Author: Ashish Kumar Singh,

Ashish.Kumar.Singh3@stolav.no

LIST of Tools

ENSEMBL-VEP

- Versions:
  - ensembl: 113.58650ec
  - ensembl-compara: 113.9cf749d
  - ensembl-funcgen: 113.e30608c
  - ensembl-io: 113.bee6816
  - ensembl-variation: 113.ebfce74
  - ensembl-vep: 113.0

1. ### Singularity Command to PULL ENSEMBL-VEP Container

   To pull the ENSEMBL-VEP container using Singularity, use the following command:

   ```
   singularity pull --name vep.sif docker://ensemblorg/ensembl-vep
   ```

   Check the sif image:

   ```
   singularity exec vep.sif vep --help
   ```

2. ### Vep cache download

   Use the Cache for same version, i.e., Version:113

   ```
   #GRCh38
   curl -O https://ftp.ensembl.org/pub/release-
   113/variation/vep/homo_sapiens_refseq_vep_113_GRCh38.tar.gz
   tar xzf homo_sapiens_refseq_vep_113_GRCh38.tar.gz

   #GRCh37
   curl -O https://ftp.ensembl.org/pub/release-
   113/variation/vep/homo_sapiens_refseq_vep_113_GRCh37.tar.gz
   tar xzf homo_sapiens_refseq_vep_113_GRCh37.tar.gz
   ```

VEP Plugins

Web link: https://www.ensembl.org/info/docs/tools/vep/script/vep_plugins.html

Final selected Plugins after discussion:

1. Functional effect
    - MaveDB
        - **STATUS: DONE**

```
#Create the respective directory:
    mkdir -p Functional_effect/MaveDB
#Pull the data:
    cd Functional_effect/MaveDB
    wget
https://ftp.ensembl.org/pub/current_variation/MaveDB/MaveDB_variants.ts
v.gz
    wget
https://ftp.ensembl.org/pub/current_variation/MaveDB/MaveDB_variants.ts
v.gz.tbi
```

2. Gene tolerance to change
    - DosageSensitivity
        - **STATUS: DONE**

```
# Pull the data:
    cd Gene_tolerance_to_change/DosageSensitivity
    wget
https://zenodo.org/record/6347673/files/Collins_rCNV_2022.dosage_sensit
ivity_scores.tsv.gz
```

    - **LOEUF {link not working "MANUAL work"}**

```
# Some work will be required to run in GRCh38
#download link
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7334197/bin/41586_2020_230
8_MOESM4_ESM.zip
# LINK IS NOT WORKING!!!

#Checkout this:
https://gnomad.broadinstitute.org/downloads#v4
```

    - **pLI {link not working}**

```

```

```
# Download link:
https://ftp.broadinstitute.org/pub/ExAC_release/release0.3/functional_gene_constraint/fordist_cleaned_exac_r03_march16_z_pli_rec_null_data.txt
# LINK IS NOT WORKING!!!
```

3. Motif
   - **FunMotifs (Note! *Link provided for this plugin is not working*)

```
# Download link
wget http://bioinf.icm.uu.se:3838/funmotifs/
```

4. Nearby features
   - Downstream
     - **STATUS: DONE**

```
# NO additional data is requried but only VEP cache
# VEP command
./vep -i variations.vcf --plugin Downstream
```

   - TSSDistance
     - **STATUS: DONE**

```
# NO additional data is requried but only VEP cache
# VEP command
./vep -i variations.vcf --plugin TSSDistance
```

5. Pathogenicity predictions
   - dbNSFP
     - **STATUS: DONE**

```
# RE-APPLY for licence (as per mail instruction)
https://www.dbnsfp.org/download
```

   - LOFTEE
     - **Note! *Messy implementation***
       - For now we **SKIP** it.

```
# INFO link: https://github.com/konradjk/loftee/tree/grch38
```

6. Phenotype data and citations

- DisGeNET
  - **Need to PURCHASE license for download**

```
# APPLY for licence
https://disgenet.com/academic-apply
```

- Mastermind
  - **Need to PURCHASE license for download**

```
# APPLY for licence
https://www.genomenon.com/indexed-variant-file
```

