Annotating and prioritising genomic variants using the Ensembl Variant Effect Predictor - a tutorial

Sarah Hunt[1], Benjamin Moore[1], M. Amode[1], Irina Armean[1], Diana Lemos[1], Aleena Mushtaq[1], Andrew Parton[1], Helen Schuilenburg[1], Michał Szpak[1], Anja Thormann[1], Emily Perry[1], Stephen Trevanion[1], Paul Flicek[1], Fiona Cunningham[1], and Andrew Yates[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute

June 25, 2021

Abstract

The Ensembl Variant Effect Predictor (VEP) is a freely available, open source tool for the annotation and filtering of genomic variants. It predicts variant molecular consequence using the Ensembl/GENCODE or RefSeq gene sets. It also reports phenotype associations from databases such as ClinVar, allele frequencies from studies including gnomAD, and predictions of deleteriousness from tools such as SIFT and CADD. Ensembl VEP includes filtering options to customise variant prioritisation. It is well supported and updated roughly quarterly to incorporate the latest gene, variant and phenotype association information. Ensembl VEP analysis can be performed using a highly configurable, extensible command-line tool, a Representational State Transfer (REST) application programming interface (API) and a user-friendly web interface. These access methods are designed to suit different levels of bioinformatics experience and meet different needs in terms of data size, visualisation and flexibility. In this tutorial, we will describe performing variant annotation using the Ensembl VEP web tool, which enables sophisticated analysis through a simple interface.

Annotating and prioritising genomic variants using the Ensembl Variant Effect Predictor - a tutorial

Benjamin Moore, Sarah E Hunt, M. Ridwan Amode, Irina M Armean, Diana Lemos, Aleena Mushtaq, Andrew Parton, Helen Schuilenburg, Michał Szpak, Anja Thormann, Emily Perry, Stephen J Trevanion, Paul Flicek, Andrew D Yates, Fiona Cunningham

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

Grant numbers

Ensembl Variation Resources receive funding from the Wellcome Trust (grant number WT108749/Z/15/Z, WT200990/Z/16/Z, WT201535/Z/16/Z, WT212925/Z/18/Z), the BBSRC (BB/S020152/1) and the European Molecular Biology Laboratory. This project has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n°825575.

Abstract

The Ensembl Variant Effect Predictor (VEP) is a freely available, open source tool for the annotation and filtering of genomic variants. It predicts variant molecular consequence using the Ensembl/GENCODE or RefSeq gene sets. It also reports phenotype associations from databases such as ClinVar, allele frequencies from studies including gnomAD, and predictions of deleteriousness from tools such as SIFT and CADD. Ensembl VEP includes filtering options to customise variant prioritisation. It is well supported and updated roughly quarterly to incorporate the latest gene, variant and phenotype association information.

Ensembl VEP analysis can be performed using a highly configurable, extensible command-line tool, a Representational State Transfer (REST) application programming interface (API) and a user-friendly web interface. These access methods are designed to suit different levels of bioinformatics experience and meet different needs in terms of data size, visualisation and flexibility. In this tutorial, we will describe performing variant annotation using the Ensembl VEP web tool, which enables sophisticated analysis through a simple interface.

Keywords

Variant annotation, filtering, VEP, "molecular consequence", variant prioritisation

Main Text

Introduction

Genome and exome sequencing are becoming routine in clinical research and diagnostic settings, as an individual's genotype may provide insight into disease mechanism, progression and treatment. Each sequenced genome contains 4.1 to 5.0 million variant sites (1000 Genomes Project Consortium et al., 2015), many of which will be rare but benign alleles, so additional information is required to enable variant interpretation and prioritisation. As the scale of data production increases, robust and efficient software tools are needed to support variant annotation and filtering.

Variant interpretation requires i) the mapping of variants to transcripts and predictions of molecular consequence; ii) the consideration of all current knowledge relating to a variant and iii) the application of predictive algorithms to evaluate impact of change at the locus. Appropriate resources are available: the reference gene sets are regularly updated; the number of assertions of phenotype association in the literature and in key databases continues to grow; population frequency studies expand to include more individuals and report more detailed catalogues of rare variants and variant pathogenicity prediction is an active area of tool development.

