BRaVa variant annotation recommendations

General

Genome build: GRCh38; report chromosome, position, reference and alternate

Even where variant frequencies below a certain threshold cannot be reported, we highly encourage the release of the above variant detail to enable cross-referencing across datasets.

Software: VEP 105 (gencode v39) + LOFTEE v1.04_GRCh38

Transcripts: MANE Select as the canonical, where available. Otherwise 'canonical' as the canonical, in protein coding genes. *Note that for VEP 105, this is equivalent to selecting the 'canonical' transcript in protein coding genes.*

Variant specific annotations

Consequence

- LOFTEE v1.04_GRCh38 annotations for all putative LoF variants (HC, LC and flags)
- Consequence with respect to MANE S
- elect (and a small number of 'canonical' transcripts in genes without a MANE select) transcripts (as above)

Allele frequency

- gnomAD: overall, major continental, popmax FAF
- Internal AC, AN and AF, number of homozygotes - split by major continental ancestry, if possible

WGS and WES curation *(BRaVa WES and WGS curation doc that used covid HGI as a starting point)*

In silico deleteriousness

- CADD v1.6
- REVEL, using dbNSFP4.3
- SpliceAI v1.3 (this fork enables variant batching on GPUs which dramatically speeds up computation)
    - Run with updated gencode v39 annotation file using build 38 for protein-coding transcripts.

- Code used to generate the above file. Terminal command and options: `get_gene_id_to_canonical_transcript_id(only_protein_coding=True,is_canonical_transcript=True`. Ensembl database used: v105 (`homo_sapiens_core_105_38`).

Variant quality summary metrics

- Mean depth of coverage at variant sites
- Mean allele balance for heterozygotes; mean allele balance for homozygotes
- Appropriate QC measures to be determined based on QC and filtering approach (see *WGS and WES curation* for suggested QC)

Variant categories for gene-based tests

Note that each variant should receive a **single** annotation. Lower numbers in the following categorisation take precedence, so e.g. if a variant is categorised as HC by LOFTEE and has a SpliceAI DS score ≥ 0.2, it should only receive a **High confidence pLoF** annotation.

1. **High confidence pLoF**: high-confidence LoF variants (LOFTEE HC)
2. **Damaging missense/protein-altering**: any variant not categorised in (1) (High confidence pLoF) with at least one of
    a. Variant annotated as missense/start-loss/stop-loss/in-frame indel and (REVEL ≥ 0.773 or CADD ≥ 28.1 (or both))
    b. Any variant with SpliceAI DS ≥ 0.2 where SpliceAI DS is max(DS_AG, DS_AL, DS_DG, DS_DL), see here.
    c. Low-confidence LoF variants (LOFTEE LC)
3. **Other missense/protein-altering**:
    a. Missense/start-loss/stop-loss/in-frame indel not categorised in (2) (Damaging missense/protein-altering)
4. **Synonymous**: synonymous variants with SpliceAI DS < 0.2 in the gene (control set)

REVEL and CADD score cut-offs based on:
https://www.biorxiv.org/content/10.1101/2022.03.17.484479v1
Finally, remove annotations from all variants with gnomAD Max broad ancestry AF > 0.01.