Ensembl Variant Effect Predictor – flexible and consistent molecular consequence prediction

Irina M. Armean, Laurent Gil, Diana Lemos, Andrew Parton, Helen Schuilenburg, Anja Thormann, Sarah E. Hunt, Paul Flicek, Fiona Cunningham

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

EMBL-EBI

Ensembl Variant Effect Predictor (VEP)

The prediction of molecular consequence of a variant is essential for both basic and clinical research.

The Ensembl VEP:

- Open-source toolset for genomic **variant annotation** and **interpretation**
- Uses the extensive Ensembl **transcriptomic**, **regulatory** and **variation** data to predict **consequences** of variants
- Reports **allele frequency** data from reference projects and results of multiple **pathogenicity predictors**

Standardisation

Standardised reporting is crucial for data sharing across tools and platforms and having a community agreed approach is essential.

- Variant consequences are reported using Sequence Ontology terms
- The output is in standardised VCF format agreed with SnpEff and ANNOVAR
- Participant in the GA4GH Variant Representation and Variant Annotation work groups

The Sequence Ontology

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

Allele normalisation

Sharing and merging data from multiple sources such as allele frequencies and phenotypic annotation relies on unambiguously defined variants and their allele's representation.

Despite its crucial role, there are different ways one can perform allele normalisation especially for insertions and deletions in repeat regions.

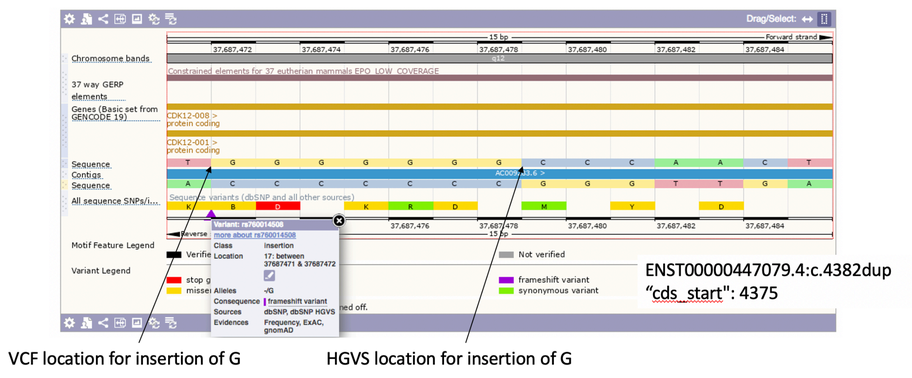
For example:

- VCF: left aligned
- HGVS: 3' aligned

VEP shift_3prime

New in VEP: The VEP '–-shift_3prime' option will right align variants relative to their associated transcripts prior to consequence calculation.

- 2.5% of ClinVar alleles would be shifted



https://www.ensembl.org/info/docs/tools/vep/script/vep_options.html

Access points

The Ensembl VEP tool functionality can be accessed via multiple routes:

- The Ensembl REST API using the VEP endpoint
- Using the VEP web interface
- Using the command line tool

https://rest.ensembl.org/

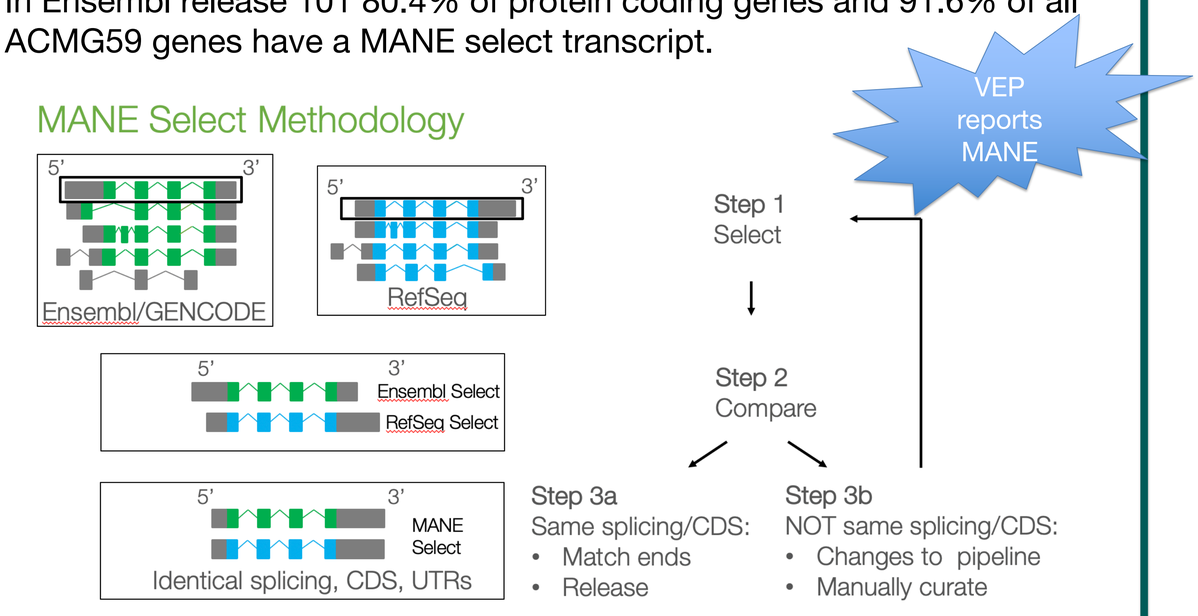
https://www.ensembl.org/vep

Transcripts Sets

The Ensembl VEP can annotate variants using Ensembl/GENCODE, RefSeq and custom gene annotation.