In the Ensembl Project (Howe et al., 2021) we create high-quality gene sets, predict genomic regions involved in gene regulation and collate large-scale sets of variant and phenotype association data. Ensembl VEP (McLaren et al., 2016) builds on these resources and integrates results from variant assessment algorithms to enable convenient but extensive variant annotation. We provide regular updates, approximately every 3 months, to both the VEP software and associated data to ensure the latest information can be used for analysis. Here we present a tutorial describing the Ensembl VEP web interface, detailing the available analyses options and filters.

Tutorial

Data Input

Navigate to the Ensembl VEP homepage by clicking on the 'VEP' link in the blue navigation bar in the Ensembl homepage (*https://www.ensembl.org/index.html*). The Ensembl VEP homepage links to the three different VEP interfaces and detailed documentation. Click on 'Launch VEP' to open the web form, which is divided into sections for data input and optional analysis configuration (Figure 1).

The human GRCh38 assembly is selected by default, but a link provides access to a dedicated GRCh37 tool. Other species can be selected using the 'Add/remove species' option. To make the management of multiple analyses simpler, a name can be assigned to the job.

Data can be input by (1) pasting into the text box, (2) uploading a file or (3) by providing a URL for a file on a public server. The text box is suitable for small-scale datasets. To analyse a larger dataset, provide a URL or use the file upload option which supports a maximum file size of 50 megabytes (or around 2 million lines in a compressed VCF).

Ensembl VEP supports a range of data input formats including;

- variant call format (VCF);
- Human Genome Variation Society (HGVS) descriptions (den Dunnen et al., 2016), using Ensembl, RefSeq or LRG accessions;
- variant identifiers (from databases including dbSNP, ClinVar and UniProt);
- ambiguous gene-based descriptions often used in literature (for example 'BRCA2:p.Val2466Ala').

VCF is the standard exchange format used in next-generation sequencing pipelines so Ensembl VEP is optimised to analyse variants in this format.

Transcript set selection

Predicting the molecular consequence of a genomic variant is an essential step in interpretation and requires extensive, accurate gene annotation. There are two commonly used human gene sets: Ensembl/GENCODE (Frankish et al., 2021) and RefSeq (O'Leary et al., 2019). Both sets are generated using similar but slightly different evidence and algorithms, and so differ slightly. VEP can analyse variants using either gene set, or the combined group or GENCODE Basic, (which contains a small subset of representative transcripts for each gene). Select your preference in the 'Transcript database to use' section (Figure 1).

The VEP algorithm compares each variant to each transcript in the selected set and reports the relative transcript location of the variant (for example exonic, upstream) with any predicted molecular consequence (for example missense, frameshift). Consequences are described using Sequence Ontology terms (SO; Cunningham et al., 2015) to enable comparison and integration with results from other systems.

Transcript-related identifiers

HUGO Gene Nomenclature Committee (HGNC) gene symbols, versioned transcript accessions and transcript types (for example: AGT, ENST00000366667.6, protein coding respectively) are returned by default. Use the 'Identifiers' section (Figure 2) to add further information, including Ensembl or RefSeq protein identifiers, UniProt protein accessions and HGVS variant descriptions at protein and transcript level to your output.

Frequencies and citations

With over seven hundred million variants in dbSNP (version 154, May 2020) alone, the majority of variants found in an individual will have already been described. This information can be crucial to interpretation. Ensembl VEP searches databases including dbSNP, COSMIC and HGMD and reports any variants at the same location as your input variants. For databases with redistribution restrictions, variants are matched on location alone (i.e., with no allele specificity) and names are reported. For fully open databases, variants are matched by allele and key additional information is reported. By default, we only report matches to variants passing our quality filtering (for example, those mapping to multiple genomic locations are excluded); to include all variants in the search check the 'Include flagged variants' option.

In rare disease studies it is useful to filter out variants using reference population frequencies, as variants common in the general population are less likely to be causative. Use the 'Variants and frequency data' section (Figure 3) section to select the reference dataset to be searched. Allele frequencies from the Genome Aggregation Database (gnomAD; Karczewski et al., 2020) and 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) are currently available.

