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

The integration of protein function knowledge in the UniProt Knowledgebase (UniProtKB) with genomic data provides unique opportunities for biomedical research by helping scientists to rapidly comprehend complex processes in biology. UniProtKB provides expert curation of functionally characterized variants reported in the literature, many of them associated with disease. Understanding the association of genetic variation with its functional consequences in proteins is essential for the interpretation of large-scale genomics data. UniProtKB protein sequences and sequence feature annotations such as active sites, modified sites, binding domains and amino acid variation have been mapped to the current build of the human genome (GRCh38). We have also created public track hubs for the Ensembl and UCSC browsers. We examine some specific biological examples in disease-related genes and proteins, illustrating the utility of combining protein and genome annotations for the functional interpretation of variants. We also present a larger comparison of UniProtKB variant and feature curation data that collocates with clinically observed SNP data from ClinVar, illustrating how protein and gene disease annotations currently compare and discuss some additional uses that might follow with combined annotation. The combination of gene and protein annotation can assist medical geneticists with the functional interpretation of variation. The genomic track hubs are downloadable from the UniProt FTP site (http://www.uniprot.org/downloads) and directly discoverable as public track hubs at the USCS 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 for missense mutations, connecting these variations to more subtle effects like the potential disruption of enzymatic active sites, protein-binding sites, or post-translational modification sites is not easily achieved. Some tools such as PolyPhen-2 (Adzhubei et al. 2013) predict the effects of variants using a subset of protein annotation from UniProtKB, such as active sites, in their score and including additional annotations in such tools should improve predictive capabilities. UniProtKB represents decades of effort in capturing protein and gene names, function, catalytic activity, cofactors, pathway information, subcellular location, protein-protein interactions, patterns of expression, and phenotypes or diseases associated with variants in a protein (Famiglietti et al. 2014) using literature-based and semi-automated expert curation. Included in this information are sequence positional features such as enzyme active sites; modified residues; binding domains; protein isoforms; glycosylation sites; protein variations and more. Aligning this valuable curated protein functional information with genomic annotation and making it seamlessly available to the genomic, proteomic and clinical communities will greatly inform studies on functional variation.

Genome browsers (Kent et al. 2002; Yates et al. 2016) that provide a web based graphical view of genomic data and utilize standard data file formats that enable users to import and visualize their own data through community track hubs, (Raney et al. 2014) provide one useful means to integrate and share UniProt's extensive protein annotation data. We have previously described the Proteins API providing searching and programmatic access to protein and associated genomics data (Nightingale et al. 2017). Here, we describe our mapping of UniProtKB sequences and sequence feature annotations to the current build of the human genome (GRCh38) and the creation of public track hubs for the Ensembl and UCSC browsers. We illustrate the utility of the combination of protein and genome annotations with some specific biological examples in disease related genes

and proteins and also a larger comparison of UniProtKB variant and feature curation data that collocates with clinically observed SNP data from ClinVar (Landrum et al. 2014).

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. For the UniProt 2017_08 release, the locations of 111,346 human reference protein sequences from UniProtKB were mapped to the GRC. This includes the UniProtKB/Swiss-Prot section with 18,626 canonical and 21,439 isoform sequences and UniProtKB/TrEMBL with 56,106 sequences. Thirty four different positional annotation types (e,g. active sites, modified residues, domains and natural 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 due to some gene families are indistinguishable at the protein level and some genes map to chromosomal regions with alternative assemblies. Examples of this include histones, MHC proteins and pseudoautosomal regions on both the X and Y chromosome (Helena Mangs and Morris 2007; Veerappa et al. 2013). The data are provided in extended BED file and binary BigBed file formats on the UniProt FTP site in the genome_annotation_tracks directory. These files contain genome coordinates and selected annotation and source attribution on each feature when available. Additionally, this data can be accessed directly from the public Genome Track hubs at the UCSC and the Ensembl browsers and also via a track hub registry search for “UniProt” in the Track Hub Registry (trackhubregistry.org), which provides links to load the tracks in either browser. Ten feature tracks are turned on by default but the rest can be enabled on the browser. We recommend using track hubs and not the BED text files on these genome browsers. The text BED files are in an extended BED format that

Annotation Type		Description (feature_name)	Features Mapped
Molecule Processing	Proteome	Location of complete protein and isoform sequences (proteome)	111,346
	Signal	Sequence targeting proteins to the secretory pathway or periplasmic space (signal)	10,107
	Transit peptide	Extent of a transit peptide for organelle targeting (transit)	474
	Chain	Extent of a polypeptide chain in the mature protein (chain)	25,084
	Peptide	Extent of an active peptide in the mature protein (peptide)	382
	Propeptide	Peptide that is cleaved during maturation or activation (propep)	803
	Initiator met	Cleavage of the initiator methionine (init_met)	2,102
Regions	Topological domain	Location of non-membrane regions of membrane-spanning proteins (topo_dom)	18,630
	Transmembrane	Extent of a membrane-spanning region (transmem)	43,175
	Intramembrane	Extent of a region located in a membrane without crossing it (intramem)	321
	Domain	Position and type of each modular protein domain (domain)	65,809
	Repeat	Positions of repeated sequence motifs or domains (repeat)	18,371
	Calcium binding	Position(s) of calcium binding region(s) within the protein (ca_bind)	729
	Zinc finger	Position(s) and type(s) of zinc fingers within the protein (zn_fing)	9,073
	DNA binding	Position and type of a DNA-binding domain (dna_bind)	1,025
	Nucleotide binding	Nucleotide phosphate binding region (np_bind)	3,696
	Region	Region of interest in the sequence (region)	9,587
	Coiled coil	Positions of regions of coiled coil within the protein (coiled)	10,947
	Motif	Short (up to 20 aa) sequence motif of biological interest (motif)	3,292
Sites	Active site	Amino acids directly involved in the activity of an enzyme (act_site)	3,987
	Metal binding	Binding site for a metal ion (metal)	2,956
	Binding site	Binding site for any chemical group (binding)	6,015
	Site	Any interesting single amino acid site on the sequence (site)	2,088
Amino Acid Modifications	Modified residue	Modified residues excluding lipids, glycans & cross-links (mod_res)	52,516
	Lipidation	Covalently attached lipid group(s) (lipid)	1,019
	Glycosylation	Covalently attached glycan group(s) (carbohyd)	16,390
	Disulfide bond	Cysteine residues participating in disulfide bonds (disulfid)	19,586
	Cross-link	Residues in covalent linkage between proteins (crosslnk)	2,688
	Non-standard residue	Occurence of non-standard amino acids (selenocysteine & pyrrolysine) in the protein sequence (non_std)	37
Structure	Helix	Helical regions within the experimentally determined structure (helix)	55,762
	Turn	Turns within the experimentally determined protein structure (turn)	14,298
	Beta strand	Beta strand regions within the experimentally determined protein structure (strand)	61,433
Variants	Mutagenesis	Site experimentally altered by mutagenesis (mutagen)	19,121
	Natural variant	Description of a natural variant of the protein (variant)	75,251