- SatMutMPRA
  - **STATUS: DONE**

```
# INFO-link: https://kircherlab.bihealth.org/satMutMPRA/
# Download link
#GRCh37/38
    # manual Web GUI download was done
    # "DOWNLOAD ALL ELEMENTS"
# File processing:
    #GRCh38
        (grep ^Chr GRCh38_ALL.tsv; grep -v ^Chr GRCh38_ALL.tsv | sort -
k1,1 -k2,2n ) | bgzip > satMutMPRA_GRCh38_ALL.gz
        tabix -s 1 -b 2 -e 2 -c C satMutMPRA_GRCh38_ALL.gz
    #GRCh37
        (grep ^Chr GRCh37_ALL.tsv; grep -v ^Chr GRCh37_ALL.tsv | sort -
k1,1 -k2,2n ) | bgzip > satMutMPRA_GRCh37_ALL.gz
        tabix -s 1 -b 2 -e 2 -c C satMutMPRA_GRCh37_ALL.gz
```

7. Regulatory impact

- Enformer
  - **STATUS: DONE**

```
# Download
https://ftp.ensembl.org/pub/current_variation/Enformer/
```

8. Splicing predictions

- SpliceAI (precalculated scores)
  - **STATUS: DONE (*Just used already downloaded*)**

```
# Stand alone installation of SpliceAI:
https://pypi.org/project/spliceai/
# Need to sign-up for user agreement for pre-calculated scores.
#Weblink: https://basespace.illumina.com/s/otSPW8hnhaZR
```

  - SpliceAI (realtime calculation using container)
    - **STATUS: DONE (*Just used already downloaded*)**

```
# Container Weblink:
https://hub.docker.com/r/cmgantwerpen/spliceai_v1.3
    #docker
        docker pull cmgantwerpen/spliceai_v1.3
    #singularity
        singularity pull docker://cmgantwerpen/spliceai_v1.3
```

9. Structural variant data

- StructuralVariantOverlap

```
# Need to download V4 SV sites for GRCh38
# Weblink: https://gnomad.broadinstitute.org/data#v4-structural-
variants

#Download link:
    https://storage.googleapis.com/gcp-public-data--
gnomad/release/4.1/genome_sv/gnomad.v4.1.sv.sites.vcf.gz
    https://storage.googleapis.com/gcp-public-data--
gnomad/release/4.1/genome_sv/gnomad.v4.1.sv.sites.vcf.gz.tbi
```

- **CNV_annotation (CUSTOM) [NEED TESTING]**

```
#Weblink:
    https://gnomad.broadinstitute.org/data#v4-copy-number-variants

#Download links:
    https://storage.googleapis.com/gcp-public-data--
gnomad/release/4.1/exome_cnv/gnomad.v4.1.cnv.all.vcf.gz
```

10. Transcript annotation

- NMD
  - **STATUS: DONE**

```
# NO additional data is requried but only VEP cache
# VEP command
./vep -i variations.vcf --plugin NMD
```

- RiboseqORFs
  - **STATUS: DONE**

```
# Web link:
    https://doi.org/10.1038/s41587-022-01369-0
# Download Link:
    https://ftp.ebi.ac.uk/pub/databases/gencode/riboseq_orfs/data/
# Data processing
    bgzip Ribo-seq_ORFs.bed
    tabix Ribo-seq_ORFs.bed.gz
```

- UTRAnnotator
  - **STATUS: DONE**

```
# Download link:
https://github.com/Ensembl/UTRannotator/blob/master/uORF_5UTR_GRCh38_PU
BLIC.txt
```