The American College of Medical Genetics and Genomics (ACMG) guidelines (Richards et al., 2015) uses 5% allele frequency as stand-alone evidence a variant allele is not pathogenic. For a single causative variant, ACMG recommend frequency filters should be selected to be higher than disease prevalence. Filter cut-offs should be higher if it is possible multiple variants are acting together.

Select the 'Variant synonyms' option to display the names of variants in databases such as ClinVar, UniProt and PharmGKB. In your results, the names will be linked to the relevant entries in the source databases,

so the details held in these resources can be examined. Check the 'PubMed identifiers' button to return a list of any publications describing the variant with links to full text resources where available. Citation and synonym information is matched on variant name or location and is not allele specific.

Transcript Selection

Transcriptomic sequencing from multiple tissues has resulted in the annotation of increasing numbers of transcript isoforms for many genes. Assessing large numbers of predictions for each variant is time-consuming but important to ensure no information is missed. To support downstream filtering VEP reports transcript type (such as protein coding or pseudogene) and, for Ensembl transcripts, two prioritisation metrics. Transcript Support Level (TSL) summarises the amount of evidence supporting a transcript into a numeric score. APPRIS (Rodriguez et al., 2017) identifies principal transcript isoforms for genes in vertebrate species using protein structural information, functionally important residues and evidence from cross-species alignments. These options are listed in the 'Transcript annotation' section and are reported in Ensembl VEP results by default.

MANE (Matched Annotation from NCBI and EMBL-EBI) transcripts are also reported by default to facilitate transcript prioritisation. MANE Select transcripts are single representative transcripts for each protein coding human gene, chosen by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) and the National Center for Biotechnology Information (NCBI). They are recommended as the default transcript where one is needed for reporting. An additional transcript is required to report all clinically relevant variants in a small number of genes, including LAMA3 and SCN2A. MANE Plus Clinical transcripts are being assigned to meet this need. MANE transcripts are identical between the RefSeq and Ensembl/GENCODE sets and match the GRCh38 reference genome sequence. MANE Select transcripts are available for 78% of protein coding genes and MANE Plus Clinical transcripts for 55 genes in Ensembl release 104 (May 2021). Selection of the MANE option flags these recommended transcripts and reports both RefSeq and Ensembl transcript identifiers.

The Ensembl canonical transcript is a single default transcript available for every gene, in every species. The same Ensembl algorithm is used to pick MANE Select transcript and the canonical transcript in human, so the two are the same where a MANE Select exists. Check the 'Identify canonical transcripts' option to highlight these transcripts in your results if you require a default for every gene.

Protein domains

When a variant maps to the protein, understanding which domain it falls in can provide clues as to possible impact on function. InterPro is an integrated resource for protein families, domains and sites, combining information from several different protein signature databases. We run InterProScan (Jones et al., 2014) on all Ensembl protein sequences to identify domains and these are reported in VEP. Check the 'Protein domains' option (Figure 4) to report these results and any overlapping PDBe structures.

Regulatory elements

Variants in the non-coding regions of the genome are more difficult to interpret than those falling within genes, and are also important in disease (Zhang et al., 2015). In the Ensembl Project, we use data from large scale projects including ENCODE, IHEC and Blueprint, to predict regions in the human genome that influence gene regulation. We classify them into types such as 'promoter' and 'enhancer' (Zerbino et al., 2015). Select the 'Regulatory data' option (Figure 4) to identify where your variants overlap such regions. This analysis can be configured to report all results or only those from specific cell types.

Phenotype and disease associations

Access to phenotype or disease associations previously reported for your variants or the genes they overlap is essential. There is a large body of information available in different databases but performing multiple

searches across different resources is time consuming. In Ensembl, we aggregate phenotype and disease associations from a variety of sources, including Orphanet, the Cancer Gene Census, OMIM, ClinVar and the NHGRI-EBI GWAS Catalog, into a standardised format (Hunt et al., 2018). This information is searched by Ensembl VEP and summary information reported. ClinVar assertions of variant clinical significance are reported by default and, importantly, these are matched by allele and not just variant location. Select the 'Phenotypes' option (Figure 4) to retrieve a list of phenotype associations for overlapping genes and previously reported variants, with links to fuller information.