Table 1: UniProtKB sequence annotations in track hubs. Annotation types, descriptions and current number of each feature mapped to the human genome are shown. UniProt release 2017_08 (Aug 2017) was used for this table. For more information on sequence features in UniProt see (www.uniprot.org/help/sequence_annotation).

is used to create the BigBed files. This extended BED format is not supported in a consistent manner on all browsers but the BigBed files are well supported. The extended BED format files are useful to extract selected genome locations and annotation for combination with other data for data integration and additional computational analysis similar to what we describe below for mapping to ClinVar SNPs.

Biological Examples: To illustrate the utility of combining UniProt protein feature and variation annotation 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 and Migeon 1970; Schiffmann 2009), a rare X-linked lysosomal storage disease where glycosphingolipid catabolism fails and glycolipids accumulate in many tissues from birth. Many protein altering variants in *GLA* are associated with FD, with UniProt biocurators recording 226 manually reviewed natural 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. S**election A** illustrates a situation where an amino acid component of the enzyme 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

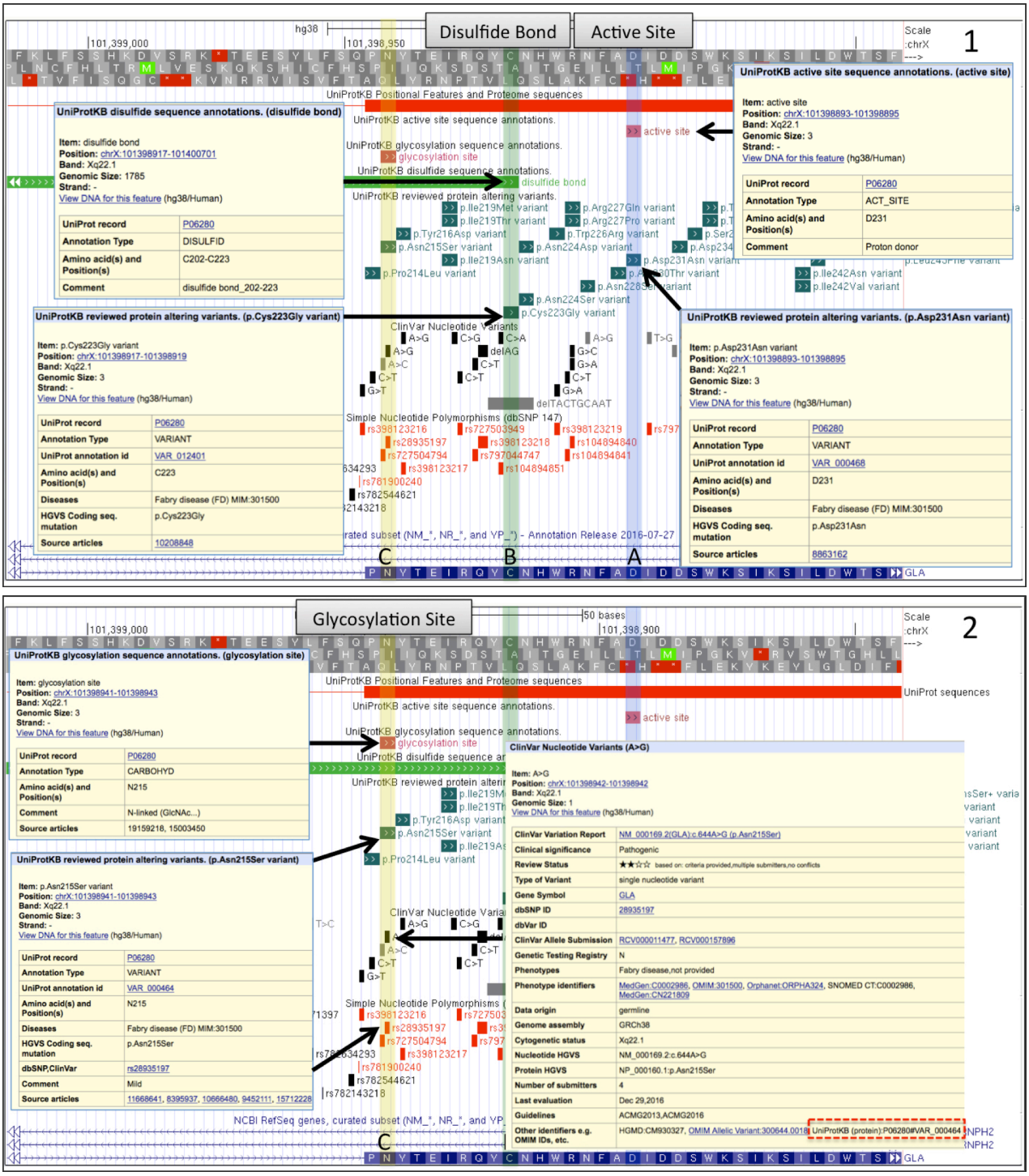

Figure1. The GLA gene (P06280, alpha-galactosidase A) associated with Fabry disease (FD) shown on the UCSC browser with UniProt genome tracks plus ClinVar and dbSNP tracks. **Panel 1 selection A** shows UniProt annotation for part of the enzyme's Active Site and an amino acid variation from a SNP associated 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.

