Variant Annotation Integrator

Select Genome Assembly and Region

clade: Mammal | genome: Human | assembly: Dec. 2013 (GRCh38/hg38)

region to annotate: genome

Select Variants

variants: User Supplied Track

maximum number of variants to be processed: 10,000

manage custom tracks | track hubs | To reset **all** user cart settings (including custom tracks), click here.

Select Genes

The gene predictions selected here will be used to determine the effect of each variant on genes, for example intronic, missense, splice site, intergenic etc.

GENCODE v29 Comprehensive Transcript Set (only Basic displayed by default)

Select Regulatory Annotations

The annotations in this section provide predicted regulatory regions based on various experimental data. When a variant overlaps an annotation selected here, the consequence term regulatory_region_variant will be assigned. Follow the links to description pages that explain how each dataset was constructed. Some datasets cover a significant portion of the genome and it may be desirable to filter these annotations by cell type and/or score in order to avoid an overabundance of hits.

- DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
  - **filter items**

Select More Annotations (optional)

Database of Non-synonymous Functional Predictions (dbNSFP)

dbNSFP (Liu *et al.* 2015) release 3.1a provides pre-computed scores and predictions of functional significance from a variety of tools. Every possible coding change to transcripts in Gencode release 22 (Ensembl 79, Mar. 2015) gene predictions has been evaluated. *Note: This may not encompass all transcripts in your selected gene set.*

Set all | Clear all

- Variant Effect Scoring Tool (VEST) (scores [0-1] predict confidence that a change is deleterious
- SIFT (D = damaging, T = tolerated)
- PolyPhen-2 with HumDiv training set (D = probably damaging, P = possibly damaging, B = benign)
- PolyPhen-2 with HumVar training set (D = probably damaging, P = possibly damaging, B = benign)
- MutationTaster (A = disease causing automatic, D = disease causing, N = polymorphism, P = polymorphism automatic)
- MutationAssessor (high or medium: predicted functional; low or neutral: predicted non-functional)
- Likelihood ratio test (LRT) (D = deleterious, N = Neutral, U = unknown)
- InterPro protein domains
- GERP++ Rejected Substitutions (RS)
- GERP++ Neutral Rate (NR)

Transcript status

HGVS variant nomenclature

The Human Genome Variation Society (HGVS) has established a sequence variant nomenclature, an international standard used to report variation in genomic, transcript and protein sequences.
Select RefSeq Genes or an official GENCODE release ("Basic Gene Annotation Set from GENCODE..." or "Comprehensive Gene Annotation Set...") in the "Select Genes" section above in order to make options appear.

Known variation

- Include dbSNP rs# ID if one exists

Conserved elements

- 100 vertebrates Conserved Elements
- 7 vertebrates Conserved Elements
- 20 mammals (17 primates) Conserved Elements
- 30 mammals Conserved Elements (27 primates)

Conservation scores

- 100 vertebrates conservation by PhastCons
- 100 vertebrates Basewise Conservation by PhyloP
- 7 vertebrates conservation by PhastCons
- 7 vertebrates Basewise Conservation by PhyloP
- 20 mammals (17 primates) conservation by PhastCons
- 20 mammals (17 primates) Basewise Conservation by PhyloP
- 30 mammals conservation by PhastCons (27 primates)
- 30 mammals Basewise Conservation by PhyloP (27 primates)

Define Filters

Functional role

Conservation

Configure Output

output format: Variant Effect Predictor (tab-separated text)

output file: fromVAI.txt (leave blank to keep output in browser)

file type returned: plain text | gzip compressed (ignored if output file is blank)

Get results

Using the Variant Annotation Integrator

Introduction

The Variant Annotation Integrator (VAI) is a research tool for associating annotations from the UCSC database with your uploaded set of variant calls. It uses gene annotations to predict functional effects of variants on transcripts. For example, a variant might be located in the coding sequence of one transcript, but in the intron of an alternatively spliced transcript of the same gene; the VAI will return the predicted functional effect for each transcript. The VAI can optionally add several other types of relevant information: the dbSNP identifier if the variant is found in dbSNP, protein damage scores for missense variants from the Database of Non-synonymous Functional Predictions (dbNSFP), and conservation scores computed from multi-species alignments. The VAI can optionally filter results to retain only specific functional effect categories, variant properties and multi-species conservation status.

NOTE:
The VAI is only a research tool, meant to be used by those who have been properly trained in the interpretation of genetic data, and should never be used to make any kind of medical decision. We urge users seeking information about a personal medical or genetic condition to consult with a qualified physician for diagnosis and for answers to personal questions.

Submitting your variant calls

In order to use the VAI, you must provide variant calls in either the Personal Genome SNP (pgSnp) or VCF format. pgSnp-formatted variants may be uploaded as a Custom Track. Compressed and indexed VCF files must be on a web server (HTTP, HTTPS or FTP) and configured as Custom Tracks, or if you happen to have a Track Hub, as hub tracks.

