of the protein peptide fragments, the genomic coordinates of a feature annotation are calculated by finding the amide (N) terminal exon and the carboxyl (C) terminal exon. All positions are mapped to the first or last position in an amino acid codon. Details and figure in Supplemental Methods.

Phase Four: UniProt BED and BigBed Files: Converting protein functional information into its genomic equivalent requires standardized formats. The Browser Extensible Data (BED) (UCSC, 2016a), a tab-delimited format, represents one format for displaying UniProtKB protein annotations on a genome browser. The binary equivalent of the BED file is BigBed (Kent, Zweig, Barber, Hinrichs, & Karolchik, 2010); this format is more flexible in allowing additional data elements, providing a greater opportunity to fully represent protein annotations and is one of the file formats used to make track hubs. A track hub is a web-accessible directory of files that can be displayed in track hub-enabled genome browsers (Raney et al., 2014). Hubs are useful, as users only need the hub URL to load all the data into the genome browser. Moreover, a public registry for track hubs is now available (https://trackhubregistry.org/) allowing users to search for track hubs in one location and providing links to multiple genome browsers.

Using the protein genomic coordinates, with additional protein feature specific annotations from UniProtKB, the BED detail (UCSC, 2016b) and BigBed formatted files, as well as track hub required files, are produced for the UniProtKB human reference proteome. Genomic coordinates are converted to the zero-based coordinates used by genome browsers.

Mapping ClinVar SNPs to protein features and variants: Data for comparing ClinVar SNPs to UniProt features comes from the ClinVar variant_summary.txt file on the NCBI FTP site, the UniProtKB feature specific BED files and the human variation file humsavar.txt on the UniProt FTP site. 1) For each feature in UniProtKB, we check the genomic position against the position for each SNP in ClinVar. If the genome positions of the protein feature overlap the chromosome and genomic coordinate of the SNP we establish a mapping. Information about the SNP and the feature, including the amino acid change are attached to the mapping file. 2) For each result in 1, we check the SNP position against the exon boundary for the protein. A flag is added if a SNP coordinate is within the exon boundary. Variants outside of exons were excluded from further analysis here. 3) For each UniProt variant in 2, we check the ClinVar RefSeq and UniProt accession numbers and check that the amino acid change reported in UniProt and ClinVar is the same.

Comparison of UniProt and ClinVar Variant Annotation: UniProt classifies variants into three categories: 1) Disease - variants reported to be implicated in disease; 2) Polymorphism - variants not reported to be implicated in disease; 3) Unclassified - variants with uncertain implication in disease as evidence for or against a pathogenic role is limited, or reports are conflicting. ClinVar does not annotate variants directly but accepts submitters assertions of clinical significance with their criteria and classifies them into 0-4 gold star groups based on levels of evidence. The predominant assertions in ClinVar and the ones we used for comparison are those recommended by the ACMG/AMP guidelines (Richards et al., 2015): Benign, Likely benign, Uncertain significance, Likely pathogenic and Pathogenic. In addition, there are a small number of disease related assertions such as ‘risk factor’ and ‘drug response’. For our comparison we only used variants with 1-4 stars but removed all 1-star variants with conflicting interpretations and those with no associated phenotype. We compared ClinVar assertions to UniProt classifications as follows: all ‘pathogenic’ assertions (pathogenic and likely pathogenic) to ‘Disease’ in UniProt; ‘Uncertain significance’ with ‘Unclassified’; and, all ‘benign’ (benign and likely benign) assertions to ‘Polymorphism’. Anything else in ClinVar was grouped as ‘other’ and were ‘drug response’ assertions.

Results

Mapping UniProtKB human protein annotations to the Genome Reference (GRC): Functional positional annotations from the UniProt human reference proteome are now being mapped to the corresponding genomic coordinates on the GRCh38 version of the human genome for each release of UniProt. These mappings are available as BED files or as part of a UniProt genomic track hub and can be downloaded from the UniProt FTP site (www.uniprot.org/downloads) or directly discoverable as a public track hub in the UCSC and Ensembl genome browsers by searching for UniProt and has been registered with the track hubs registry (trackhubregistry.org/search?q=uniprot&type=genomics). which provides links to load the tracks in either browser.

For the UniProt 2018_01 release, the locations of 112,093 human reference protein sequences from UniProtKB were mapped to the GRCh38. This includes the UniProtKB/Swiss-Prot 18,687 canonical and 14,783 isoform sequences and UniProtKB/TrEMBL 56,363 sequences. Thirty-four different positional annotation types (e.g. active sites, modified residues, domains and amino acid variants) with associated curated information from the literature are currently aligned with the genome sequence. **Table 1** shows the full list of positional annotations (features) and the number of each feature currently mapped to the human genome. Some proteins and features map to multiple locations, as some gene families are indistinguishable at the protein level, and others map to chromosomal regions with multiple alternative assemblies which are included. Examples of these duplicate mappings include histones, MHC proteins and the pseudoautosomal regions on both the X and Y chromosome (Helena Mangs & Morris, 2007; Veerappa, Padakannaya, & Ramachandra, 2013).

We recommend using track hubs and not the BED text files on genome browsers. The extended BED format is used to create the BigBed files. This extended BED format is not supported in a consistent manner on all browsers but BigBed files are supported and provide enhanced functionality. When using the track hubs at the UCSC or ENSEMBL browsers, ten feature tracks are turned on by default but the rest can be enabled in the browser controls. The BED tab-delimited files are useful to extract genome locations and annotation for data integration and computational analysis similar to that describe below for mapping to ClinVar SNPs.