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 and 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 a C-terminal end of an α-helix encoded by exons four and five. Disruption of a disulfide bond would disrupt the secondary 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. Of the ten cysteine's involved in the five disulfide bonds in alpha-galactosidase A, seven 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 (https://www.ncbi.nlm.nih.gov/clinvar/variation/10730/) and in UniProtKB as mildly pathogenic 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 over the years and, in ClinVar, multiple submitters have annotated this variant as pathogenic, citing at least 10 publications describing this variant and others seen in patients and family pedigrees with Fabry disease (Davies et al. 1993; Eng et al. 1993; Eng and Desnick 1994; Topaloglu et al. 1999; Mills et al. 2005; Ishii et al. 2007; Wu et al. 2011; Ebrahim et al. 2012; Bean et al. 2013; Tian et al. 2013). UniProtKB has two of the same publications and three publications (Guffon et al. 1998; Blaydon et al. 2001; Shabbeer et al. 2005) currently not in ClinVar. However due to UniProt's focus 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 (Garman and Garboczi 2004; Chen et al. 2009). 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 et al. 1998) and targeting to the lysosome (Ghosh et al. 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 proteins produced in different 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) α- Alpha-galactosidase A protein preparations used for treatment (Lee et al. 2003).

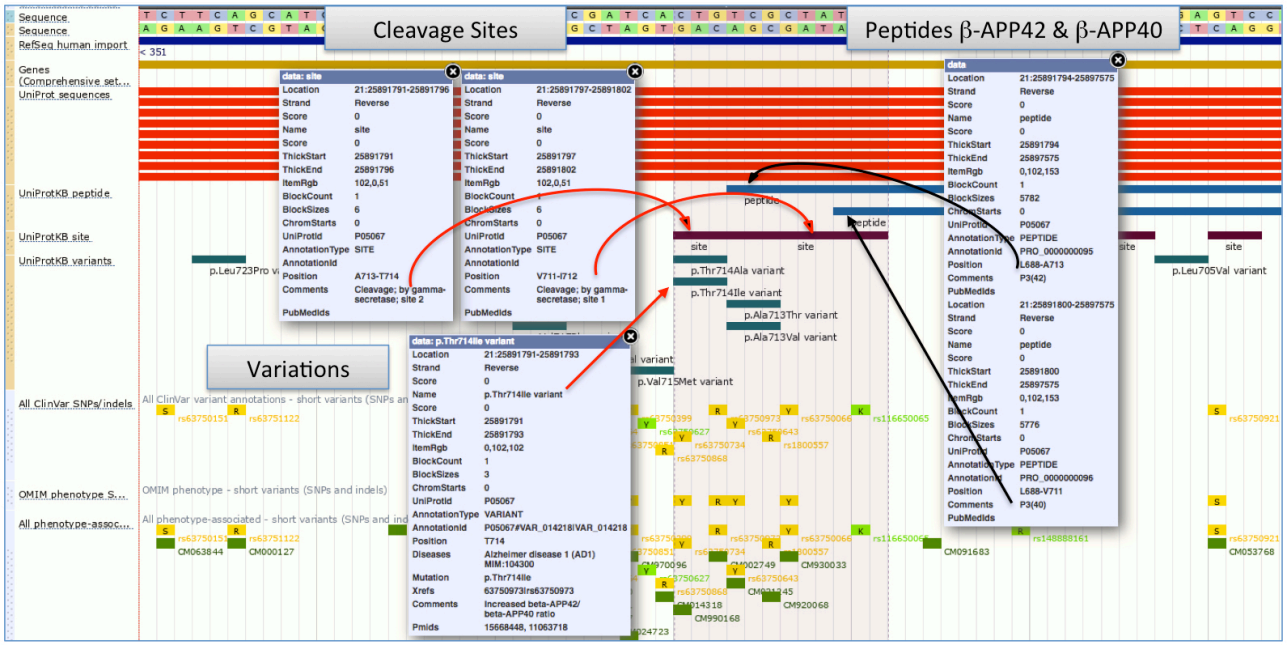

Figure 2. The APP gene (Amyloid beta A4 protein, UniProt Acc. P05067) associated with Alzheimer's disease shown on the Ensembl browser with selected UniProt Genome Tracks. Shows how pathogenic variations overlap protein cleavage sites contributing to the aberrant ratios of the β-APP42 and β -APP40 peptides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

Figure 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 found on the surface of neurons relevant to neurite growth. The protein has multiple functions and peptides derived from the mature protein are found in amyloid plaques in brain tissue from patients with Alzheimer's disease (Chartier-Harlin et al. 1991; Goate et al. 1991; Murrell et al. 1991; Hendriks et al. 1992; Mullan et al. 1992; Denman et al. 1993; Liepnieks et al. 1993; Eckman et al. 1997; Cras et al. 1998; Ancolio et al. 1999; Kwok et al. 2000; 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 2 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 UniProt, OMIM and ClinVar also 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. This coincided in brain with deposition nonfibrillar pre-amyloid plaques composed primarily of N-truncated β-APP42 (Kumar-Singh et al. 2000).

