



Sep 21, 2020

# VEPextended

Israel Aguilar Ordoñez<sup>1</sup>

<sup>1</sup>Instituto Nacional de Medicina Genómica (INMEGEN)

1 Works for me dx.d

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

Whole genome variation in 27 Mexican indigenous populations, demographic and biomedical insights



Judith Ballesteros Villascan

Centro de Investigación y de Estudios Avanzados del IPN (Cin...

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.

**EXTERNAL LINK** 

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

DOI

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

PROTOCOL CITATION

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

**EXTERNAL LINK** 

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

LICENSE

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

CREATED

Aug 30, 2020

LAST MODIFIED

Sep 21, 2020

PROTOCOL INTEGER ID

41237

GUIDELINES

#### Instalation

Download VEPextended from Github repository:

git clone <a href="https://github.com/Iaguilaror/nf-VEPextended.git">https://github.com/Iaguilaror/nf-VEPextended.git</a>

#### Compatible OS\*:

mprotocols.io

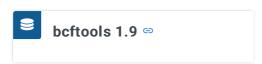
- Ubuntu 18.04.03 LTS
- \* VEPextended may run in other UNIX based OS and versions, but testing is required.

Citation: Israel Aguilar Ordoñez (09/21/2020). VEPextended. https://dx.doi.org/10.17504/protocols.io.bkhvkt66

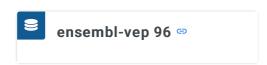
09/21/2020

A

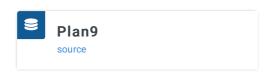
#### **Software Requirements:**











#### MATERIALS TEXT

#### **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=
##INF0=
##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
                                      Τ
                                                      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.

protocols.io
2
09/21/2020

	dbSNPb152_GRCH38_for_GATK.vcf.gz
■ Pre-s	scored files for SNVs and InDels compressed and indexed provided from CADD. Availables in
	*_InDels.tsv.gz, *_SNVs.tsv.gz

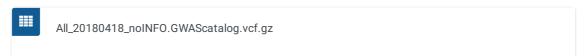
• Liftedover of gnomAD release 2.1.1 to GRCh38



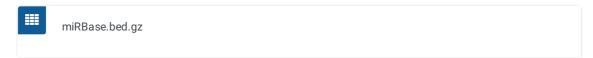
Coverages from gnomAD v 2.1.1



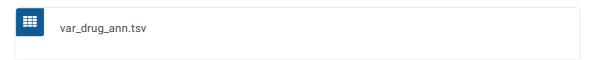
• GWAS association for SNVs (no haplotypes) compiled by iaguilar from GWAScat database at Spring 2019



 $\qquad \hbox{\bf Coordinates for every pre-miRNA, mature miRNA and seed region from miRBase v22}. \\$ 



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



BEFORE STARTING

### 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\_dbSN
P\_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

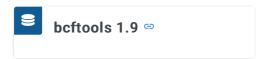
#### 1 Filter VCF

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



- 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:



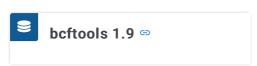
#### 2 Extract chromosomes

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



- 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:





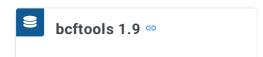
#### 3 Split chromosomes

Split in chunks a vcf file, keeping its format.



- 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:



#### Core-processing

### 4 Untangle multiallelic

Split multiallelic sites.



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

#### Dependencies:



### 5 Annotate rsID

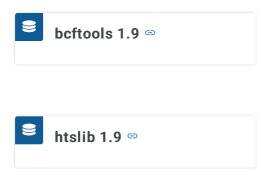
Annotate rsID to each variant in ID column of a VCF.



- 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:

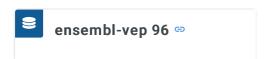
**⋈** protocols.io 5 09/21/2020



#### 6 Vep Extended

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

## Dependencies:

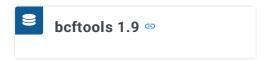


#### Pos-processing

#### 7 Rejoin chromosomes

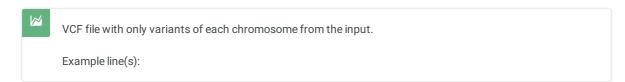
Concatenate annotated chunks in a single vcf file.

#### Dependencies:





#### **Final Output:**



##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 gnomADq 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