Variant Annotation

Overview

CellBase takes advantage of the data integrated to implement a rich and high-performance variant annotator. The variant annotation tool is integrated within the CellBase code and can be accessed in several different ways:

- **CellBase clients**: a number of client libraries are provided which make intensive use of the CellBase RESTful API. They provide fast programmatic access for genome-scale data analysis, therefore discouraging massive downloads of data to local computers. Currently supported languages include Python, R, Java and JavaScript. A similar design has been used in all of them in order to facilitate their use, external contributions and maintenance. Again, all of them provide an exhaustive API for accessing the whole CellBase RESTful API. Please, refer to the corresponding Tutorials to find details on how to download, install, configure the libraries
- **Using remote RESTful web services**: both GET and POST annotation web services are available (see http://bioinfo.hpc.cam.ac.uk/cellbase/webservices/). **The best way to use of the RESTful Web Services is through the client libraries** implemented for different programming languages. Nevertheless, under certain circumstances it may be required to directly access the RESTful API.Web services based annotation results are returned in the form of JSON objects.
- **Using the Java command line**: current Java CLI can connect to either remote web services or efficiently fetch annotation data directly from a custom installation of the database. Even when connecting to remote web services, the annotation CLI provides a lightweight efficient multi-threaded implementation which outperforms other local variant annotators (see _Benchmark_ results below)

The typical input for the CellBase variant annotator will be a VCF file, although the CLI also offers the possibility to explicitly provide a short list of variants as an argument for fast annotation. Two different output formats can be currently generated by the annotator: a .json file with a list of VariantAnnotation objects (see Variant and VariantAnnotation models at https://github.com/opencb/biodata/tree/develop/biodata-models/src/main/resources/avro), or a tab separated values file with the VEP formatted output.

Data sources

Data provided by the variant annotator is the result of integrating most of the annotations available at the CellBase knowledge base: ENSEMBL's core transcript annotation such as location, id, strand, biotype,etc.; protein annotation provided by UniProt, InterPro, SIFT and PolyPhen; population frequencies provided by the European Variation Archive for The 1000 Genomes Project Phase 3, The Exome Server Project (EVS), The Exome Aggregation Consortium v3 (ExaC), gnomAD exomes, gnomAD genomes and The Genomes of the Netherlands (GoNL); sequence conservation from PhastCons and PhyloP; gene expression values from The Genome Expression Atlas; gene drug interaction data from The Drug Gene Interaction Database (DGIdb) and the Human Phenotype Ontology database (HPO); clinical variants annotation from ClinVar. Sequence effect prediction is also calculated on the fly and described by Sequence Ontology (SO) terms. We are constantly working to integrate new data sources in the knowledgebase.

Benchmark

Exhaustive comparison of sequence effect predictions was made against VEP (83) results for the whole 1000 Genome Phase 3 variant set (83 million variants, 346 million effect predictions), yielding a 99.999% of concordance with Ensembl VEP Consequence Types.

- VEP annotations: 346M
- CellBase annotations: 346M
- Coincidence at SO term level (346M annotations)
    - Annotations provided by VEP and not provided by CellBase:
        - 3364 (99.999% coincidence)
        - 61% (2060) of these due to differences on miRNA data sources
        - 39% Difficulties with VEP output format parsing
    - Annotations provided by CellBase and not provided by VEP:
        - 4918 (99.999% coincidence)
        - 60% (2970) of these due to differences on miRNA data sources
        - Difficulties with VEP output format parsing
- Coincidence at variant level (83M variants)
    - Variants with conflicting annotation: 4990 (99.994% coincidence)

Custom annotations

CellBase variant annotations can be complemented with custom annotations provided by the user. The variant annotation CLI allows to provide a VCF file with custom annotation in the INFO column.

How to annotate variants

Please, refer to the VCF and Variant Annotation tutorial.