Mapping UniProt Features to ClinVar SNPs: For a larger scale analysis, we examined SNPs from ClinVar that overlap selected protein features and compared ClinVar SNPs to UniProt amino acid variations. 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; in the case of a feature spanning an intron, positions in the intron are excluded. For this comparison, a pathogneic ClinVar SNP is broadly defined as one with at least one or more pathogenic or likely pathogenic assertions by a submitter. **Table 2** shows how many ClinVar SNPs exist within the boundaries of a UniProtKB feature and how many of these SNPs have assertions of pathogenicity. It is reasonable to expect that some classes of protein functional features are critical to function and one might expect to see less variation in these regions and that variations that do occur in the feature are more likely to be pathogenic as compared to regions outside functional features. **Table 2** 'suggests' some features where disruption by a SNP may be more or less likely to cause harm. It is interesting to note that of

UniProt Feature Type	Number of UniProt Features	ClinVar SNPs in feature	Pathogenic SNPs in feature	% Pathogenic SNPs in Feature
Non Standard AA	37	2	2	100.0%
Initiator Met.	2,102	74	56	75.7%
Binding Sites	6,015	247	169	68.4%
DNA Binding dom.	1,025	1,803	1,185	65.7%
Natural variant	75,251	48,526	27,967	57.6%
Active Site	3,987	92	53	57.6%
Nucleotide binding	3,696	883	483	54.7%
Site	2,088	147	75	51.0%
Intramembrane	321	456	220	48.2%
Mutagen	19,121	1,067	513	48.1%
Zn Finger	9,073	1,107	495	44.7%
Lipid	1,019	17	7	41.2%
Metal Binding	2,956	3,750	1,481	39.5%
Strand	61,433	9,918	3,910	39.4%
Ca Binding Site	729	363	136	37.5%
Turn	14,298	1,640	612	37.3%
Peptide	382	235	86	36.6%
Transmembrane	43,175	14,063	5,129	36.5%
Motif	3,292	521	188	36.1%
Helix	55,762	14,722	5,223	35.5%
Cross Link	2,688	35	11	31.4%
Modified Residue	52,516	1,079	334	31.0%
Region	9,587	27,481	8,231	30.0%
Domain	65,809	137,265	41,037	29.9%
Topological dom.	18,630	36,331	9,677	26.6%
Signal Peptide	10,107	2,921	775	26.5%
Repeat	18,371	13,076	3,375	25.8%
Coiled Coil	10,947	13,824	3,427	24.8%
Transit peptide	474	618	131	21.2%
Carbohydrate Site	16,390	292	58	19.9%
Propeptide	803	952	174	18.3%

Table 2. ClinVar SNPs that overlap UniProt Features. Pathogenic SNPs are those with one or more assertions of being pathogenic.

the eight features where currently over 50% of the ClinVar SNPs have pathogenic assertions six are single amino acid features where it appears many changes may not be tolerated. However, one single amino acid feature carbohydrate/glycosylation sites are at the bottom of the table with lees than 20% pathogenic assertions. However compared to other features the number of all SNPs seen in carbohydrate/glycosylation sites is low with less that 2% have any overlapping SNP. Though it is tempting to speculate on biological reasons why certain features show more or less pathogenic SNPs, there are differences in annotation practices, incomplete annotation, different priorities and biases in both databases, plus details such as overlapping features, that require a more rigorous statistical comparison before drawing any conclusions. A recent comparison of 10,000 human genomes (Telenti et al. 2016) looked at one computational feature, 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. Comparable features in Table 2 are the transmembrane and intramembrane features. A more rigorous analysis on this topic is in progress.

Comparison of ClinVar SNPs with UniProtKB natural amino acid variations (**Table 3**) shows that currently 23,344 UniProt variants collocate on the genome with ClinVar SNPs, which is 30% of all UniProtKB variants and 9% of all ClinVar SNPs. This mapping shows that currently 52% of UniProtKB disease associated variants exist in ClinVar and 27% of ClinVars SNPs with one or more pathogenic assertions are present as amino acid variants in UniProtKB. There is general agreement in annotation between variations that map with 84% of UniProtKB disease associated variants mapping to pathogenic SNPs in ClinVar and 79% of UniProtKB variants not associated with a disease mapping to ClinVar SNPS without a pathogenic assertion. That some annotations are discordant between the databases is not surprising as the standards, methods and levels of evidence employed by protein curators and medical geneticists are different and evolving. Recent efforts in the medical community to standardize methods and the levels of evidence required for

annotation of genetic variants (Richards et al. 2015; Amr et al. 2016; Manrai et al. 2016; O'Daniel et al. 2016; Walsh et al. 2016) along with increasing amounts of population data (Amr et al. 2016; Walsh et al. 2016) are leading to the revaluation of previous assertions of pathogenicity. Comparisons like this and investigation into the discordant annotations should help improve both genomic and proteomic annotation. UniProt is collaborating with ClinVar by making reciprocal links between our databases for variants that exist in both databases as a first step in making this data more accessible to both communities.

	All CinVar SNPs (255,857)	**Pathogenic SNPs (54,467)**	**Non-Pathogenic SNPs (201,390)**
All UniProt Variants (76,612)	23,344	14,551	8,793
UniProt Disease Variants (29,614)	15,323	12,830	2,493
UniProt non-Disease Variants (46,998)	8,021	1,721	6,300

Table 3. Comparison of number and type of ClinVar SNPs and UniProt amino acid variants that overlap in genome position and result in the same amino acid change.

