Sep 21, 2020

VEPextended

DOI
dx.doi.org/10.17504/protocols.io.bkhvkt66

Israel Aguilar Ordoñez[1]

[1]Instituto Nacional de Medicina Genómica (INMEGEN)

Whole genome variation...

Judith Ballesteros Villascan
Centro de Investigación y de Estudios Avanzados del IPN (Cin...

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.bkhvkt66

External link: https://github.com/laguilaror/nf-VEPextended

Protocol Citation: Israel Aguilar Ordoñez 2020. VEPextended. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.bkhvkt66

Manuscript citation:
Aguilar-Ordoñez I, Pérez-Villatoro F, García-Ortiz H, Barajas-Olmos F, Ballesteros-Villascán J, González-Buenfil R, Fresno C, Garcíarrubio A, Fernández-López JC, Tovar H, Hernández-Lemus E, Orozco L, Soberón X, Morett E (2021) Whole genome variation in 27 Mexican indigenous populations, demographic and biomedical insights. PLoS ONE 16(4): e0249773. doi: 10.1371/journal.pone.0249773

Protocol status: Working
We use this protocol and it's working

Created: August 30, 2020

Last Modified: September 21, 2020

Protocol Integer ID: 41237

Abstract

'VEPextended' is a tool, implemented in Nextflow, that annotates called variants using Variant Effect Predictor (VEP) and additional plugins that implement functionalities, that are not included in variation API.

All steps described are mk modules of code that will be done automatically through Nextflow pipeline.

Guidelines

Instalation

Download VEPextended from Github repository:

```
git clone https://github.com/Iaguilaror/nf-VEPextended.git
```

Compatible OS*:

- Ubuntu 18.04.03 LTS

* VEPextended may run in other UNIX based OS and versions, but testing is required.

Software Requirements:

Software	
bcftools	NAME

Software	
htslib	NAME

Software	
ensembl-vep	NAME

Software

Nextflow

NAME

Software

Plan9

NAME

https://github.com/9fans/plan9port

SOURCE LINK

Materials

Pipeline Inputs

- A compressed vcf file with extension '.vcf.gz', which must have a TABIX index with .tbi extension, located in the same directory as the vcf file.

Note(s): INFO must cointain AC
Example line(s):

```
##fileformat=VCFv4.2
##FILTER=
##GATKCommandLine.ApplyRecalibration.2=##INFO=
##INFO=
##contig=
##contig=
##contig=
##bcftools_viewCommand=view --compression-level 0 --output-type z --min-ac 1 --threads
1 ./test/data/sample.vcf.gz;
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
chr22   30000353        .       G       T       .       PASS
AC=107;AF_mx=0.708;AN=150;DP=883;nhomalt_mx=39
```

- *_dbSNP.vcf.gz: A reference dbSNP b152 file, that contains rsID of variants.

Dataset	
dbSNPb152_GRCH38_for_GATK.vcf.gz	NAME
https://ftp.ncbi.nlm.nih.gov/snp/	LINK

- Pre-scored files for SNVs and InDels compressed and indexed provided from CADD. Availables in

Dataset	
*_InDels.tsv.gz, *_SNVs.tsv.gz	NAME
https://cadd.gs.washington.edu/download	LINK

- Liftedover of gnomAD release 2.1.1 to GRCh38

Dataset	
completegenome_gnomAD.vcf.bgz	NAME
https://gnomad.broadinstitute.org/downloads	LINK

- Coverages from gnomAD v 2.1.1

Dataset	
gnomad.genomes.coverage.summary.bed.gz	NAME
https://storage.googleapis.com/gnomad-public/release/2.1/coverage/genomes/gnomad.genomes.coverage.summary.tsv.bgz	LINK

- GWAS association for SNVs (no haplotypes) compiled by iaguilar from GWAScat database at Spring 2019

Dataset	
All_20180418_noINFO.GWAScatalog.vcf.gz	NAME
https://www.ebi.ac.uk/gwas/docs/file-downloads	LINK

- Coordinates for every pre-miRNA, mature miRNA and seed region from miRBase v22.