Results from additional sources are available. DisGeNET (Piñero et al., 2020) is a database of gene and variant disease associations. Select this option to view summary results including disease names and PubMed identifiers, which are linked to full text publications. The Mastermind Genomic Search Engine (Chunn et al., 2020) (https://www.genomenon.com/mastermind) holds gene, variant, disease, phenotype and therapy evidence mined from millions of scientific articles. Select this option to return links to the Mastermind website, which is free to access with registration.

Prediction packages

An increasing number of pathogenicity scoring algorithms are being developed to aid variant interpretation. It must however be remembered that predictions often use the same training sets and/or evidence so agreement between two algorithms does not necessarily provide additional evidence for a rating. We calculate scores for all possible amino acid substitutions in all Ensembl proteins using SIFT (Kumar et al., 2009) and PolyPhen-2 (Adzhubei et al., 2010). These results are returned by default.

dbNSFP, the database for nonsynonymous SNPs' functional predictions (Liu at al., 2020) contains pre-calculated scores for over 20 algorithms. Select this option (Figure 5), to browse the 'Fields to include' menu and configure the precise results set to be returned. Combined Annotation-Dependent Depletion (CADD; Rentzsch et al., 2019) is a framework for scoring the deleteriousness of genomic variants using a wide range of different information including conservation, functional information and protein level pathogenicity predictions. Select this option to view scores for variants in both coding and non-coding loci.

Variants which disrupt splicing have also been implicated in human disease (Ward et al., 2010). We optionally report results from the well-established MaxEntScan (Yeo et al., 2014); SpliceAI (Jaganathan et al., 2019), which takes a machine learning approach; and the ensemble scores provided in the dbscSNV (Liu et al., 2020) database. Select these options in the 'Splicing predictions' section (Figure 5).

Filtering and Advanced options

The options in these sections will not be required for the majority of analyses. The 'Filters' section (Figure 6) allows the results returned to be restricted by allele frequency, to contain only variants in coding sequence or to be reduced to a subset of the available variant-transcript combinations. However, we recommend instead to filter results after the analysis, which allows greater flexibility. The 'Advanced options' allow you to change the way VEP analyses variants internally (a smaller batch size will reduce memory requirements but increase run time) and control whether insertion and deletions in repetitive sequence are expressed at their most 3' position prior to consequence evaluation.

Results

Having configured your analysis, click the 'Run' button at the bottom of the form. Analysis jobs run on our compute farm and the time required will depend on the number of input variants and range of options chosen. The 'Recent jobs' table displays the status of all your analyses and has options to edit and resubmit, share or discard jobs. Results can be saved by logging into an Ensembl account. Once a job has the status of 'Done', clicking on 'View Results' will display the results table.

Summary statistics and charts display an overview of the results on the output page (Figure 7). There is also a table with a preview of the detailed results and a simple interface to configure filtering of the output.

To aid variant prioritisation, multiple filters can be combined using basic logical relationships, allowing the creation of complex customised queries. For example, 'Consequence is protein_altering_variant' plus 'CADD PHRED >=30' plus 'gnomAD AF is not defined' will report variants which are predicted to change protein sequence, are in the 0.1% most deleterious changes predicted by CADD and are not seen in the gnomAD exome variant set. Importantly, we report the most specific SO term but enable querying by parent terms. For example, when the consequence of 'protein altering variant' is selected, missense and frameshift variants are reported.

The results interface allows you to download your output in VCF and other formats for further analysis or export the variation or gene list to the Ensembl BioMart tool to extract additional data, such as gene homologues and sequences.