Discussion

The combination of protein and genome annotation is made easier by accurate mapping and allows a more detailed look at the effects of variation on protein function. Exome sequencing for 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 help inform variant curation by providing a functional explanation for a variant's effect. The results suggest that location of a variant within some functional features may correlate with pathogenicity. UniProtKB features have been mapped to the genome before, as

the UCSC genome browser has provided selected UniProtKB/Swiss-Prot features for GRCh37 for several years and recently updated them to GRCh38. The mappings described here contain additional annotation 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 and automated annotation provided by a combination of automated prediction algorithms with human curated rules (Vasudevan et al. 2010; Dimmer et al. 2012; Famiglietti et al. 2014; Pedruzzi et al. 2015). This includes sequence features such as active sites, binding sites and functional domains. 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 the future releases. We hope to extend genome mapping to other model organisms such as mouse or rat in the future. UniProt is working with the UCSC and Ensembl browser teams to pool our knowledge and improve the presentation of protein annotation on the respective browsers. In addition, the data provided here are available programmatically via a REST API (Nightingale et al. 2017).

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.

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.

Phase One: Mapping Ensembl identifiers and translations to UniProtKB protein sequences. UniProt imports Ensembl translated sequences and associated identifiers, including gene symbol and the HGNC identifier. An Ensembl translated sequence is mapped to a UniProtKB/Swiss-Prot sequence if and only if the Ensembl translated sequence is 100% identical to the UniProtKB sequence. No insertions, deletions or extensions are allowed in the mapped Ensembl translated sequence. All mapped Ensembl translations from a single gene are collated and defined as separate isoforms of the UniProtKB/Swiss-Prot entry where more than one Ensembl translation can be mapped to a single protein isoform sequence. One protein isoform is defined as the canonical sequence. Where an Ensembl translation does not match an existing UniProtKB/Swiss-Prot protein sequence, the Ensembl translation is added as a new UniProtKB/TrEMBL entry. The 100% identity threshold to an Ensembl translated sequence that is currently enforced is strict and means that a relatively small number of well-annotated proteins in UniProtKB/Swiss-Prot are not currently mapped to the genome (~3%). We expect to develop means make some of these available in the future, others like 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 with these genomic coordinates is the 3' and 5' UTR offsets in the translation and exon splice phasing. With this collated genomic 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 effectively gives UniProt a set of

peptide fragments with exon identifiers and genomic coordinates. This set of peptide fragments is stored as the basis for all protein to genomic mappings in UniProt.

Phase Three: Converting UniProt Position Annotations to their Genomic Coordinates: UniProt position annotations or features all have either a single amino acid location or amino acid range within the UniProtKB canonical protein sequence. With the exon genomic coordinates mapped to the protein peptide fragment, the genomic coordinates of a positional annotation are calculated by finding the amide (N) terminal exon and the carboxyl (C) terminal exon. The N-terminal or 5' genomic coordinate (UP_{Gcoord}) of the positional annotation is calculated by 1. calculating the amino acid offset (N_{aa}) from the N-terminal protein peptide fragment start amino acid. 2. Taking the 5' genomic coordinate (G_{coord}) of the mapped exon and exon splice phasing (phase); the genomic coordinate is calculated as:

$$UP_{Gcoord} = G_{coord} + (N_{aa} * 3) + phase.$$

Likewise, the C-terminal, 3' genomic coordinate is calculated in the same way but with the 3' exon. For reverse strand mapped genes the above formula is modified to take into account the negative direction and phasing. Where a positional feature is spread out over multiple exons, the introns will be automatically included in the mapping. This process is illustrated in **Figure 3**. If the positional feature is composed of a single amino acid, the three bases that denote that amino acid are given as the genomic coordinate. This is a limitation for the UniProt reviewed natural variants as UniProt does not independently define the specific allele change responsible for the missense, protein-altering variant. Therefore, UniProt is providing cross-references to dbSNP.

Phase 4: UniProt BED and BigBed Files: Converting protein functional information into its genomic equivalent requires standardized file formats. Genomic data is collated in files based upon a simple tab delimited text format or the SAM (Sequence Alignment/Map) format (Li et al 2009). The Browser Extensible Data (BED), a tabulated based format, represents the best format type for

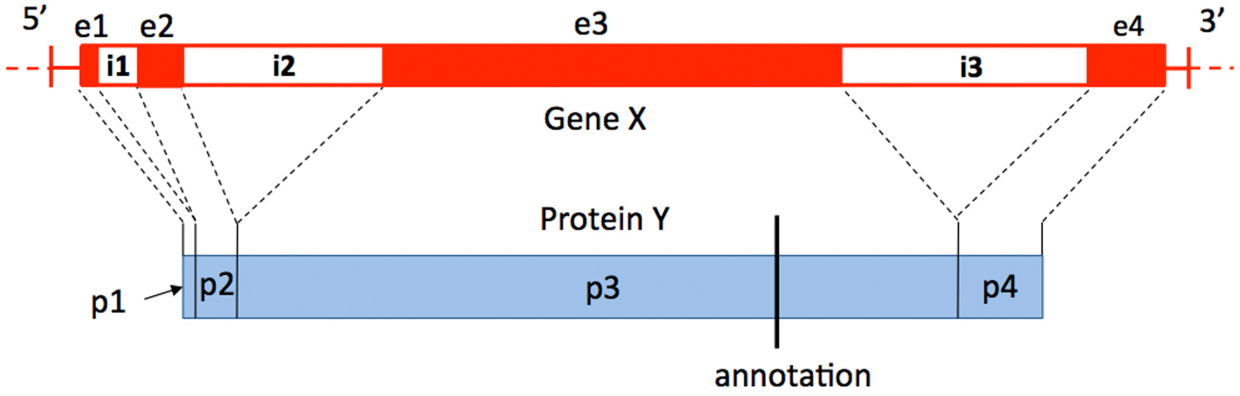

Figure 3. Converting UniProt Position Annotations to Genomic Coordinates. The genomic coordinates of a positional annotation are calculated by finding the N-terminal and the C-terminal of its exon. The genomic coordinate of the annotation in peptide p3 is calculated as:

$UP_{Gcoord} = G_{coord}$ of e3 start + N_{aa} p3 peptide start to annotation * 3 + phase

Likewise the C-terminal, 3' genomic coordinate is calculated in the same way but with the 3' exon.

converting UniProt annotations into genomic features for display in a genome browser. A BED file is interpreted as an individual horizontal feature 'track' when uploaded into a genome browser; this allows users to choose specific UniProtKB annotations most relevant to their analysis. The binary equivalent of the BED file is BigBed (Kent et al. 2010). This file is more flexible in allowing for additional tabulated data elements providing UniProt a greater opportunity to fully represent its protein annotations and one of the file formats used to make track hubs (Raney et al. 2014). A track hub is a web-accessible directory of files that can be displayed in track hub enabled genome browsers. 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 directly through the genome browser rather than searching for hubs at the institute or bioinformatics resource that generated the data.

Using the protein and genomic coordinates with additional feature specific annotations from UniProtKB, BED (UCSC 2016a) and BED detail (UCSC 2016b) formatted files were produced for

the UniProtKB human reference proteome by converting the genomic mappings to the zero based genomic coordinates used by genome browsers. A region of DNA (or block), its size and offset from the start of DNA being annotated is calculated for a protein annotation or sequence. If a protein annotation or sequence is defined by a range (e.g. chain, domain, region) or is composed of more than one amino acid or sub-region (e.g. exon) one block element is defined in a comma separated list for each sub-region (e.g. each exon). This means that a protein annotation could be represented as a single block or more than one block depending upon its composition. Blocks and block sizes for the UniProt proteome sequences define the specific exons and the sizes of those exons that are translated into the protein sequence. A standard BED file is generated for the proteome sequences and BED detail files are generated for each positional annotation type listed in Table 1. In both file formats the UniProtKB accession for the sequence or individual annotation is provided in the BED name column (column 4) to provide a convenient link to the original UniProt entry. In the BED detail files the last two columns are used for UniProtKB annotation identifiers and description, when available.

UniProt BigBed files differ from the BED detail files with the addition of extra columns that separate the final description column of the BED detail file. For all UniProt positional annotations except variant annotations an additional literature evidence column is defined. Variant BigBed files have a further additional column defining the protein sequence change as a consequence of the genomic variant. BigBed files (Kent et al. 2010) are produced from the text BED detail files using the UCSC bedToBigBed converter program (http://hgdownload.cse.ucsc.edu/admin/exe/). bedToBigBed requires additional information defining the column structure of the BED detail file as an autoSQL (Kent and Brumbaugh 2002) (http://hgwdev.cse.ucsc.edu/~kent/exe/doc/autoSql.doc) and the chromosome names and sizes for the genome assembly where UniProt use the chromosome names and sizes available from Ensembl's latest assembly release. Binary BigBed files are

generated for each type sequence annotation and the UniProtKB human proteome protein sequence set.

Mapping ClinVar SNPs to protein features and variants: Data for comparing ClinVar SNPs to UniProt features come from the ClinVar variant_summary.txt file on the NCBI FTP site, the feature specific BED files and the human variation file humsavar.txt on the UniProt FTP site. UniProt data is from release 2017_08 and the ClinVar August 2017 release. 1) For each feature in UniProtKB, we check the genomic position against the position for each record in ClinVar. If the chromosome numbers match and the genome positions of the protein feature overlap the 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 same Protein in the same chromosome. A flag is added if a SNP coordinate is within the exon boundary. Variants outside of exons were excluded from further analysis. 3) For each UniProt variant in 2, we check the protein accession number and that the amino acid change reported in UniProt and ClinVar is the same, if they match a mapping between a UniProtKB Variant and ClinVar SNP is established. A pathogenic ClinVar SNP is broadly defined as one with any assertion that it is pathogenic or likely pathogenic. Those classified as risk factors or associated with a disease condition were not included but represent a very small portion (<2%) of the harmful ClinVar assertions.

Data access

Tab-delimited text files in an extended BED file format and binary BigBed files for use as genome Track Hubs are located in the UniProt FTP site in the genome_annotation_tracks directory (http://www.uniprot.org/downloads). Public Track hubs are available at the UCSC genome browser (Tyner et al. 2016) at (genome.ucsc.edu/cgi-bin/hgHubConnect?) and the Ensembl genome browser (Hubbard et al. 2007; Aken et al. 2016) via a track hub registry search

UniProt Genomic Mapping

(ensembl.org/Homo_sapiens/Info/Index#modal_user_data-Track_Hub_Registry_Search) for "UniProt". Both can be found through a direct Track Hub Registry (trackhubregistry.org) search, which provides links to load the tracks at either browser. Track hubs and related files are updated with each release of UniProt KB.

Links from trackhubregistry.org that load the default UniProt tracks automatically are shown below. Additional tracks can be selected for display on each browser. Note if some tracks fail to display on first view please click to refresh browser page until all are displayed.

UCSC Browser: (http://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)

Ensembl Browser:

(http://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 and U01HG007437.

References

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]* **Chapter 7**: Unit7 20.

Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P et al. 2016. Ensembl 2017. *Nucleic Acids Res* doi:10.1093/nar/gkw1104.

Amr SS, Al Turki SH, Lebo M, Sarmady M, Rehm HL, Abou Tayoun AN. 2016. Using large sequencing data sets to refine intragenic disease regions and prioritize clinical variant interpretation. *Genetics in medicine : official journal of the American College of Medical Genetics* doi:10.1038/gim.2016.134.