Dataset	
miRBase.bed.gz	NAME
ftp://mirbase.org/pub/mirbase/CURRENT/	LINK

- _coverages.bed.gz: contains covergage of your sample.
- genotype - drug associations from PGKB

Dataset

NAME

var_drug_ann.tsv

LINK

https://www.pharmgkb.org/downloads

Before start

Test

To test VEPextended's execution using test data, run:

```
./runtest.sh
```

Your console should print the Nextflow log for the run, once every process has been submitted, the following message will appear:

```
======
VEP annotator: Basic pipeline TEST SUCCESSFUL
======
```

VEPextended results for test data should be in the following file:

```
test/results/_pos1_rejoin_chromosomes/sample.filtered.untangled_multiallelics.anno_dbSNP_vep.vcf.gz
```

Usage

To run VEPextended go to the pipeline directory and execute:

```
nextflow run vep-annotator.nf --vcffile  [--output_dir path to results ] [-resume]
```

For information about options and parameters, run:

```
nextflow run vep-annotator --help
```

Pre-processing

1 **Filter VCF**

Remove the variants that did not have any copy of the alternative allele.

Note

a) Filter and remove when there was not an alternative allele in the VCF file.vcf.gz, to only conserve found variants.
b) Compress the filtered file using one thread for compression.
c) Make and index output file using the filtered and compressed file.

Dependencies:

Software	
bcftools	NAME

2 **Extract chromosomes**

Extract variants per chromosome vcf file that have at least one variant.

Note

a) Using the index of the compressed vcf file, list the chromosomes names.
b) If there is at least one variant per chromosome, separate variants per chromosome.

Dependencies:

Software	
bcftools	NAME

Software	
htslib	NAME

3 **Split chromosomes**

Split in chunks a vcf file, keeping its format.

Note

a) Save the header of a vcf in a temporary file.
b) Save the body of a vcf in a temporary file.
c) Make chunks of the body of the vcf file.
d) Add the header to each body chunk.

Dependencies:

Software	
bcftools	NAME

Core-processing

4 **Untangle multiallelic**

Split multiallelic sites.

Note

a) Separate multiallelic sites and conserve vcf format.
b) Do not print bcftools version.

Dependencies:

Software

bcftools NAME

5 **Annotate rsID**

Annotate rsID to each variant in ID column of a VCF.

Note

a) Make a Reference file with a define range.
b) Compress input file.
c) Annotate rsID in the compressed input file using a Reference.

Dependencies:

Software

bcftools NAME

Software

htslib NAME

6 **Vep Extended**

*Annotate variants with Variant Effector Predictor tool (VEP). For more information about, see**VEP***

Dependencies:

Software	
ensembl-vep	NAME

Pos-processing

7 **Rejoin chromosomes**

Concatenate annotated chunks in a single vcf file.

Dependencies:

Software	
bcftools	NAME

Software	
htslib	NAME

Final Output:

Expected result
VCF file with only variants of each chromosome from the input. Example line(s):

```
##fileformat=VCFv4.2 #CHROM     POS     ID      REF     ALT     AC
AN      DP      AF_mx   nhomalt_mx      Allele  Consequence
IMPACT  SYMBOL  Gene    Feature_type    Feature BIOTYPE EXON
INTRON  HGVSc   HGVSp   cDNA_position   CDS_position
Protein_position        Amino_acids     Codons  Existing_variation
DISTANCE        STRAND  FLAGS   VARIANT_CLASS   SYMBOL_SOURCE
HGNC_ID CANONICAL       TSL     APPRIS  CCDS    ENSP    SWISSPROT
TREMBL  UNIPARC SOURCE  GENE_PHENO      SIFT    PolyPhen
DOMAINS HGVS_OFFSET     HGVSg   AF      AFR_AF  AMR_AF  EAS_AF
EUR_AF  SAS_AF  AA_AF   EA_AF   gnomAD_AF       gnomAD_AFR_AF
gnomAD_AMR_AF   gnomAD_ASJ_AF   gnomAD_EAS_AF   gnomAD_FIN_AF
gnomAD_NFE_AF   gnomAD_OTH_AF   gnomAD_SAS_AF   MAX_AF
MAX_AF_POPS     CLIN_SIG        SOMATIC PHENO   PUBMED  MOTIF_NAME
MOTIF_POS       HIGH_INF_POS    MOTIF_SCORE_CHANGE
GeneHancer_type_and_Genes       gnomADg gnomADg_AC      gnomADg_AN
gnomADg_AF      gnomADg_DP      gnomADg_AC_nfe_seu
gnomADg_AN_nfe_seu      gnomADg_AF_nfe_seu
gnomADg_nhomalt_nfe_seu gnomADg_AC_raw  gnomADg_AN_raw
gnomADg_AF_raw  gnomADg_nhomalt_raw     gnomADg_AC_afr
gnomADg_AN_afr  gnomADg_AF_afr  gnomADg_nhomalt_afr
gnomADg_AC_nfe_onf      gnomADg_AN_nfe_onf      gnomADg_AF_nfe_onf
gnomADg_nhomalt_nfe_onf gnomADg_AC_amr  gnomADg_AN_amr
gnomADg_AF_amr  gnomADg_nhomalt_amr     gnomADg_AC_eas
gnomADg_AN_eas  gnomADg_AF_eas  gnomADg_nhomalt_eas
gnomADg_nhomalt gnomADg_AC_nfe_nwe      gnomADg_AN_nfe_nwe
gnomADg_AF_nfe_nwe      gnomADg_nhomalt_nfe_nwe gnomADg_AC_nfe_est
gnomADg_AN_nfe_est      gnomADg_AF_nfe_est
gnomADg_nhomalt_nfe_est gnomADg_AC_nfe  gnomADg_AN_nfe
gnomADg_AF_nfe  gnomADg_nhomalt_nfe     gnomADg_AC_fin
gnomADg_AN_fin  gnomADg_AF_fin  gnomADg_nhomalt_fin
gnomADg_AC_asj  gnomADg_AN_asj  gnomADg_AF_asj
gnomADg_nhomalt_asj     gnomADg_AC_oth  gnomADg_AN_oth
gnomADg_AF_oth  gnomADg_nhomalt_oth     gnomADg_popmax
gnomADg_AC_popmax       gnomADg_AN_popmax       gnomADg_AF_popmax
gnomADg_nhomalt_popmax  gnomADg_cov     gwascatalog
gwascatalog_GWASC_trait gwascatalog_GWASC_pvalue
gwascatalog_GWASC_study clinvar clinvar_CLNDN   clinvar_CLNSIG
clinvar_GENEINFO        clinvar_CLNDISDB        miRBase
pharmgkb_drug   pharmgkb_drug_PGKB_annid
pharmgkb_drug_PGKB_gene pharmgkb_drug_PGKB_chem
pharmgkb_drug_PGKB_phencat chr22        16132524        .       G
C       27      156     1962    0.173   3       C
intergenic_variant      MODIFIER        .       .       .       .
.       .       .       .       .       .       .       .       .
```

```
.       .       .       .       .       SNV     .       .       .
.       .       .       .       .       .       .       .       .
.       .       .       .       chr22:g.16132524G>C     .       .
.       .       .       .       .       .       .       .       .
.       .       .       .       .       .       .       .       .
.       .       .       .       .       .       .       .       .
.       .       .       .       .       .       .       .       .
.       .       .       .       .       .       .       .       .
.       .       .       .       .       .       .       .       .
.       .       .       .       .       .       .       .       .
.       .       .       .       .       .       .       .       .
.       .       .       .       .       .       .       .       .
.       . chr22 27840687        .       G       T       69
156     2311    0.442   18      T
```