Andrew Nightingale[1], Jie luo, Michele Magrane[1], Peter McGarvey[2], Sandra Orchard[1], Maria Martin[1], UniProt Consortium[1,2,3]
[1]EMBL-European Bioinformatics Institute, Cambridge, UK
[2]SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
[3]Protein Information Resource, Georgetown University, Washington DC & University of Delaware, USA

Enabling interpretation of protein variation effects with UniProt

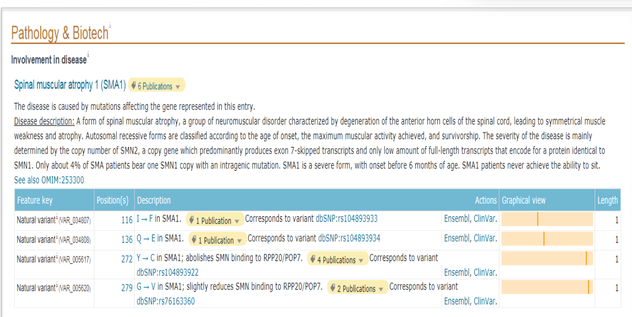
Introduction

Understanding the effect of genetic variants on protein function is crucial to thoroughly understand the role of proteins in disease biology. UniProt aims to support the scientific community, computational biologists and clinical researchers, by providing a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. This includes a comprehensive catalogue of protein altering variation data coupled with information about how these variants affect protein function.

UniProt variant data sources

Variant data from literature

Variants are captured from the scientific literature and manually reviewed for addition to UniProtKB/Swiss-Prot

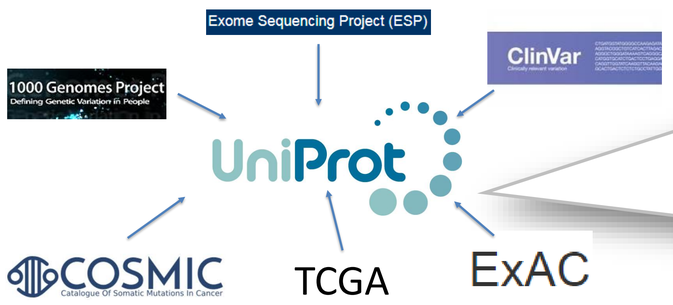


Description of disease associated with genetic variations in a protein

Variant data including effects of the variant on the protein and links to variant resources

Category	Number
Total reviewed variants	79,284
Disease-associated variants	30,471
Number of proteins with variants	12,886

Large-scale variant data



Database	Total Imported Variant	Total Unique
1000Genomes	859,757	81,216
ClinVar	183,655	76,218
COSMIC	184,237	18,863
ESP	939,238	68,803
ExAC	4,333,620	2,776,617
TCGA	1,202,700	920,549
UniProt	80,224	49,971
Total	7,781,431	3,992,437

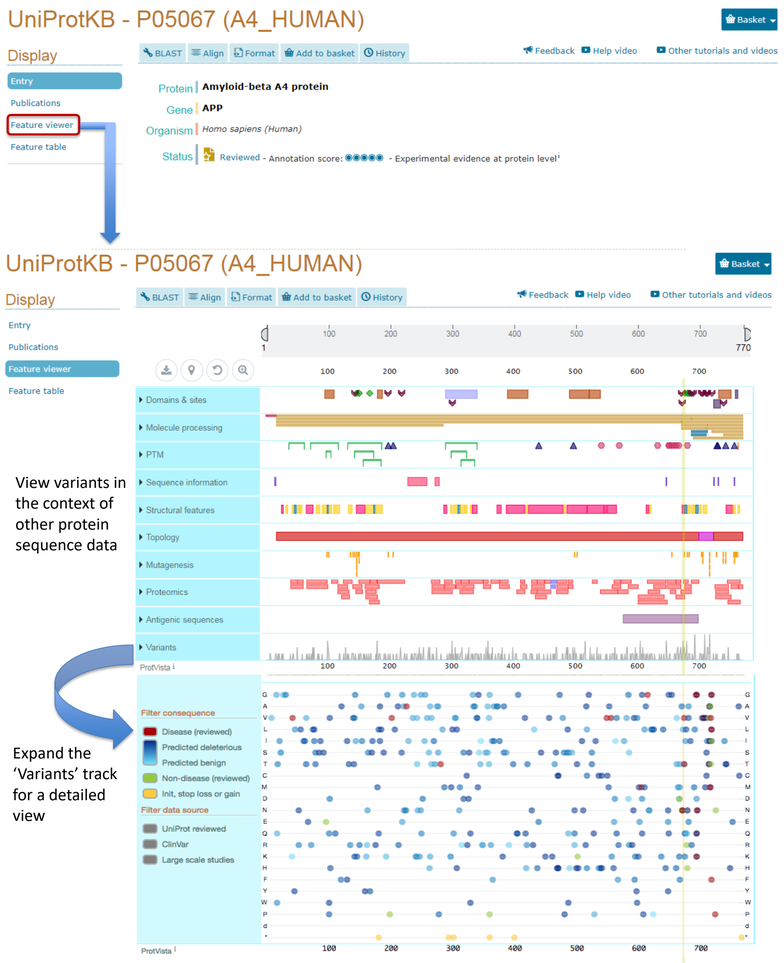
*Represents the number of UniProt variants with a dbSNP identifier

1. Imported variant data is dependent upon exact mapping between the reference proteome and genome.
2. Variant data is imported from a variety of resources to complement the set of variants captured from the literature

Interpretation protein variant effect with UniProt

1. Variants in the context of other protein functional annotation data

The UniProt feature viewer, ProtVista, provides a graphical view of protein variants in the context of other functional annotations such as domains, active sites and post-translational modifications. Possible variant effects can be identified by investigating co-localised protein functional residues.