Protein-coding gene transcript effect predictions

Any gene prediction track in the UCSC Genome Browser database or in a track hub can be selected as the VAI's source of transcript annotations for prediction of functional effects. Sequence Ontology (SO) terms are used to describe the effect of each variant on genes in terms of transcript structure as follows:

SO term	description
intergenic_variant	A sequence variant located in the intergenic region, between genes.
upstream_gene_variant	A sequence variant located 5' of a gene. (VAI searches within 5,000 bases.)
downstream_gene_variant	A sequence variant located 3' of a gene. (VAI searches within 5,000 bases.)
5_prime_UTR_variant	A variant located in the 5' untranslated region (UTR) of a gene.
3_prime_UTR_variant	A variant located in the 3' untranslated region (UTR) of a gene.
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid.
missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved.
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence.
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence.
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three.
initiator_codon_variant	A codon variant that changes at least one base of the first codon of a transcript

initiator_codon_variant	A codon variant that changes at least one base of the first codon of a transcript.
incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed.
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript.
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains.
exon_loss	A sequence variant whereby an exon is lost from the transcript. (VAI assigns this term when an entire exon is deleted.)
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript.
NMD_transcript_variant	A variant in a transcript that is already the target of nonsense-mediated decay (NMD), i.e. stop codon is not in last exon nor within 50 bases of the end of the second-to-last exon.
intron_variant	A transcript variant occurring within an intron.
splice_donor_variant	A splice variant that changes the 2-base region at the 5' end of an intron.
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron.
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron.
complex_transcript_variant	A transcript variant with a complex insertion or deletion (indel) that spans an exon/intron border or a coding sequence/UTR border.
non_coding_exon_variant	A sequence variant that changes exon sequence of a non-coding gene.
no_sequence_alteration	A variant that causes no change to the transcript sequence and/or specifies only the reference allele, no alternate allele. In rare cases when the transcript sequence (e.g. from RefSeq) differs from the reference genome assembly, a difference from the reference genome may restore the transcript sequence instead of altering it.

Optional annotations

In addition to protein-coding genes, some genome assemblies offer other sources of annotations that can be included in the output for each variant.

Database of Non-synonymous Functional Predictions (dbNSFP)

dbNSFP annotations are available only for hg19/GRCh37 (dbNSFP release 2.0) and hg38/GRCh38 (release 3.1a). dbNSFP (Liu *et al.* 2011) provides pre-computed scores and predictions of functional significance from a variety of tools. Every possible coding change to transcripts in GENCODE (for hg19: release 9, Ensembl 64, Dec. 2011; for hg38, release 22, Ensembl 79, Mar. 2015) gene predictions has been evaluated. dbNSFP includes only single-nucleotide missense changes; its data do not apply to indels, multi-nucleotide variants, non-coding or synonymous changes.

dbNSFP provides scores and predictions from several tools that use various machine learning techniques to estimate the likelihood that a single-nucleotide missense variant would damage a protein's structure and function:

- **SIFT (Sorting Intolerant From Tolerant)** uses sequence homology and the physical properties of amino acids to predict whether an amino acid substitution affects protein function. Scores less than 0.05 are classified as Damaging ("D" in output); higher scores are classified as Tolerated ("T"). (Ng and Henikoff, 2003)
- **PolyPhen-2 (Polymorphism Phenotyping v2)** applies a naive Bayes classifier using several sequence-based and structure-based predictive features including refined multi-species alignments. PolyPhen-2 was trained on two datasets, and dbNSFP provides scores for both. The HumDiv training set is intended for evaluating rare alleles potentially involved in complex phenotypes, for example in genome-wide association studies (GWAS). Predictions are derived from scores, with these ranges for HumDiv: "probably damaging" ("D") for scores in [0.957, 1]; "possibly damaging" ("P") for scores in [0.453, 0.956]; "benign" ("B") for scores in [0, 0.452]. HumVar is intended for studies of Mendelian diseases, for which mutations with drastic effects must be sorted out from abundant mildly deleterious variants. Predictions are derived from scores, with these ranges for HumDiv: "probably damaging" ("D") for scores in [0.909, 1]; "possibly damaging" ("P") for scores in [0.447, 0.908]; "benign" ("B") for scores in [0, 0.446]. (Adzhubei *et al.*, 2010)
- **MutationTaster** applies a naive Bayes classifier trained on a large dataset (>390,000 known disease mutations from HGMD Professional and >6,800,000 presumably harmless SNP and Indel polymorphisms from the 1000 Genomes Project). Variants that cause a premature stop codon resulting in nonsense-mediated decay (NMD), as well as variants marked as probable-pathogenic or pathogenic in ClinVar, are automatically presumed to be disease-causing ("A"). Variants with all three genotypes present in HapMap or with at least 4 heterozygous genotypes in 1000 Genomes are automatically presumed to be harmless polymorphisms ("P"). Variants not automatically determined to be disease-causing or polymorphic are predicted to be "disease-causing" ("D") or polymorphisms ("N") by the classifier. Probability scores close to 1 indicate high "security" of the prediction; probabilities close to 0 for an automatic prediction ("A" or "P") can indicate that the classifier predicted a different outcome. (Schwarz *et al.*, 2010)
- **MutationAssessor** uses sequence homologs grouped into families and sub-families by combinatorial entropy formalism to compute a Functional Impact Score (FIS). It is intended for use in cancer studies, in which both gain of function and loss of function are important; the authors also identify a third category, "switch of function." A prediction of "high" or "medium" indicates that the variant probably has some functional impact, while "low" or "neutral" indicate that the variant is probably function-neutral. (Reva *et al.*, 2011)
- **LRT (Likelihood Ratio Test)** uses comparative genomics to identify variants that disrupt highly conserved amino acids. Variants are predicted to be deleterious ("D"), neutral ("N") or unknown ("U"). (Chun and Fay, 2009)
- **VEST (Variant Effect Scoring Tool)** (available only for hg38/GRCh38) uses a classifier that was trained with ~45,000 disease mutations from HGMD and ~45,000 high frequency missense variants (putatively neutral) from the Exome Sequencing Project. (Carter *et al.*, 2013)