Ancolio K, Dumanchin C, Barelli H, Warter JM, Brice A, Campion D, Frebourg T, 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. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 4119-4124.

Bean LJ, Tinker SW, da Silva C, Hegde MR. 2013. Free the data: one laboratory's approach to knowledge-based genomic variant classification and preparation for EMR integration of genomic data. *Human mutation* **34**: 1183-1188.

Blaydon D, Hill J, Winchester B. 2001. Fabry disease: 20 novel GLA mutations in 35 families. *Human mutation* **18**: 459.

Chartier-Harlin MC, Crawford F, Houlden H, Warren A, Hughes D, Fidani L, Goate A, Rossor M, Roques P, Hardy J et al. 1991. Early-onset Alzheimer's disease caused by mutations at codon 717 of the beta-amyloid precursor protein gene. *Nature* **353**: 844-846.

Chen R, Jiang X, Sun D, Han G, Wang F, Ye M, Wang L, Zou H. 2009. Glycoproteomics analysis of human liver tissue by combination of multiple enzyme digestion and hydrazide chemistry. *Journal of proteome research* **8**: 651-661.

Cras P, van Harskamp F, Hendriks L, Ceuterick C, van Duijn CM, Stefanko SZ, Hofman A, Kros JM, Van Broeckhoven C, Martin JJ. 1998. Presenile Alzheimer dementia characterized by amyloid angiopathy and large amyloid core type senile plaques in the APP 692Ala-->Gly mutation. *Acta neuropathologica* **96**: 253-260.

Davies JP, Winchester BG, Malcolm S. 1993. Mutation analysis in patients with the typical form of Anderson-Fabry disease. *Human molecular genetics* **2**: 1051-1053.

Denman RB, Rosenzcwaig R, Miller DL. 1993. A system for studying the effect(s) of familial Alzheimer disease mutations on the processing of the beta-amyloid peptide precursor. *Biochemical and biophysical research communications* **192**: 96-103.

Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, Bely B, Browne P, Mun Chan W, Eberhardt R et al. 2012. The UniProt-GO Annotation database in 2011. *Nucleic Acids Res* **40**: D565-570.

Ebrahim HY, Baker RJ, Mehta AB, Hughes DA. 2012. Functional analysis of variant lysosomal acid glycosidases of Anderson-Fabry and Pompe disease in a human embryonic kidney epithelial cell line (HEK 293 T). *Journal of inherited metabolic disease* **35**: 325-334.

Eckman CB, Mehta ND, Crook R, Perez-tur J, Prihar G, Pfeiffer E, Graff-Radford N, Hinder P, Yager D, Zenk B et al. 1997. A new pathogenic mutation in the APP gene (I716V) increases the relative proportion of A beta 42(43). *Human molecular genetics* **6**: 2087-2089.

Eng CM, Desnick RJ. 1994. Molecular basis of Fabry disease: mutations and polymorphisms in the human alpha-galactosidase A gene. *Human mutation* **3**: 103-111.

Eng CM, Resnick-Silverman LA, Niehaus DJ, Astrin KH, Desnick RJ. 1993. Nature and frequency of mutations in the alpha-galactosidase A gene that cause Fabry disease. *American journal of human genetics* **53**: 1186-1197.

Famiglietti ML, Estreicher A, Gos A, Bolleman J, Gehant S, Breuza L, Bridge A, Poux S, Redaschi N, Bougueleret L et al. 2014. Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Human mutation* **35**: 927-935.

Garman SC, Garboczi DN. 2004. The molecular defect leading to Fabry disease: structure of human alpha-galactosidase. *J Mol Biol* **337**: 319-335.

Germain DP, Poenaru L. 1999. Fabry disease: identification of novel alpha-galactosidase A mutations and molecular carrier detection by use of fluorescent chemical cleavage of mismatches. *Biochemical and biophysical research communications* **257**: 708-713.

Ghosh P, Dahms NM, Kornfeld S. 2003. Mannose 6-phosphate receptors: new twists in the tale. *Nat Rev Mol Cell Biol* **4**: 202-212.

Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L et al. 1991. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349**: 704-706.

Guffon N, Froissart R, Chevalier-Porst F, Maire I. 1998. Mutation analysis in 11 French patients with Fabry disease. *Human mutation* **Suppl 1**: S288-290.

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

Hendriks L, van Duijn CM, Cras P, Cruts M, Van Hul W, van Harskamp F, Warren A, McInnis MG, Antonarakis SE, Martin JJ et al. 1992. Presenile dementia and cerebral haemorrhage linked to a mutation at codon 692 of the beta-amyloid precursor protein gene. *Nature genetics* **1**: 218-221.

Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T et al. 2007. Ensembl 2007. *Nucleic Acids Res* **35**: D610-617.

Ioannou YA, Zeidner KM, Grace ME, Desnick RJ. 1998. Human alpha-galactosidase A: glycosylation site 3 is essential for enzyme solubility. *Biochem J* **332 (Pt 3)**: 789-797.

Ishii S, Chang HH, Kawasaki K, Yasuda K, Wu HL, Garman SC, Fan JQ. 2007. Mutant alpha-galactosidase A enzymes identified in Fabry disease patients with residual enzyme activity: biochemical characterization and restoration of normal intracellular processing by 1-deoxygalactonojirimycin. *Biochem J* **406**: 285-295.

Kent J, Brumbaugh H. 2002. autoSql and autoXml: Code Generators from the Genome Project. In *Linux Journal*. Linux Journal, Houston, Texas.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome research* **12**: 996-1006.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**: 2204-2207.