11. Variant data

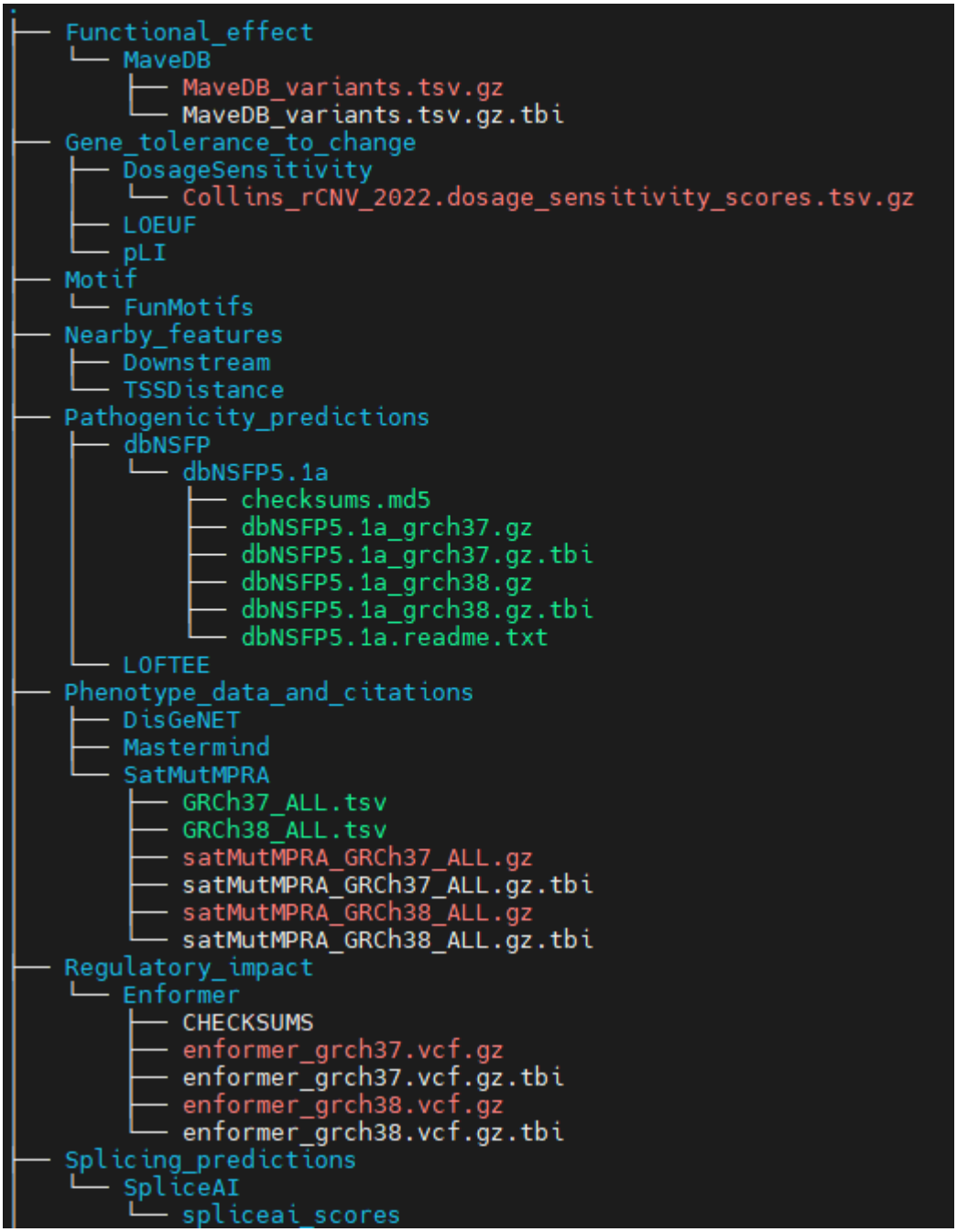
- LOVD
  - **STATUS: DONE**

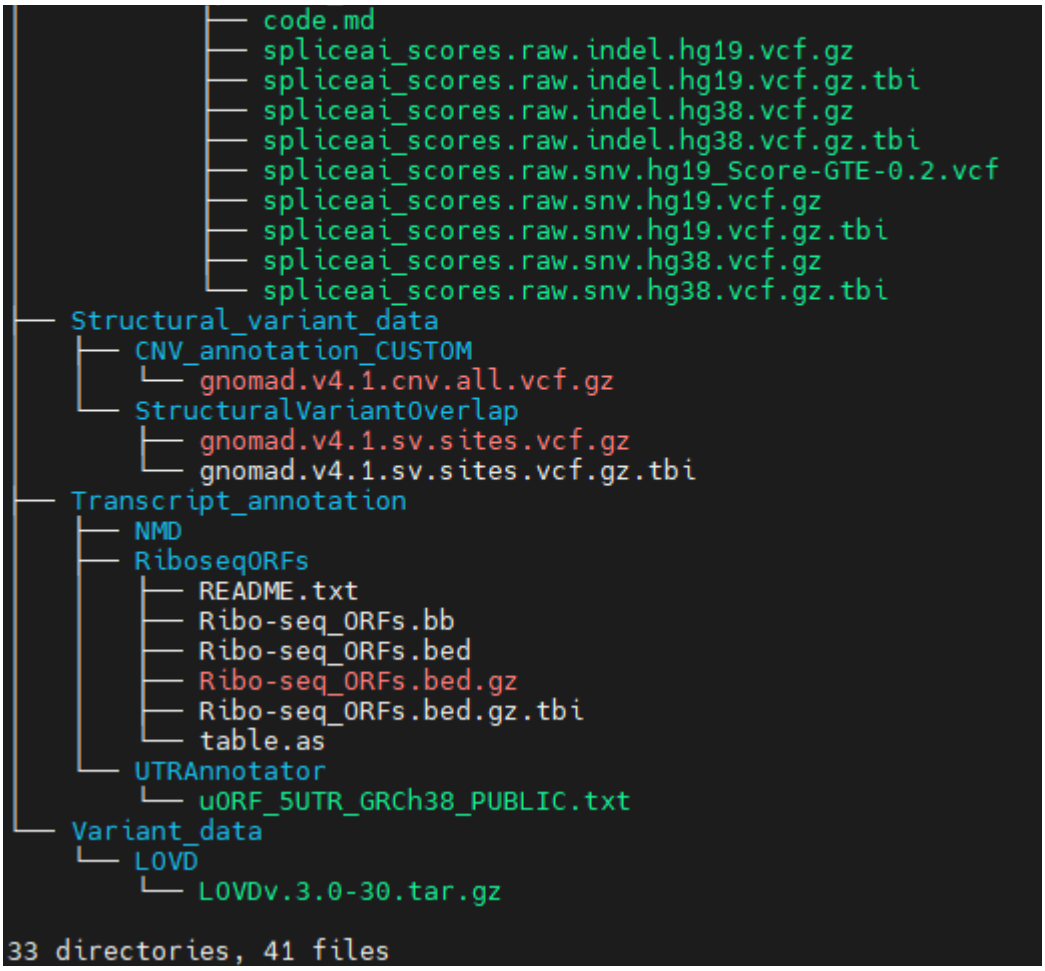
```
# NO additional data is requried but only VEP cache
# VEP command
./vep -i variations.vcf --plugin LOVD
```

File structuring

```
mkdir -p Functional_effect/MaveDB
mkdir -p Gene_tolerance_to_change/DosageSensitivity
```

```
mkdir -p Gene_tolerance_to_change/LOEUF
mkdir -p Gene_tolerance_to_change/pLI
mkdir -p Motif/FunMotifs
mkdir -p Nearby_features/Downstream
mkdir -p Nearby_features/TSSDistance
mkdir -p Pathogenicity_predictions/dbNSFP
mkdir -p Pathogenicity_predictions/LOFTEE
mkdir -p Phenotype_data_and_citations/DisGeNET
mkdir -p Phenotype_data_and_citations/Mastermind
mkdir -p Phenotype_data_and_citations/SatMutMPRA
mkdir -p Regulatory_impact/Enformer
mkdir -p Splicing_predictions/SpliceAI
mkdir -p Structural_variant_data/StructuralVariantOverlap
mkdir -p Transcript_annotation/NMD
mkdir -p Transcript_annotation/RiboseqORFs