Biological Examples: To illustrate the utility of combining UniProt protein positional (feature) and variation annotations to quickly determine a probable mechanism of action, we looked at two well-studied disease-associated proteins. The alpha-galactosidase A (*GLA*) gene (HGNC:4296, UniProtKB P06280) has been linked to Fabry disease (FD) (OMIM:301500) (Romeo & Migeon, 1970; Schiffmann, 2009), a rare X-linked lysosomal storage disease where glycosphingolipid catabolism fails and glycolipids accumulate in many tissues from birth. Many of the protein-altering variants in *GLA* are associated with FD, with UniProt biocurators recording 226 manually reviewed variants, 220 SNPs and 6 deletions, in *GLA*; of which, 219 are associated with FD. ClinVar has 193 SNPs and 34 deletions associated with FD and *GLA*; of which, 155 have an assertion of either pathogenic or likely pathogenic. 64 SNPs are identical between UniProt and ClinVar in that they cause the same amino acid change. These variants are distributed evenly over the entire protein sequence, indicating that disruption of the wild type protein sequence by any of these variants results in alteration or loss of function, which leads to FD. **Figure 1** shows a portion of exon 5 (ENSE00003671052;X:101,398,946-101,398,785) of the *GLA* gene on the UCSC genome browser. **Panel 1 selection A** illustrates a situation where an amino acid component of the enzyme's active site overlaps variant P06280:p.Asp231Asn (uniprot.org/uniprot/P06280#VAR_000468) where the acidic proton donor aspartic acid (Asp, D) is replaced by the neutral asparagine (Asn, N) suggesting it no longer functions properly as a proton donor in the active site. This missense variant is associated with Fabry disease in the literature (Redonnet-Vernhet et al., 1996). **Selection B** shows a UniProt annotated cysteine (Cys, C) residue involved in one of the five disulfide bonds present in the mature protein. The Cys aligns with an amino acid variant P06280:p.Cys223Gly (uniprot.org/uniprot/P06280 #VAR_012401) associated with Fabry disease in the literature (Germain & Poenaru, 1999). P06280:p.Cys223Gly is a missense variant that converts a cysteine to a glycine resulting in the loss of the wild-type disulfide bond between a β-strand and the C-terminal end of an α-helix encoded by exons four and five. Disruption of a disulfide bond would disrupt the structure of the alpha-galactosidase A protein and is an obvious mechanism of action for the pathogenicity of this variant. Currently the P06280:p.Cys223Gly variant is not annotated in ClinVar or dbSNP. Seven of the ten cysteines involved in the five disulfide bonds in alpha-galactosidase A have annotated variations that remove a cysteine and are associated with the disease. **Panel 2 selection C** shows an N-linked glycosylation site overlapping multiple variants annotated in ClinVar as a pathogenic allele (www.ncbi.nlm.nih.gov/clinvar/variation/10730/) and in UniProtKB as associated with Fabry disease P06280:p.N215S (uniprot.org/uniprot/P06280#VAR_000464). These missense variants all modify the required asparagine (Asn, N) to another amino acid. The P06280:p.Asn215Ser variant has been described in the literature multiple times and, in ClinVar, multiple submitters have annotated this variant as pathogenic, citing at least ten publications describing this variant and others seen in patients and family pedigrees with Fabry disease. UniProtKB has five publications, some the same as ClinVars list and others unique. However, since UniProtKB's focus is on protein structural and functional annotation, the publications that review the molecular details on the function of glycosylation at this and other sites are annotated in UniProtKB (Chen et al., 2009; Garman & Garboczi, 2004) and not in ClinVar. Evidence here shows that the oligomannose-containing carbohydrate at this Asn215 site plus the Asn192 site (not shown) are responsible for secretion of the active enzyme (Ioannou, Zeidner, Grace, & Desnick, 1998) and targeting to the lysosome

(Ghosh, Dahms, & Kornfeld, 2003). Mutation of Asn215 to Ser eliminates the carbohydrate attachment site, causing inefficient trafficking of the enzyme to the lysosome. The treatment for Fabry disease is enzyme replacement therapy using recombinant alpha-galactosidase A proteins produced in mammalian cell lines that differ in their glycosylation profile. Since all use the mannose-6-phosphate receptor pathway to enter the cells, the difference in glycoforms at these sites also explains the pharmacological differences between the Replagal (human cell line) and Fabrazyme (CHO cell lines) protein preparations used for treatment (Lee et al., 2003).

Figure 2 shows a portion of the *APP* gene that encodes the amyloid beta A4 protein (UniProtKB P05067) on the Ensembl genome browser. Amyloid beta A4 protein is a cell surface receptor with multiple functions related to neurite growth. Peptides derived from cleavage of the mature protein are found in amyloid plaques in brain tissue from patients with Alzheimer's disease (Ancolio et al., 1999; Chartier-Harlin et al., 1991; Cras et al., 1998; Denman, Rosenzcwaig, & Miller, 1993; Eckman et al., 1997; Goate et al., 1991; Hendriks et al., 1992; Kwok et al., 2000; Liepnieks, Ghetti, Farlow, Roses, & Benson, 1993; Mullan et al., 1992; Murrell, Farlow, Ghetti, & Benson, 1991; Nilsberth et al., 2001). Cleavage into multiple chains and peptides occurs at eight cleavage sites by α-, β-, γ- and θ- secretase and various caspase enzymes. Non-amyloid forming proteolysis occurs with α and γ secretases resulting in alpha species sAPPβ and a C-terminal fragment (αCTF). **Figure 3** shows two γ-secretase cleavage sites between positions 711-712 and 713-714 on the protein responsible for producing the N terminals of β-APP42 and β-APP40 peptides. The β-APP40 form aggregates to generate amyloid beta filaments that are the major component of the toxic amyloid plaques found in the brains of Alzheimer's disease (AD) sufferers (Selkoe, 1998). Disease-associated variants annotated in UniProtKB and ClinVar align with the cleavage sites, suggesting that alteration of cleavage can affect disease progression. The associated annotation shows that variant P05067:p.Thr714Ile (uniprot.org/uniprot/P05067#VAR_014218) was found in a family showing autosomal dominant inheritance of early-onset Alzheimer's disease. The variation resulted in an ~11-fold increase in the β-APP42/β-APP40 ratio *in vitro* as measured by multiple methods and coincided with deposition of nonfibrillar pre-amyloid plaques composed primarily of N-truncated β-APP42 in the brain (Kumar-Singh et al., 2000).

UniProt Features containing ClinVar SNPs: As the examples above illustrate, a missense variant in a key functional region (feature) of a protein may alter a proteins structure and function. Alternatively a feature might be so important that few variations can be tolerated. To get an overview of what might be happening in different functional features we conducted a larger analysis. First we examined SNPs from ClinVar that overlap selected protein features. Overlap in this context means that a SNP overlaps any of the three nucleotides in an amino acid codon that is part of the feature. For this comparison, we grouped the five category ACMG/AMP assertions into three 1) pathogenic, 2) uncertain significance, and 3) benign and only 2-4 gold stars and selected 1 star rated ClinVar SNPs were used. (see methods). **Figure 3** plots the percentage of ClinVar SNPs in each annotation category that exist in each feature type (Original data in Supplemental Methods). Excluding UniProtKB/Swiss-Prot Natural Variant annotations, four features have more pathogenic variants than either benign or uncertain (Initiator Methionine, Intramembrane Region, DNA Binding Domain, Active Site) and three more have pathogenic classified variants greater than or equal to benign classified SNPs but less than variants of uncertain significance classification (Nucleotide binding region, Lipid attachment site, Cross Link attachment site). Of these seven features, six are single amino acid features; where, it appears, many changes may not be tolerated (Note: 98% of UniProtKB/Swiss-Prot reviewed natural variants are single amino acid features). Initiator methionine, which refers to the removal of the initiator methionine, has the highest proportion of pathogenic variants. In the reported 36 SNPs 67% are pathogenic, 33% uncertain and 0% benign; indicating that this type of variant results in the loss of translation of protein. In comparison, variants that collocate at carbohydrate/glycosylation sites are tolerated the most with less than 10% pathogenic assertions. The next two types of features affected most severely by a SNP are the intramembrane and transmembrane features, Figure 3. These features describe a sequence of amino acids entirely within a membrane but not crossing it or spanning the extent of the membrane, respectively. A recent comparison of 10,000 human genomes (Telenti et al., 2016) analysed transmembrane regions in a few thousand proteins, and observed less variation than other regions, suggesting a structural reason for amino acid conservation in those regions.