The vertical yellow line acts as a positional ruler aligning protein features from the different categories. Ala673Thr is a known Alzheimer's disease; it aligns to a cleavage site and close to metal binding sites, indicating that it disrupts protein function or protein degradation.

The UniProt feature viewer can be integrated into any website. You can keep all available tracks or only those most relevant to you. Instructions can be found here: https://github.com/ebi-webcomponents*.

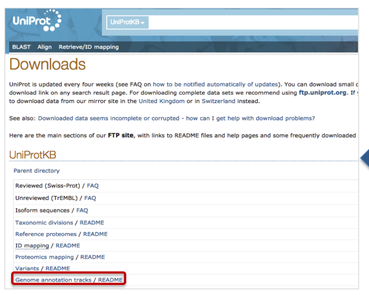
*Under active development

2. Integrate UniProt data into genome browsers

UniProt produces genome browser tracks which allow users to integrate UniProt sequence information including UniProt reviewed protein variants into genome browsers such as the Ensembl and UCSC browsers. You can couple UniProt data with other available genomic information and also with your own sequencing data. This makes it easier to see how changes in the genome can contribute to altered protein function and lead to disruptive disease phenotypes.

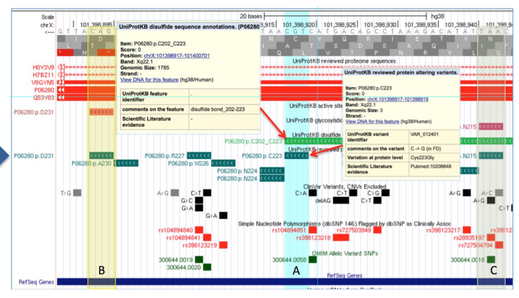
UniProt genome tracks can be directly loaded into a genome browser as a track hub from: https://trackhubregistry.org

Or downloaded from



http://www.uniprot.org/downloads

GLA gene associated with Fabry disease (FD) in the UCSC browser showing UniProt genome tracks plus variations from ClinVar, dbSNP and OMIM. Panel A shows shows UniProt data for a disulfide bond and an amino acid variation associated with FD that removes the cysteine required for a structural fold. Similar situations exist in panel B where part of the enzyme's active site is disrupted and panel C where an N-linked carbohydrate is located. Only the pathogenic variation in C is in other public resources.

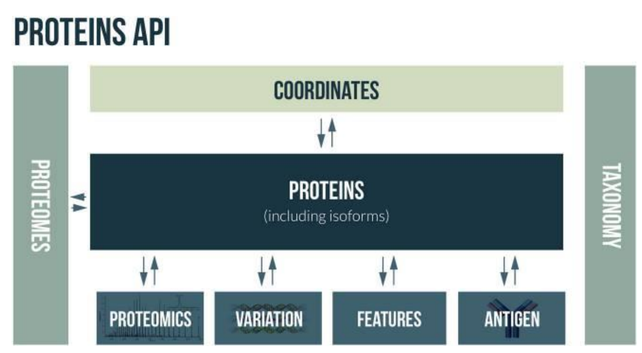


UniProt sequence information for FGFR2 displayed in the Ensembl genome browser

3. Access data programmatically via the Proteins API

The Proteins API provides programmatic access to protein and associated genomics data such as curated protein sequence positional annotations from UniProtKB as well as mapped variation and proteomics data from large-scale data sources.

https://www.ebi.ac.uk/proteins/api



Proteins API services and their data exchange relationship. Starting from any service, a user can retrieve information from another service using inter-relationships between the services.

Example use case

I'm interested in Wilson disease. How can I use the Proteins API to learn more about this?

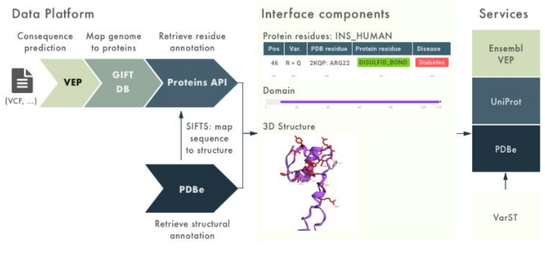
1. Use the 'Variation' end-point with the disease filter set to 'Wilson' to retrieve all UniProtKB records that are annotated with Wilson disease and have associated variants. This returns a single record, copper-transporting ATPase 2 from the ATP7B gene.
2. Using Wilson disease variants protein positions, retrieve sequence annotations via features end-point and determine if any of the variants align to functionally important residues such as binding sites or active sites.
3. The protein contains copper-binding sites and an active site; there are a number of variants from both reviewed and large-scale data, including variants which disrupt copper-binding sites as well as a variant from COSMIC which disrupts the active site.
4. Unique proteomics peptides are found with the protein which can be used to identify it in mass spectrometry experiments.

Read more in the publication PMID: 28383659

Coming soon - PepVEP

Protein Variant Effect Predictor

A Platform for interpreting protein, structure and clinical information with genomic variants effect prediction from Ensembl variation.



The PepVEP will integrate:

- Genomic and Variant effect information is taken from Ensembl Variation.
- Protein functional annotations from UniProt
- Protein structure functional annotations from PDBe
- Clinical annotations from all three services and others

To provide a more comprehensive interpretation of the functional effect of a genomic variant.

Funding
UniProt is funded by National Institutes of Health, European Molecular Biology Laboratory, Swiss Federal Government, British Heart Foundation, Parkinson's Disease United Kingdom and National Science Foundation

EMBL-EBI PIR SIB Swiss Institute of Bioinformatics

http://insideuniprot.blogspot.co.uk/ @uniprot http://www.ebi.ac.uk/training

www.uniprot.org
help@uniprot.org