mkdir -p Transcript_annotation/UTRAnnotator
mkdir -p Variant_data/LOVD
```




├── code.md
├── spliceai_scores.raw.indel.hg19.vcf.gz
├── spliceai_scores.raw.indel.hg19.vcf.gz.tbi
├── spliceai_scores.raw.indel.hg38.vcf.gz
├── spliceai_scores.raw.indel.hg38.vcf.gz.tbi
├── spliceai_scores.raw.snv.hg19_Score-GTE-0.2.vcf
├── spliceai_scores.raw.snv.hg19.vcf.gz
├── spliceai_scores.raw.snv.hg19.vcf.gz.tbi
├── spliceai_scores.raw.snv.hg38.vcf.gz
└── spliceai_scores.raw.snv.hg38.vcf.gz.tbi
├── Structural_variant_data
├── CNV_annotation_CUSTOM
└── gnomad.v4.1.cnv.all.vcf.gz
└── StructuralVariantOverlap
├── gnomad.v4.1.sv.sites.vcf.gz
└── gnomad.v4.1.sv.sites.vcf.gz.tbi
├── Transcript_annotation
├── NMD
├── RiboseqORFs
├── README.txt
├── Ribo-seq_ORFs.bb
├── Ribo-seq_ORFs.bed
├── Ribo-seq_ORFs.bed.gz
├── Ribo-seq_ORFs.bed.gz.tbi
└── table.as
└── UTRAnnotator
└── uORF_5UTR_GRCh38_PUBLIC.txt
└── Variant_data
└── LOVD
└── LOVDv.3.0-30.tar.gz
33 directories, 41 files