Comparison of ClinVar SNPs and UniProtKB Natural Variant Annotation: To survey genomic and protein annotations on variation we compared ClinVar SNPs with UniProtKB natural amino acid variations. Currently 24,585 UniProt variants collocate on the genome with ClinVar SNPs, (0-4 stars), which is 31.6% of all UniProtKB variants and 8% of all ClinVar SNP variants. Currently 53% of UniProtKB disease variants exist in ClinVar and 24.5% of ClinVar SNPs with pathogenic assertions are present as amino acid variants in UniProtKB. **Table 2** shows a comparison between the UniProt variant classification of Disease, Unclassified and Polymorphism with a combination of ClinVar's ACMG/AMP based assertions of 'Pathogenic or likely pathogenic', 'Uncertain Significance' and 'Benign or likely benign'. This comparison only included SNPs with ClinVar's 1-4 gold star evidence levels but removed any 1-star variants with conflicting interpretations and any that did not report a disease or phenotype in the record. This subset was 35% of the total variants mapped. The table shows there is general agreement among similar annotations between the databases, with 86% of UniProtKB disease associated variants mapping to 'pathogenic' SNPs in ClinVar and with 10% falling into the middle 'Uncertain Significance' category. The remaining 4% fall mainly into the benign category. UniProt's 'Polymorphism' category is closest in meaning to the 'Benign' categories in ClinVar; here, again, there is 85% agreement. For the remaining 15% 'Polymorphism' variants 11% match the 'Uncertain Significance' category in ClinVar, 3% are classified as 'pathogenic' in ClinVar and 1% as 'drug response'. UniProt's 'Unclassified' category is closest in meaning to ClinVar's 'Uncertain Significance'; which are "grey" areas in each classification system and as such the agreement between the two databases is lowest with, in comparison to the previous two classifications, only 54% aligning and the rest split between 'pathogenic' and 'benign' in ClinVar. The large number of variants annotated with 'Uncertain Significance' status is currently a general problem in the field (Hoffman-Andrews, 2017). In the ACMG/AMP framework, uncertain occupies a middle ground between benign and pathogenic, often there is some evidence of a functional defect or harmful effect but it does not rise to clinical relevance or there is conflicting evidence.

Though annotations in UniProt and ClinVar are in general agreement, there is still a significant level of disagreement between the databases, which is similar to that seen in recent analyses that compared variant pathogenicity interpretations by several laboratories within the ClinGen framework (Amendola et al., 2016; Gelb et al., 2018). These discrepancies may arise when variants have been assessed at different times, for different populations, and using different data types; pathogenicity may sometimes be inferred from functional assays. Protein curators traditionally looked more at functional biochemical evidence that related function to a disease through evidence of genetic inheritance. In comparison, medical geneticists put more weight on genetic studies, variant frequencies, penetrance and increasingly statistical models (InSiGHT; Plon et al., 2008) or variant classification. A detailed evaluation of these reasons will be the subject of a forthcoming study in which UniProt curators are performing a systematic re-curation of a randomly chosen set of variants from UniProt and ClinVar (in preparation Famiglietti et al.).

Comparisons can get even more complicated when clinically relevant somatic mutations in cancer are considered which may not formally be pathogenic but are relevant to the efficacy of any treatment and need their own curation guidelines (Boca, 2018; Li et al., 2017; Madhavan, 2018; Ritter et al., 2016). Pharmagenomic variants also may get different classifications. An example of this would be Q9BY32:p.Pro32Thr (uniprot.org/uniprot/Q9BY32#VAR_015576) a variation in the ITPA gene classified as a ‘disease’ in UniProt due to its association with heritable inosine triphosphatase deficiency in the literature with a notation that “It might have pharmacogenomic implications and be related to increased drug toxicity of purine analog drugs”. In ClinVar the same variation NM_033453.3(ITPA):c.94C>A (www.ncbi.nlm.nih.gov/clinvar/variation/14746/) is annotated by an expert panel at PharmaGKB (Whirl-Carrillo et al., 2012) as “drug, response” citing literature on the variants effects on some antiviral drugs (Azakami et al., 2011; Chayama et al., 2011) only. Whilst the notation is different, both annotations are correct from the publications cited and each group’s area of interest.

Conclusion

The combination of protein and genome annotation is made easier by accurate mapping between the two and allows for more detailed analysis on the effects of variation on protein function. Exome sequencing for clinical diagnosis is becoming more common but usually uncovers many non-synonymous SNP variations of unknown significance (VUS). Distinguishing which, if any, of these variants could be causal is difficult. Protein annotation can aid in variant curation by providing a functional explanation for a variant’s effect. The results reported here suggest that the location of a variant within some functional features may be related with pathogenicity. However, this assertion requires more rigorous statistical analysis and possibly a structural analysis. We plan to investigate this topic in the future and expect that others may do the same now that this data is more readily available.

UniProtKB features have been mapped to the genome before, as the UCSC genome browser has provided selected UniProtKB/Swiss-Prot features for several years. The mappings described here contain additional annotation beyond that previously available, including isoform sequences from Swiss-Prot and also sequences and features from the TrEMBL section of UniProtKB. UniProtKB/TrEMBL contains additional isoforms and predicted genes not currently in UniProtKB/Swiss-Prot, along with automated feature annotations from a combination of prediction algorithms and human curated rules (Dimmer et al., 2012; Famiglietti et al., 2014; Pedruzzi et al., 2015; Vasudevan et al., 2010). The data files and track hubs described here will be updated with each release of UniProtKB, making any new annotation available immediately. The 34 features currently provided are not all the positional annotations in UniProtKB and we may add additional features in future releases. We plan to extend genome mapping to other model organisms starting with Yeast in the future. UniProt is working with the UCSC and Ensembl browser teams to improve the presentation of protein annotation on the respective browsers. In addition, some of the data provided here are available programmatically via a REST API (Nightingale et al., 2017).

UniProt collaborates with ClinVar to provide reciprocal links between variants that exist in both databases. We have used these links to identify variants where the interpretation of clinical significance differs between UniProt and ClinVar and are currently performing a more detailed investigation of the underlying reasons for annotation differences using a more extensive set of variants (Famiglietti et al., manuscript in preparation). Recent efforts in the medical community to standardize the methods and levels of evidence required for the annotation of genetic variants (Amr et al., 2016; Manrai et al., 2016; O'Daniel et al., 2016; Richards et al., 2015; Walsh et al., 2016), along with increasing amounts of population data (Amr et al., 2016; Walsh et al., 2016), are leading to the widespread re-evaluation of previous assertions of pathogenicity. The work described here provides the basis for a re-evaluation of UniProt annotation and the further standardization of this annotation with ClinVar and ClinGen.

In summary, these new genomic mappings should help better integrate protein and genomic analyses and should improve interoperability between the genomic and proteomic communities to better determine the functional effects of genome variation on proteins. The location of a variant within functional features may correlate with pathogenicity and would be a useful attribute for use in variant prediction algorithms including machine-learning approaches.

Data access

The extended BED text files and binary BigBed files used for genome Track Hubs are located in the UniProt FTP site in the genome_annotation_tracks directory (uniprot.org/downloads). Public Track hubs are available at the UCSC genome browser (Tyner et al., 2016) at (genome.ucsc.edu/cgi-bin/hgHubConnect?hubSearchTerms=uniprot&hubDbFilter=hg38) and the Ensembl genome browser (Aken et al., 2016; Hubbard et al., 2007) via a track hub registry search (ensembl.org/Homo_sapiens/Info/Index#modal_user_data-Track_Hub_Registry_Search) for “UniProt”. Both can be found by a Track Hub Registry search, (trackhubregistry.org/search?q=uniprot&type=genomics) which provides links to load the tracks at either browser. All files are updated with each release of UniProtKB.
Links from trackhubregistry.org that load the default UniProt tracks automatically are shown below. Additional tracks can be selected for display on each browser.
UCSC Browser:
genome.ucsc.edu/cgi-bin/hgHubConnect?db=hg38&hubUrl=ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/genome_annotation_tracks/UP000005640_9606_hub/hub.txt&hgHub_do_redirect=on&hgHubConnect.remakeTrackHub=on