Results are displayed in a table (Figure 8) with a single line per combination of variant allele and transcript or regulatory element. Click on the "Show/hide columns" button to configure which columns are displayed if you wish to view a subset of the results. Cells containing many records (as can happen for example for PubMed IDs) will initially be compressed and need expanding to view. The results table displays only a summary of the information available for a variant. You can easily examine evidence for your variants of interest in greater detail. Links enable you to access relevant publications in Europe PMC or view details in resources such as UniProt, ClinVar and PDBe. The table is also a convenient access point to data held in Ensembl: it has links to the variant location on the genome browser and detailed information about any genes, transcripts or variants the input variant overlaps.

Ensembl VEP interfaces

The Ensembl VEP web tool enables analysis configuration and results filtering via a simple interface. It is ideal for analysing small sets of variants and interactively assessing the results. We provide two other interfaces that are more appropriate for the integration of VEP annotations in web views or for large scale analyses. Here we briefly describe these REST and command line interfaces.

Language-agnostic computational access to VEP analysis is available through the Ensembl REST API. The VEP REST service (https://rest.ensembl.org) supports similar options to the web tool and is suitable for programmatic integration into web pages or analysis pipelines. HGVS notation, position and allele-based descriptions and a range of common variant names are supported as input and up to 200 variants can be submitted in a single request.

The command line tool is the most powerful and flexible way to use Ensembl VEP. It supports more analysis options than the other interfaces. There is also no limit on input file size, making it suitable for the annotation of large variant sets identified through whole genome sequencing. The use of custom gene, variant and other annotation sets is supported, enabling analysis against private data. While VEP can be run by anyone comfortable with command line tools, those with basic programming skills can simply create extensions to add novel, custom functionality. Run time depends on the number and complexity of options selected: basic analysis of a whole exome (~200,000 variants) takes under 5 minutes while a single genome (~4.5 million variants) will take around an hour. A Docker image is available to simplify installation. A results-filtering tool is also available in the Ensembl VEP command line package. Full instructions for installation and options for running Ensembl VEP locally can be found in our online documentation (https://www.ensembl.org/vep).

Conclusion

The Ensembl VEP web tool enables the flexible configuration of variant analysis from an extensive range of options via a simple interface. It allows customisable filtering so you can interrogate and understand your results. It links out to detailed resources, both within the Ensembl browser and other key websites. The regular updating of the reference data and analysis tools supported within Ensembl VEP make it an essential tool for variant annotation, filtering and prioritisation.

Acknowledgments

We thank members of the Ensembl team for gene, regulatory and comparative genomics annotation, and web development. We thank previous team members, in particular William McLaren and Laurent Gil, for their contributions to Ensembl VEP. We also wish to thank the of EMBL-EBI's technical services cluster for their support and the VEP community who have helped to improve Ensembl VEP by suggesting new functionality, giving feedback and bug reports.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature* , *526* (7571), 68–74. https://doi.org/10.1038/nature15393

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods* , *7* (4), 248–249. https://doi.org/10.1038/nmeth0410-248

Chunn, L. M., Nefcy, D. C., Scouten, R. W., Tarpey, R. P., Chauhan, G., Lim, M. S., Elenitoba-Johnson, K., Schwartz, S. A., & Kiel, M. J. (2020). Mastermind: A Comprehensive Genomic Association Search Engine for Empirical Evidence Curation and Genetic Variant Interpretation.*Frontiers in genetics* , *11* , 577152. https://doi.org/10.3389/fgene.2020.577152

Cunningham, F., Moore, B., Ruiz-Schultz, N., Ritchie, G. R., & Eilbeck, K. (2015). Improving the Sequence Ontology terminology for genomic variant annotation. *Journal of biomedical semantics* , *6* , 32. https://doi.org/10.1186/s13326-015-0030-4

Format:den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., Roux, A. F., Smith, T., Antonarakis, S. E., & Taschner, P. E. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human mutation* , *37* (6), 564–569. https://doi.org/10.1002/humu.22981

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology* ,*6* (5), R44. https://doi.org/10.1186/gb-2005-6-5-r44

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., ... Flicek, P. (2021). GENCODE 2021. *Nucleic acids research* , *49* (D1), D916–D923. https://doi.org/10.1093/nar/gkaa1087