Standardised reporting is one of the aims for the RefSeq and Ensembl joint MANE (Matched Annotation from NCBI and EMBL-EBI) project. The MANE Select transcript set is to include one well-supported transcript per protein coding locus.

In Ensembl release 101 80.4% of protein coding genes and 91.6% of all ACMG59 genes have a MANE select transcript.



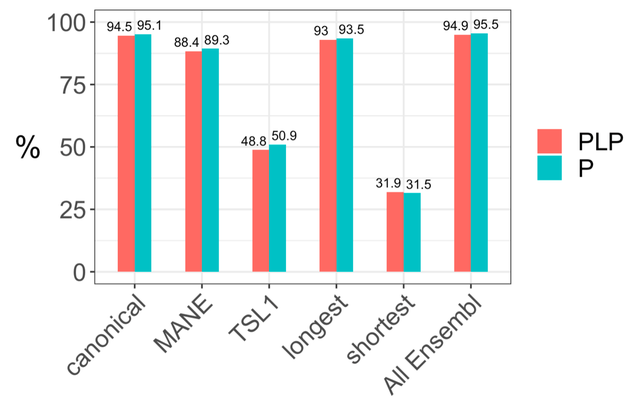
Data used for transcript selection:

- Ensembl/GENCODE: conservation, expression, length, representation in UniProt and RefSeq and clinical information if available.
- RefSeq: expression, conservation, representation in UniProt and Ensembl, length, prior manual curation (LRG)

Impact of transcript choice

Variant molecular consequence prediction is highly dependent on the transcript set used. It is therefore crucial to use all transcripts to not miss any possible annotation!

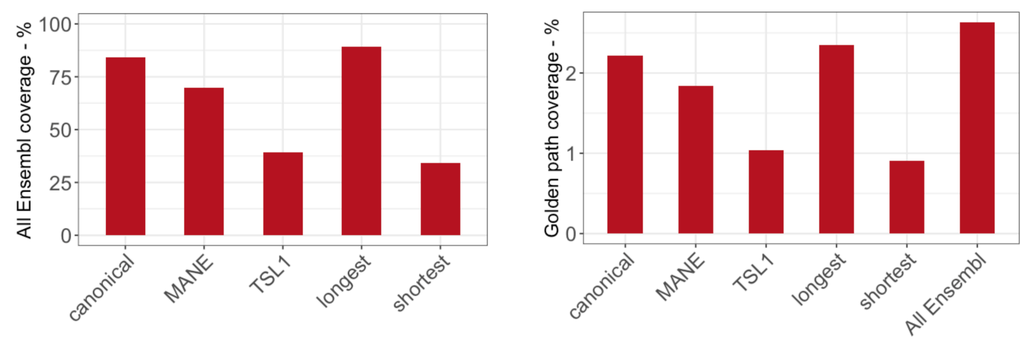
Experiment: given different **protein coding** transcript sets, how many of the pathogenic (P) or pathogenic and likely pathogenic (PLP) ClinVar variants will be annotated as **missense** or **protein truncating variants** (PTVs)?



Left: percent of pathogenic and likely pathogenic variants annotated as missense or PTVs based on different transcript sets.

Data: ClinVar VCF release 24.08.2020, transcript sets Ensembl release 101

The highest coverage of 95.5% is achieved when all transcripts for a gene are used compared to using only the canonical transcript 95.1% or longest transcript 93.5%.



Genomic coverage of different protein coding transcript subsets in respect of the genomic regions annotated as within protein coding transcripts in Ensembl release 101 (left) and in relation to the golden path for GRCh38.p13 (right).

Ensembl receives majority funding from Wellcome (grant number WT108749/Z/15/Z) with additional funding for specific project components from the National Human Genome Research Institute (U41HG007823 and 2U41HG007234), the Biotechnology and Biological Sciences Research Council (BB/N019563/1, BB/M011615/1, BB/S020152/1, BB/P016855/1, BB/S02011X/1 and BB/P024602/1), Open Targets, Wellcome (WT104947/Z/14/Z, WT200990/Z/16/Z, WT201535/Z/16/Z, WT108749/Z/15/A, WT212925/Z/18/Z, WT218328/B/19/Z, WT201535/Z/16/Z), British Council (414710385), ELIXIR: the research infrastructure for life-science data, and the European Molecular Biology Laboratory. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 733161 (MultipleMS), n° 731060 (INFRAVEC2), n° 825575 (EJP RD), n° 817923 (AQUA-FAANG), n° 817998 (GENESWitCH) and n° 815668 (BovReg). We also receive funding from the 'Save the Tasmanian Devil Program'.

European Bioinformatics Institute
Wellcome Genome Campus
Hinxton, Cambridge, CB10 1SD, UK

Tel. +44 (0) 1223 494 444
comms@ebi.ac.uk
www.ebi.ac.uk