UniProt Genomic Mapping

Ensembl Browser:
www.ensembl.org/TrackHub?url=ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/genome_annotation_tracks/UP000005640_9606_hub/hub.txt;species=Homo_sapiens;name=UniProt_Features;registry=1

Acknowledgments

Funding: NIH grants U41HG007822, U41HG007822-02S1 for UniProt and U01HG007437.for ClinGen

References

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet, Chapter 7*, Unit7 20. doi:10.1002/0471142905.hg0720s76

Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., . . . Flicek, P. (2016). Ensembl 2017. *Nucleic Acids Res*. doi:10.1093/nar/gkw1104

Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., . . . Rehm, H. L. (2016). Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am J Hum Genet, 98*(6), 1067-1076. doi:10.1016/j.ajhg.2016.03.024

Amr, S. S., Al Turki, S. H., Lebo, M., Sarmady, M., Rehm, H. L., & Abou Tayoun, A. N. (2016). Using large sequencing data sets to refine intragenic disease regions and prioritize clinical variant interpretation. *Genet Med*. doi:10.1038/gim.2016.134

Ancolio, K., Dumanchin, C., Barelli, H., Warter, J. M., Brice, A., Campion, D., . . . Checler, F. (1999). Unusual phenotypic alteration of beta amyloid precursor protein (betaAPP) maturation by a new Val-715 --> Met betaAPP-770 mutation responsible for probable early-onset Alzheimer's disease. *Proc Natl Acad Sci U S A, 96*(7), 4119-4124.

Azakami, T., Hayes, C. N., Sezaki, H., Kobayashi, M., Akuta, N., Suzuki, F., . . . Chayama, K. (2011). Common genetic polymorphism of ITPA gene affects ribavirin-induced anemia and effect of peg-interferon plus ribavirin therapy. *J Med Virol, 83*(6), 1048-1057. doi:10.1002/jmv.22069

Boca, S. M., Panagiotou, O. A., Rao, S., McGarvey, P. B., Madhavan, S. (2018). Future of Evidence Synthesis in Precision Oncology: Between Systematic Reviews and Biocuration. *JCO Precision Oncology, In Press*.

Chartier-Harlin, M. C., Crawford, F., Houlden, H., Warren, A., Hughes, D., Fidani, L., . . . et al. (1991). Early-onset Alzheimer's disease caused by mutations at codon 717 of the beta-amyloid precursor protein gene. *Nature, 353*(6347), 844-846. doi:10.1038/353844a0

Chayama, K., Hayes, C. N., Abe, H., Miki, D., Ochi, H., Karino, Y., . . . Kumada, H. (2011). IL28B but not ITPA polymorphism is predictive of response to pegylated interferon, ribavirin, and telaprevir triple therapy in patients with genotype 1 hepatitis C. *J Infect Dis, 204*(1), 84-93. doi:10.1093/infdis/jir210

Chen, R., Jiang, X., Sun, D., Han, G., Wang, F., Ye, M., . . . Zou, H. (2009). Glycoproteomics analysis of human liver tissue by combination of multiple enzyme digestion and hydrazide chemistry. *J Proteome Res, 8*(2), 651-661. doi:10.1021/pr8008012

Cras, P., van Harskamp, F., Hendriks, L., Ceuterick, C., van Duijn, C. M., Stefanko, S. Z., . . . Martin, J. J. (1998). Presenile Alzheimer dementia characterized by amyloid angiopathy and large amyloid core type senile plaques in the APP 692Ala-->Gly mutation. *Acta Neuropathol, 96*(3), 253-260.

Denman, R. B., Rosenzcwaig, R., & Miller, D. L. (1993). A system for studying the effect(s) of familial Alzheimer disease mutations on the processing of the beta-amyloid peptide precursor. *Biochem Biophys Res Commun, 192*(1), 96-103. doi:10.1006/bbrc.1993.1386

Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J., . . . Apweiler, R. (2012). The UniProt-GO Annotation database in 2011. *Nucleic Acids Res, 40*(Database issue), D565-570. doi:10.1093/nar/gkr1048

Eckman, C. B., Mehta, N. D., Crook, R., Perez-tur, J., Prihar, G., Pfeiffer, E., . . . Hardy, J. (1997). A new pathogenic mutation in the APP gene (I716V) increases the relative proportion of A beta 42(43). *Hum Mol Genet, 6*(12), 2087-2089.

Famiglietti, M. L., Estreicher, A., Gos, A., Bolleman, J., Gehant, S., Breuza, L., . . . UniProt, C. (2014). Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum Mutat, 35*(8), 927-935. doi:10.1002/humu.22594

Garman, S. C., & Garboczi, D. N. (2004). The molecular defect leading to Fabry disease: structure of human alpha-galactosidase. *J Mol Biol, 337*(2), 319-335. doi:10.1016/j.jmb.2004.01.035

Gelb, B. D., Cave, H., Dillon, M. W., Gripp, K. W., Lee, J. A., Mason-Suares, H., . . . Vincent, L. M. (2018). ClinGen's RASopathy Expert Panel consensus methods for variant interpretation. *Genet Med*. doi:10.1038/gim.2018.3

Germain, D. P., & Poenaru, L. (1999). Fabry disease: identification of novel alpha-galactosidase A mutations and molecular carrier detection by use of fluorescent chemical cleavage of mismatches. *Biochem Biophys Res Commun, 257*(3), 708-713. doi:10.1006/bbrc.1999.0310

Ghosh, P., Dahms, N. M., & Kornfeld, S. (2003). Mannose 6-phosphate receptors: new twists in the tale. *Nat Rev Mol Cell Biol, 4*(3), 202-212. doi:10.1038/nrm1050

Goate, A., Chartier-Harlin, M. C., Mullan, M., Brown, J., Crawford, F., Fidani, L., . . . et al. (1991). Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature, 349*(6311), 704-706. doi:10.1038/349704a0

Helena Mangs, A., & Morris, B. J. (2007). The Human Pseudoautosomal Region (PAR): Origin, Function and Future. *Curr Genomics, 8*(2), 129-136.

Hendriks, L., van Duijn, C. M., Cras, P., Cruts, M., Van Hul, W., van Harskamp, F., . . . et al. (1992). Presenile dementia and cerebral haemorrhage linked to a mutation at codon 692 of the beta-amyloid precursor protein gene. *Nat Genet, 1*(3), 218-221. doi:10.1038/ng0692-218

Hoffman-Andrews, L. (2017). The known unknown: the challenges of genetic variants of uncertain significance in clinical practice. *Journal of Law and the Biosciences, 4*(3), 648-657. doi:10.1093/jlb/lsx038

Hubbard, T. J., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., . . . Birney, E. (2007). Ensembl 2007. *Nucleic Acids Res, 35*(Database issue), D610-617. doi:10.1093/nar/gkl996

InSiGHT. InSiGHT Variant Interpretation Committee: Mismatch Repair Gene Variant Classification Criteria. Retrieved from https://www.insight-group.org/content/uploads/2017/05/2013-08_InSiGHT_VIC_v1.9.pdf

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., . . . Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet, 99*(4), 877-885. doi:10.1016/j.ajhg.2016.08.016

Ioannou, Y. A., Zeidner, K. M., Grace, M. E., & Desnick, R. J. (1998). Human alpha-galactosidase A: glycosylation site 3 is essential for enzyme solubility. *Biochem J, 332 (Pt 3)*, 789-797.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res, 12*(6), 996-1006. doi:10.1101/gr.229102. Article published online before print in May 2002

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics, 26*(17), 2204-2207. doi:10.1093/bioinformatics/btq351

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46*(3), 310-315. doi:10.1038/ng.2892

Kumar-Singh, S., De Jonghe, C., Cruts, M., Kleinert, R., Wang, R., Mercken, M., . . . Van Broeckhoven, C. (2000). Nonfibrillar diffuse amyloid deposition due to a gamma(42)-secretase site mutation points to an essential role for N-truncated A beta(42) in Alzheimer's disease. *Hum Mol Genet, 9*(18), 2589-2598.