In addition, dbNSFP provides InterPro protein domains where available (Hunter *et al.*, 2012) and two measures of conservation computed by GERP++ (Davydov *et al.*, 2010).

Transcript status

Some of the gene prediction tracks have additional annotations to indicate the amount or quality of supporting evidence for each transcript. When the track selected in the "Select Genes" section has such annotations, these can be enabled under "Transcript Status". The options depend on which gene prediction track is selected.

- **GENCODE tags:** when GENCODE Genes are selected in the "Select Genes" section, any GENCODE tags associated with a transcript can be added to output.
- **RefSeq status:** when RefSeq Genes are selected in the "Select Genes" section, the transcript's status can be included in output.
- **Canonical UCSC transcripts:** when UCSC Genes (labeled GENCODE V22 in hg38/GRCh38) are selected in the "Select Genes" section, the flag "CANONICAL=YES" is added when the transcript has been chosen as "canonical" (see the "Related Data" section of the UCSC Genes track description).

Known variation

If the selected genome assembly has a SNPs track (derived from dbSNP), when a variant has the same start and end coordinates as a variant in dbSNP, the VAI includes the reference SNP (rs#) identifier in the output. Currently, the VAI does not compare alleles due to the frequency of strand anomalies in dbSNP.

Conservation

If the selected genome assembly has a Conservation track with phyloP scores and/or phastCons scores and conserved elements, those can be included in the output. Both phastCons and phyloP are part of the PHAST package; see the Conservation track description in the Genome Browser for more details.

Filters

The volume of unrestricted output can be quite large, making it difficult to identify variants of particular interest. Several filters can be applied to keep only those variants that have specific properties.

Functional role

By default, all variants are included in the output regardless of predicted functional effect. If you would like to keep only variants that have a particular type of effect, you can uncheck the checkboxes of other effect types. The detailed functional effect predictions are categorized as follows:

- **intergenic:** intergenic_variant
- **upstream/downstream of gene:** upstream_gene_variant, downstream_gene_variant
- **5' or 3' UTR:** 5_prime_UTR_variant, 3_prime_UTR_variant
- **CDS - synonymous coding change:** synonymous_variant
- **CDS - non-synonymous (missense, stop gain/loss, frameshift etc):** missense_variant, inframe_insertion, inframe_deletion, frameshift_variant, initiator_codon_variant, incomplete_terminal_codon_variant, stop_lost, stop_retained_variant, stop_gained, NMD_transcript_variant
- **intron:** intron_variant
- **splice site or splice region:** splice_donor_variant, splice_acceptor_variant, splice_region_variant
- **exon of non-coding gene:** non_coding_exon_variant

Known variation

(applicable only to assemblies that have "Common SNPs" and "Mult SNPs" tracks) By default, all variants appear in output regardless of overlap with known dbSNP variants that map to multiple locations (a possible red flag), or that have a global minor allele frequency (MAF) of 1% or higher. Those categories of known variants can be used to exclude overlapping variants from output by unchecking the corresponding checkbox.

Conservation

(applicable only to assemblies that have "Conservation" tracks) If desired, output can be restricted to only those variants that overlap conserved elements computed by phastCons.

Output format

Currently, the VAI produces output comparable to Ensembl's Variant Effect Predictor (VEP), in either tab-separated text format or HTML. Columns are described here. When text output is selected, entering an output file name causes output to be saved in a local file instead of appearing in the browser, optionally compressed by gzip (compression reduces file size and network traffic, which results in faster downloads). When HTML is selected, output always appears in the browser window and the output file name is ignored.

Acknowledgments

Anyone familiar with Ensembl's Variant Effect Predictor (VEP) will doubtless notice similarities in options and interface. In collaboration with our colleagues at Ensembl, we have made an effort to limit the differences between the tools by using Sequence Ontology terms to describe variants' functional effects and by creating a "VEP" output format. Any bugs in the VAI, however, are in the VAI only.