Kumar-Singh S, De Jonghe C, Cruts M, Kleinert R, Wang R, Mercken M, De Strooper B, Vanderstichele H, Lofgren A, Vanderhoeven I et al. 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. *Human molecular genetics* **9**: 2589-2598.

Kwok JB, Li QX, Hallupp M, Whyte S, Ames D, Beyreuther K, Masters CL, Schofield PR. 2000. Novel Leu723Pro amyloid precursor protein mutation increases amyloid beta42(43) peptide levels and induces apoptosis. *Annals of neurology* **47**: 249-253.

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**: D980-985.

Lee K, Jin X, Zhang K, Copertino L, Andrews L, Baker-Malcolm J, Geagan L, Qiu H, Seiger K, Barngrover D et al. 2003. A biochemical and pharmacological comparison of enzyme replacement therapies for the glycolipid storage disorder Fabry disease. *Glycobiology* **13**: 305-313.

Liepnieks JJ, Ghetti B, Farlow M, Roses AD, Benson MD. 1993. Characterization of amyloid fibril beta-peptide in familial Alzheimer's disease with APP717 mutations. *Biochemical and biophysical research communications* **197**: 386-392.

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

Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, Margulies DM, Loscalzo J, Kohane IS. 2016. Genetic Misdiagnoses and the Potential for Health Disparities. *The New England journal of medicine* **375**: 655-665.

Mills K, Morris P, Lee P, Vellodi A, Waldek S, Young E, Winchester B. 2005. Measurement of urinary CDH and CTH by tandem mass spectrometry in patients hemizygous and heterozygous for Fabry disease. *Journal of inherited metabolic disease* **28**: 35-48.

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. *Nature genetics* **1**: 345-347.

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

Nightingale A, Antunes R, Alpi E, Bursteinas B, Gonzales L, Liu W, Luo J, Qi G, Turner E, 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 CB, Condron MM, Axelman K, Forsell C, Stenh C, Luthman J, Teplow DB, Younkin SG et al. 2001. The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Abeta protofibril formation. *Nature neuroscience* **4**: 887-893.

O'Daniel JM, McLaughlin HM, Amendola LM, Bale SJ, Berg JS, Bick D, Bowling KM, Chao EC, Chung WK, Conlin LK et al. 2016. A survey of current practices for genomic sequencing

test interpretation and reporting processes in US laboratories. *Genetics in medicine : official journal of the American College of Medical Genetics* doi:10.1038/gim.2016.152.

Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, Baratin D, Cuche BA, Bougueleret L, Poux S et al. 2015. HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res* **43**: D1064-1070.

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D et al. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**: 1003-1005.

Redonnet-Vernhet I, Ploos van Amstel JK, Jansen RP, Wevers RA, 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. *Journal of medical genetics* **33**: 682-688.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E et al. 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. *Genetics in medicine : official journal of the American College of Medical Genetics* doi:10.1038/gim.2015.30.

Romeo G, Migeon BR. 1970. Genetic inactivation of the alpha-galactosidase locus in carriers of Fabry's disease. *Science* **170**: 180-181.

Schiffmann R. 2009. Fabry disease. *Pharmacol Ther* **122**: 65-77.

Selkoe DJ. 1998. The cell biology of beta-amyloid precursor protein and presenilin in Alzheimer's disease. *Trends Cell Biol* **8**: 447-453.

Shabbeer J, Robinson M, Desnick RJ. 2005. Detection of alpha-galactosidase a mutations causing Fabry disease by denaturing high performance liquid chromatography. *Human mutation* **25**: 299-305.

Telenti A, Pierce LCT, Biggs WH, di Iulio J, Wong EHM, Fabani MM, Kirkness EF, Moustafa A, Shah N, Xie C et al. 2016. Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences* **113**: 11901-11906.

Tian ML, Yan YL, Xiong JC, Liu XX, Yang Y, Hu ZX. 2013. [Genetic and clinical study of three Chinese pedigrees with Fabry disease]. *Zhonghua yi xue yi chuan xue za zhi = Zhonghua yixue yichuanxue zazhi = Chinese journal of medical genetics* **30**: 185-188.

Topaloglu AK, Ashley GA, Tong B, Shabbeer J, Astrin KH, Eng CM, Desnick RJ. 1999. Twenty novel mutations in the alpha-galactosidase A gene causing Fabry disease. *Molecular medicine (Cambridge, Mass)* **5**: 806-811.

Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L et al. 2016. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* doi:10.1093/nar/gkw1134.

UCSC. 2016a. BED (Browser Extensible Data) format. Vol 2016, https://genome.ucsc.edu/FAQ/FAQformat - format1.

UCSC. 2016b. BED detail format. Vol 2016, https://genome.ucsc.edu/FAQ/FAQformat.html - format1.7.

Vasudevan A, Vinayaka CR, Natale DA, Huang H, Kahsay RY, Wu CH. 2010. *Structure-guided rule-based annotation of protein functional sites in UniProtKB*. Springer.

Veerappa AM, Padakannaya P, Ramachandra NB. 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**: 285-293.

Walsh R, Thomson KL, Ware JS, Funke BH, Woodley J, McGuire KJ, Mazzarotto F, Blair E, Seller A, Taylor JC et al. 2016. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genetics in medicine : official journal of the American College of Medical Genetics* doi:10.1038/gim.2016.90.

Wu X, Katz E, Della Valle MC, Mascioli K, Flanagan JJ, Castelli JP, Schiffmann R, Boudes P, Lockhart DJ, Valenzano KJ et al. 2011. A pharmacogenetic approach to identify mutant forms of alpha-galactosidase A that respond to a pharmacological chaperone for Fabry disease. *Human mutation* **32**: 965-977.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L et al. 2016. Ensembl 2016. *Nucleic Acids Res* **44**: D710-716.