Kwok, J. B., Li, Q. X., Hallupp, M., Whyte, S., Ames, D., Beyreuther, K., . . . Schofield, P. R. (2000). Novel Leu723Pro amyloid precursor protein mutation increases amyloid beta42(43) peptide levels and induces apoptosis. *Ann Neurol, 47*(2), 249-253.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., . . . Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res, 46*(D1), D1062-d1067. doi:10.1093/nar/gkx1153

Lee, K., Jin, X., Zhang, K., Copertino, L., Andrews, L., Baker-Malcolm, J., . . . Edmunds, T. (2003). A biochemical and pharmacological comparison of enzyme replacement therapies for the glycolipid storage disorder Fabry disease. *Glycobiology, 13*(4), 305-313. doi:10.1093/glycob/cwg034

Li, M. M., Datto, M., Duncavage, E. J., Kulkarni, S., Lindeman, N. I., Roy, S., . . . Nikiforova, M. N. (2017). Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn, 19*(1), 4-23. doi:10.1016/j.jmoldx.2016.10.002

Liepnieks, J. J., Ghetti, B., Farlow, M., Roses, A. D., & Benson, M. D. (1993). Characterization of amyloid fibril beta-peptide in familial Alzheimer's disease with APP717 mutations. *Biochem Biophys Res Commun, 197*(2), 386-392.

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., . . . Gunter, C. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature, 508*(7497), 469-476. doi:10.1038/nature13127

Madhavan, S., Ritter, D., Micheel, C., Rao, S., Roy, A., Sonkin, D., Sidiropoulos, N., McCoy, M., Griffith, O., Griffith, M., Mcgarvey, P., Kulkarni, S. On Behalf of The Clingen Somatic Working Group. (2018). Clingen Somatic Working Group – Standardizing And Democratizing Access To Cancer Molecular Diagnostic Data To Drive Translational Research. *Pacific Symposium on Biocomputing*.

Mahmood, K., Jung, C. H., Philip, G., Georgeson, P., Chung, J., Pope, B. J., & Park, D. J. (2017). Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum Genomics, 11*(1), 10. doi:10.1186/s40246-017-0104-8

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., . . . Kohane, I. S. (2016). Genetic Misdiagnoses and the Potential for Health Disparities. *N Engl J Med, 375*(7), 655-665. doi:10.1056/NEJMsa1507092

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., . . . Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol, 17*(1), 122. doi:10.1186/s13059-016-0974-4

Mullan, M., Crawford, F., Axelman, K., Houlden, H., Lilius, L., Winblad, B., & Lannfelt, L. (1992). A pathogenic mutation for probable Alzheimer's disease in the APP gene at the N-terminus of beta-amyloid. *Nat Genet, 1*(5), 345-347. doi:10.1038/ng0892-345

Murrell, J., Farlow, M., Ghetti, B., & Benson, M. D. (1991). A mutation in the amyloid precursor protein associated with hereditary Alzheimer's disease. *Science, 254*(5028), 97-99.

Nightingale, A., Antunes, R., Alpi, E., Bursteinas, B., Gonzales, L., Liu, W., . . . Martin, M. (2017). The Proteins API: accessing key integrated protein and genome information. *Nucleic Acids Res*. doi:10.1093/nar/gkx237

Nilsberth, C., Westlind-Danielsson, A., Eckman, C. B., Condron, M. M., Axelman, K., Forsell, C., . . . Lannfelt, L. (2001). The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Abeta protofibril formation. *Nat Neurosci, 4*(9), 887-893. doi:10.1038/nn0901-887

O'Daniel, J. M., McLaughlin, H. M., Amendola, L. M., Bale, S. J., Berg, J. S., Bick, D., . . . Rehm, H. L. (2016). A survey of current practices for genomic sequencing test interpretation and reporting processes in US laboratories. *Genet Med*. doi:10.1038/gim.2016.152

OMIM. (2018). Online Mendelian Inheritance in Man, OMIM®. Retrieved from http://omim.org/

Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., de Castro, E., . . . Bridge, A. (2015). HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res, 43*(Database issue), D1064-1070. doi:10.1093/nar/gku1002

Plon, S. E., Eccles, D. M., Easton, D., Foulkes, W. D., Genuardi, M., Greenblatt, M. S., . . . Tavtigian, S. (2008). Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat, 29*(11), 1282-1291. doi:10.1002/humu.20880

Raney, B. J., Dreszer, T. R., Barber, G. P., Clawson, H., Fujita, P. A., Wang, T., . . . Kent, W. J. (2014). Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics, 30*(7), 1003-1005. doi:10.1093/bioinformatics/btt637

Redonnet-Vernhet, I., Ploos van Amstel, J. K., Jansen, R. P., Wevers, R. A., Salvayre, R., & Levade, T. (1996). Uneven X inactivation in a female monozygotic twin pair with Fabry disease and discordant expression of a novel mutation in the alpha-galactosidase A gene. *J Med Genet, 33*(8), 682-688.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Rehm, H. L. (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*. doi:10.1038/gim.2015.30

Ritter, D. I., Roychowdhury, S., Roy, A., Rao, S., Landrum, M. J., Sonkin, D., . . . ClinGen Somatic Cancer Working, G. (2016). Somatic cancer variant curation and harmonization through consensus minimum variant level data. *Genome Med, 8*(1), 117. doi:10.1186/s13073-016-0367-z

Romeo, G., & Migeon, B. R. (1970). Genetic inactivation of the alpha-galactosidase locus in carriers of Fabry's disease. *Science, 170*(3954), 180-181.

Schiffmann, R. (2009). Fabry disease. *Pharmacol Ther, 122*(1), 65-77. doi:10.1016/j.pharmthera.2009.01.003

Selkoe, D. J. (1998). The cell biology of beta-amyloid precursor protein and presenilin in Alzheimer's disease. *Trends Cell Biol, 8*(11), 447-453.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., . . . Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat, 34*(1), 57-65. doi:10.1002/humu.22225

Telenti, A., Pierce, L. C. T., Biggs, W. H., di Iulio, J., Wong, E. H. M., Fabani, M. M., . . . Venter, J. C. (2016). Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences, 113*(42), 11901-11906. doi:10.1073/pnas.1613365113

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., . . . Kent, W. J. (2016). The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res*. doi:10.1093/nar/gkw1134

UCSC. (2016a). BED (Browser Extensible Data) format. Retrieved from https://genome.ucsc.edu/FAQ/FAQformat#format1

UCSC. (2016b). BED detail format. Retrieved from https://genome.ucsc.edu/FAQ/FAQformat.html#format1.7

Vasudevan, A., Vinayaka, C. R., Natale, D. A., Huang, H., Kahsay, R. Y., & Wu, C. H. (2010). *Structure-guided rule-based annotation of protein functional sites in UniProtKB*: Springer.

