Bringing protein functional positional annotation knowledge into reference genomes

Introduction

UniProt\(^1\) acts as the global central hub for protein information by providing comprehensive and high quality manually curated annotations and cross-references from, and connecting to, 144 biological data resources. In collaboration with Ensembl\(^2\), we have mapped all human protein sequences in UniProtKB to the human reference GRCh38 genome\(^3\). By utilising these mappings UniProt has produced a new resource where positional annotations from the UniProtKB are remapped to their genomic coordinates. Adding protein annotations to genomic browsers provides users with an enhanced understanding of the architecture of the gene and its translated protein product and how changes in the genome can contribute to altered protein function that can lead to disruptive disease phenotypes.

Remapping UniProtKB Positional Annotations to Genomic Coordinates

1. Reviewed sequences are mapped to equivalent Ensembl translations and new Ensembl translations are imported as new unreviewed entries. 2. Annotations such as variants, active sites, disulfide bonds and metal binding sites have known protein sequence locations. These protein positions are converted into genomic coordinates by utilizing the genomic mapping from 1. 3. The now genomic mapped protein annotations are represented using UCSC track file formats: BED detail and bigBed\(^4\). UniProt use the BED formats as they provide a globally recognized, simple format that can be extended to incorporate additional data fields; which we have used to add UniProtKB specific annotation detail.

Disease Related Annotations and their Influence on a Protein’s Function

1. The UniProtKB Natural Variant has been identified in Alzheimer's disease 1 (AD 1) patients. It is within the metal binding site and adjacent to a carbohydrate post translational modification (PTM) site. Inferring that the variant is either disrupting metal binding or the carbohydrate PTM.

Conclusions

The UniProt BED files are a new valuable resource for translational biology research. By remapping UniProtKB’s protein positional annotations and disease related variants into the genomic assembly context; UniProt now provides additional information for genomic analysis that contributes knowledge to the functional role of a gene, its role within biological processes and how modifications to the gene or protein product can lead to a disease state.

References