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., Gall, A., ... Flicek, P. (2021). Ensembl 2021. *Nucleic acids research* ,*49* (D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., Parton, A., Armean, I. M., Trevanion, S. J., Flicek, P., & Cunningham, F. (2018). Ensembl variation resources. *Database : the journal of biological databases and curation* , *2018* , bay119. https://doi.org/10.1093/database/bay119

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., & Farh, K. K. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell* ,*176* (3), 535–548.e24. https://doi.org/10.1016/j.cell.2018.12.015

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)* , *30* (9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans.*Nature* , *581* (7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* , *4* (7), 1073–1081. https://doi.org/10.1038/nprot.2009.86

Liu, X., Li, C., Mou, C., Dong, Y., & Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome medicine* , *12* (1), 103. https://doi.org/10.1186/s13073-020-00803-9

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology* , *17* (1), 122. https://doi.org/10.1186/s13059-016-0974-4

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., ... Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.*Nucleic acids research* , *44* (D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research* , *48* (D1), D845–D855. https://doi.org/10.1093/nar/gkz1021

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research* , *47* (D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., & ACMG Laboratory Quality Assurance Committee (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* , *17* (5), 405–424. https://doi.org/10.1038/gim.2015.30

Rodriguez, J. M., Rodriguez-Rivas, J., Di Domenico, T., Vázquez, J., Valencia, A., & Tress, M. L. (2018). APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic acids research* , *46* (D1), D213–D217. https://doi.org/10.1093/nar/gkx997

Ward, A. J., & Cooper, T. A. (2010). The pathobiology of splicing.*The Journal of pathology* , *220* (2), 152–163. https://doi.org/10.1002/path.2649

Yeo, G., & Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology: a journal of computational molecular cell biology* , *11* (2-3), 377–394. https://doi.org/10.1089/1066527041410418

Zhang, F., & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human molecular genetics* , *24* (R1), R102–R110. https://doi.org/10.1093/hmg/ddv259

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., & Flicek, P. R. (2015). The ensembl regulatory build. *Genome biology* ,*16* (1), 56. https://doi.org/10.1186/s13059-015-0621-5

Figure legends

Figure 1. The Ensembl VEP web interface showing species/assembly selection, data input, transcript set selection and additional groups of configuration options.

Figure 2. The 'Identifiers' section which allows the selection of gene, protein and HGVS identifiers.

Figure 3. The 'Variants and frequency data' section which allows the selection of information known about variants at the same location.

Figure 4. The 'Additional annotations' section which allows the selection of transcript, protein domain, regulatory region and phenotype annotations.

Figure 5. The 'Predictions' section, which allows the selection of different pathogenicity, splicing and conservation predictions.

Figure 6. Filtering and advanced options

Figure 7. The results page with summary statistics and options for filtering and downloading the results table.

Figure 8. The results table showing predicted molecular consequences and links to the location and overlapping genes and variant displays within the Ensembl genome browser.

Conflicts of interest statement

Paul Flicek is a member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd.

Data Availability Statement

No new data were created or analysed in this study.

Publicly available data is integrated into the Ensembl variation resources. Reference data packaged for use in Ensembl VEP is available from our FTP site in release-specific directories for example: http://ftp.ensembl.org/pub/release-103/variation/vep/.

The Ensembl VEP command line tool is available from https://github.com/Ensembl/ensembl-vep

The Ensembl VEP plugins are available from https://github.com/Ensembl/VEP_plugins

Ensembl VEP plugins are created to integrate datasets with redistribution restrictions. These plugins contain full instructions for data collection and formatting. We have here described the use of the following data sets via plugins:

CADD (https://cadd.gs.washington.edu/download)

dbNSFP (ftp://dbnsfp:dbnsfp@dbnsfp.softgenetics.com/dbNSFP4.2a.zip)

dbscSNV (https://drive.google.com/file/d/0B60wROKy6OqcQ0IyYnh5bmdHMW8/view)

DisGeNET (https://www.disgenet.org/downloads)

Mastermind (https://www.genomenon.com/cvr/)

SpliceAI (https://pypi.org/project/spliceai/)