Veerappa, A. M., Padakannaya, P., & Ramachandra, N. B. (2013). Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. *Funct Integr Genomics, 13*(3), 285-293. doi:10.1007/s10142-013-0323-6

Walsh, R., Thomson, K. L., Ware, J. S., Funke, B. H., Woodley, J., McGuire, K. J., . . . Watkins, H. (2016). Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med*. doi:10.1038/gim.2016.90

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., . . . Klein, T. E. (2012). Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther, 92*(4), 414-417. doi:10.1038/clpt.2012.96

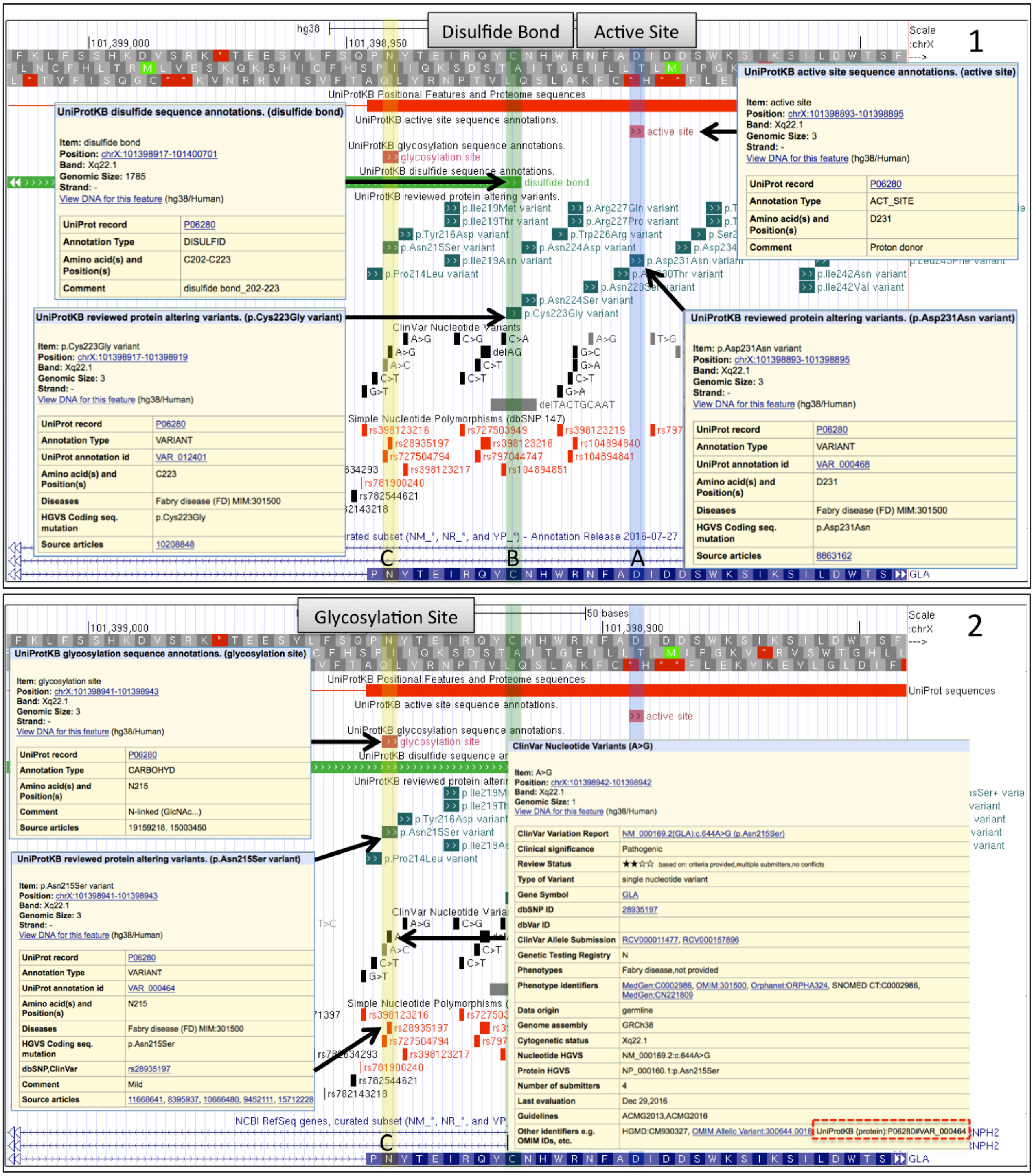
Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., . . . Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Res, 44*(D1), D710-716. doi:10.1093/nar/gkv1157

Tables and Figures

Annotation Type		Description (feature_name)	Features Mapped
Molecule Processing	Proteome	Location of complete protein and isoform sequences (proteome)	112,093
	Signal	Sequence targeting proteins to the secretory pathway or periplasmic space (signal)	10,360
	Transit peptide	Extent of a transit peptide for organelle targeting (transit)	482
	Chain	Extent of a polypeptide chain in the mature protein (chain)	25,339
	Peptide	Extent of an active peptide in the mature protein (peptide)	383
	Propeptide	Peptide that is cleaved during maturation or activation (propep)	802
	Initiator met	Cleavage of the initiator methionine (init_met)	1,992
Regions	Topological domain	Location of non-membrane regions of membrane-spanning proteins (topo_dom)	18,773
	Transmembrane	Extent of a membrane-spanning region (transmem)	43,734
	Intramembrane	Extent of a region located in a membrane without crossing it (intramem)	329
	Domain	Position and type of each modular protein domain (domain)	66,315
	Repeat	Positions of repeated sequence motifs or domains (repeat)	19,314
	Calcium binding	Position(s) of calcium binding region(s) within the protein (ca_bind)	731
	Zinc finger	Position(s) and type(s) of zinc fingers within the protein (zn_fing)	9,127
	DNA binding	Position and type of a DNA-binding domain (dna_bind)	1,267
	Nucleotide binding	Nucleotide phosphate binding region (np_bind)	3,826
	Region	Region of interest in the sequence (region)	9,894
	Coiled coil	Positions of regions of coiled coil within the protein (coiled)	16,909
	Motif	Short (up to 20 aa) sequence motif of biological interest (motif)	3,332
Sites	Active site	Amino acids directly involved in the activity of an enzyme (act_site)	4,190
	Metal binding	Binding site for a metal ion (metal)	3,031
	Binding site	Binding site for any chemical group (binding)	6,275
	Site	Any interesting single amino acid site on the sequence (site)	2,183
Amino Acid Modifications	Modified residue	Modified residues excluding lipids, glycans & cross-links (mod_res)	54,743
	Lipidation	Covalently attached lipid group(s) (lipid)	1,035
	Glycosylation	Covalently attached glycan group(s) (carbohyd)	16,474
	Disulfide bond	Cysteine residues participating in disulfide bonds (disulfid)	19,816
	Cross-link	Residues in covalent linkage between proteins (crosslnk)	6,829
	Non-standard residue	Occurence of non-standard amino acids (selenocysteine & pyrrolysine) in the protein sequence (non_std)	36
Structure	Helix	Helical regions in the experimentally determined structure (helix)	57,596
	Turn	Turns within the experimentally determined protein structure (turn)	14,813
	Beta strand	Beta strand regions within the experimentally determined protein structure (strand)	63,579
Variants	Mutagenesis	Sites experimentally altered by mutagenesis (mutagen)	20,335
	Natural variant	Description of a natural variant of the protein (variant)	76,678

Table 1: UniProtKB sequence annotations in track hubs. Annotation types, descriptions and current number of each feature mapped to the human genome are shown. UniProtKB release 2018_01 (Jan 2018) was used for this table. For more information on sequence features in UniProt see (www.uniprot.org/help/sequence_annotation).

UniProt Genomic Mapping



UniProtKB disulfide sequence annotations. (disulfide bond)

Item: disulfide bond
Position: chrX:101398917-101400701
Band: Xq22.1
Genomic Size: 1785
Strand: -
View DNA for this feature (hg38/Human)

UniProt record	P06280
Annotation Type	DISULFID
Amino acid(s) and Position(s)	C202-C223
Comment	disulfide bond_202-223

UniProtKB reviewed protein altering variants. (p.Cys223Gly variant)

Item: p.Cys223Gly variant
Position: chrX:101398917-101398919
Band: Xq22.1
Genomic Size: 3
Strand: -
View DNA for this feature (hg38/Human)

UniProt record	P06280
Annotation Type	VARIANT
UniProt annotation id	VAR_012401
Amino acid(s) and Position(s)	C223
Diseases	Fabry disease (FD) MIM:301500
HGVS Coding seq. mutation	p.Cys223Gly
Source articles	10208848

UniProtKB active site sequence annotations. (active site)

Item: active site
Position: chrX:101398893-101398895
Band: Xq22.1
Genomic Size: 3
Strand: -
View DNA for this feature (hg38/Human)

UniProt record	P06280
Annotation Type	ACT_SITE
Amino acid(s) and Position(s)	D231
Comment	Proton donor

UniProtKB reviewed protein altering variants. (p.Asp231Asn variant)

Item: p.Asp231Asn variant
Position: chrX:101398893-101398895
Band: Xq22.1
Genomic Size: 3
Strand: -
View DNA for this feature (hg38/Human)

UniProt record	P06280
Annotation Type	VARIANT
UniProt annotation id	VAR_000468
Amino acid(s) and Position(s)	D231
Diseases	Fabry disease (FD) MIM:301500
HGVS Coding seq. mutation	p.Asp231Asn
Source articles	8863162

UniProtKB glycosylation sequence annotations. (glycosylation site)

Item: glycosylation site
Position: chrX:101398941-101398943
Band: Xq22.1
Genomic Size: 3
Strand: -
View DNA for this feature (hg38/Human)

UniProt record	P06280
Annotation Type	CARBOHYD
Amino acid(s) and Position(s)	N215
Comment	N-linked (GlcNAc...)
Source articles	19159218, 15003450

UniProtKB reviewed protein altering variants. (p.Asn215Ser variant)

Item: p.Asn215Ser variant
Position: chrX:101398941-101398943
Band: Xq22.1
Genomic Size: 3
Strand: -
View DNA for this feature (hg38/Human)

UniProt record	P06280
Annotation Type	VARIANT
UniProt annotation id	VAR_000464
Amino acid(s) and Position(s)	N215
Diseases	Fabry disease (FD) MIM:301500
HGVS Coding seq. mutation	p.Asn215Ser
dbSNP,ClinVar	rs28935197
Comment	Mild
Source articles	11668641, 8395937, 10666480, 9452111, 15712228

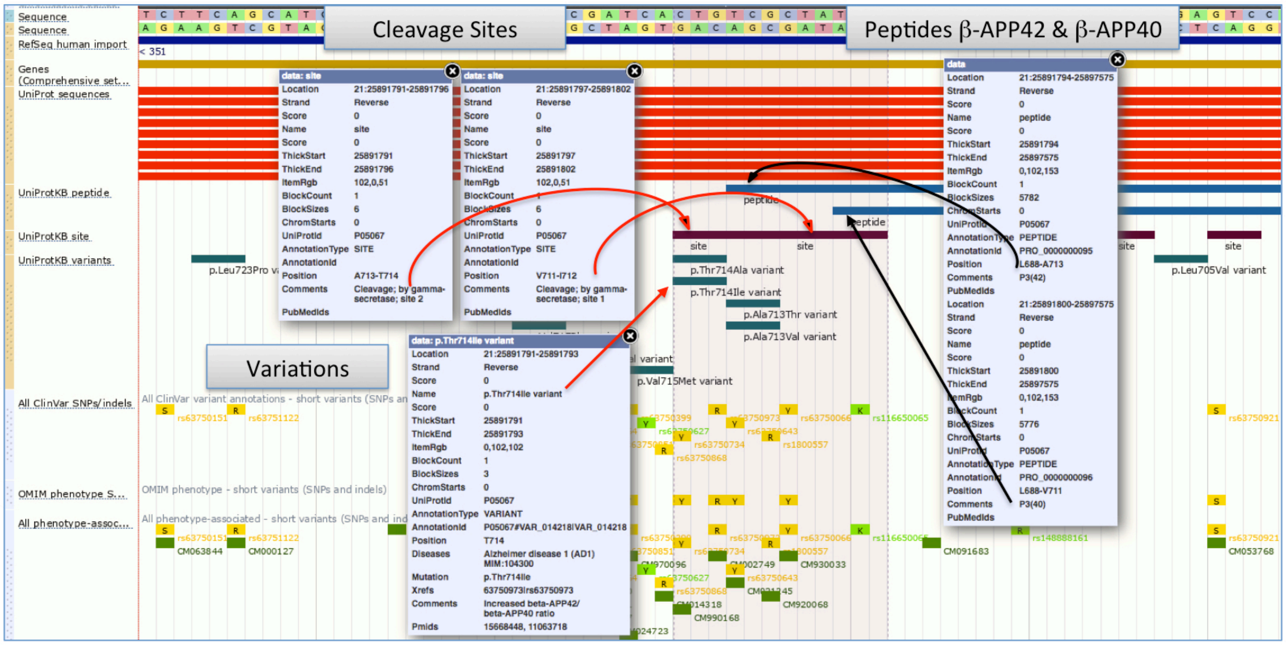
ClinVar Nucleotide Variants (A>G)

Item: A>G
Position: chrX:101398942-101398942
Band: Xq22.1
Genomic Size: 1
View DNA for this feature (hg38/Human)

ClinVar Variation Report	NM_000169.2(GLA):c.644A>G (p.Asn215Ser)
Clinical significance	Pathogenic
Review Status	★★☆☆ based on: criteria provided,multiple submitters,no conflicts
Type of Variant	single nucleotide variant
Gene Symbol	GLA
dbSNP ID	28935197
dbVar ID	
ClinVar Allele Submission	RCV000011477, RCV000157896
Genetic Testing Registry	N
Phenotypes	Fabry disease,not provided
Phenotype identifiers	MedGen:C0002986, OMIM:301500, Orphanet:ORPHA324, SNOMED CT:C0002986, MedGen:CN221809
Data origin	germline
Genome assembly	GRCh38
Cytogenetic status	Xq22.1
Nucleotide HGVS	NM_000169.2:c.644A>G
Protein HGVS	NP_000160.1:p.Asn215Ser
Number of submitters	4
Last evaluation	Dec 29,2016
Guidelines	ACMG2013,ACMG2016
Other identifiers e.g. OMIM IDs, etc.	HGMD:CM930327, OMIM Allelic Variant:300644.0018 UniProtKB (protein):P06280#VAR_000464

Figure 1. The GLA gene (P06280, alpha-galactosidase A) associated with Fabry disease (FD) shown on the UCSC browser with UniProt genome tracks plus ClinVar and dbSNP tracks. **Panel 1 selection A** shows UniProt annotation for part of the enzyme's Active Site and an amino acid variation from a SNP associated with FD that removes an acidic proton donor (Asp, D) is replaced by the neutral (Asn, N). In selection **B** another variant disrupts an annotated disulfide bond by removing a cysteine required for a structural fold. SNPs are not observed in the other data resources in these positions. **Panel 2 selection C** shows an N-linked glycosylation site disrupted by another UniProt amino acid variant that does overlap pathogenic variants in ClinVar and other databases. Links between UniProt and ClinVar are illustrated in the display.

UniProt Genomic Mapping



data: site	
Location	21:25891791-25891796
Strand	Reverse
Score	0
Name	site
Score	0
ThickStart	25891791
ThickEnd	25891796
ItemRgb	102,0,51
BlockCount	1
BlockSizes	6
ChromStarts	0
UniProtId	P05067
AnnotationType	SITE
AnnotationId	
Position	A713-T714
Comments	Cleavage; by gamma-secretase; site 2
PubMedIds	

data: site	
Location	21:25891797-25891802
Strand	Reverse
Score	0
Name	site
Score	0
ThickStart	25891797
ThickEnd	25891802
ItemRgb	102,0,51
BlockCount	1
BlockSizes	6
ChromStarts	0
UniProtId	P05067
AnnotationType	SITE
AnnotationId	
Position	V711-I712
Comments	Cleavage; by gamma-secretase; site 1
PubMedIds	

data: p.Thr714Ile variant	
Location	21:25891791-25891793
Strand	Reverse
Score	0
Name	p.Thr714Ile variant
Score	0
ThickStart	25891791
ThickEnd	25891793
ItemRgb	0,102,102
BlockCount	1
BlockSizes	3
ChromStarts	0
UniProtId	P05067
AnnotationType	VARIANT
AnnotationId	P05067#VAR_014218IVAR_014218
Position	T714
Diseases	Alzheimer disease 1 (AD1) MIM:104300
Mutation	p.Thr714Ile
Xrefs	63750973Irs63750973
Comments	Increased beta-APP42/ beta-APP40 ratio
Pmids	15668448, 11063718

data	
Location	21:25891794-25897575
Strand	Reverse
Score	0
Name	peptide
Score	0
ThickStart	25891794
ThickEnd	25897575
ItemRgb	0,102,153
BlockCount	1
BlockSizes	5782
ChromStarts	0
UniProtId	P05067
AnnotationType	PEPTIDE
AnnotationId	PRO_0000000095
Position	L688-A713
Comments	P3(42)
PubMedIds	
Location	21:25891800-25897575
Strand	Reverse
Score	0
Name	peptide
Score	0
ThickStart	25891800
ThickEnd	25897575
ItemRgb	0,102,153
BlockCount	1
BlockSizes	5776
ChromStarts	0
UniProtId	P05067
AnnotationType	PEPTIDE
AnnotationId	PRO_0000000096
Position	L688-V711
Comments	P3(40)
PubMedIds	

Figure 2. The APP gene (amyloid beta A4 protein, UniProt Acc. P05067) associated with Alzheimer's disease shown on the Ensembl browser with selected UniProt Genome Tracks. Shows how pathogenic variations overlap protein cleavage sites contributing to the aberrant ratios of the β-APP42 and β -APP40 peptides observed in some Alzheimer's cases.

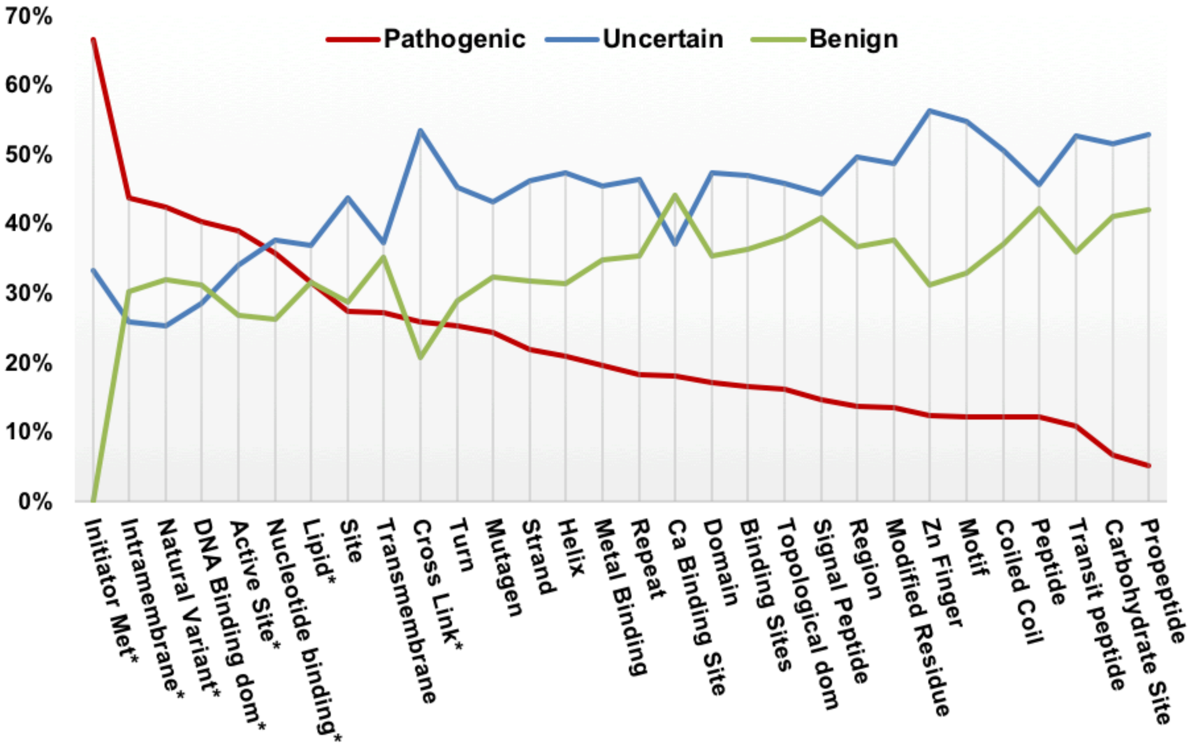


Figure 3 Percentage of ClinVar SNPs in each annotation category that exist in each feature type, underlying data table in supplemental methods. Features with " * " are discussed in the text.

	All ClinVar SNPs (249,784)	Pathogenic SNPs (27,819)	Uncertain SNPS (132,904)	Benign SNPS (94,541)	Other SNPS (11,489)
All UniProt Variants (77,647)	**8609**	3918	1291	3360	40
Disease Variants (30,220)	4135	**3562**	412	159	2
Unclassified Variants (7,579)	876	245	**476**	155	0
Polymorphism Variants (39,848)	3598	111	403	**3046**	38

Table 2. Mapping of Variants and annotation between ClinVar SNPs and UniProt amino acid variants that overlap in genome position and result in the same amino acid change. Only gold star rated ClinVar variants were included with evaluation criteria and no conflicts in assertions. Numbers in **Bold** face are key comparisons discussed in the text. Numbers in () are totals for database.