



Sep 21, 2020

nf-vcf-cataloguer

Israel Aguilar Ordoñez¹

¹Instituto Nacional de Medicina Genómica (INMEGEN)

1 Works for me

dx.doi.org/10.17504/protocols.io.bkmzku76

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

'nf-vcf-cataloguer' is a tool, implemented in Nextflow, that generates a general table description in TSV format of the description of each category and subgroup of a VCF with the extended annotation made by VEP. Furthermore, it plots each subset of the consequences of variants.

EXTERNAL LINK

https://github.com/laguilaror/nf-VCF-cataloguer

DOI

dx.doi.org/10.17504/protocols.io.bkmzku76

PROTOCOL CITATION

Israel Aguilar Ordoñez 2020. nf-vcf-cataloguer. **protocols.io** https://dx.doi.org/10.17504/protocols.io.bkmzku76

EXTERNAL LINK

https://github.com/laguilaror/nf-VCF-cataloguer

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

Sep 01, 2020

LAST MODIFIED

Sep 21, 2020

PROTOCOL INTEGER ID

41369

GUIDELINES

Instalation

Download nf-vcf-cataloguer from Github repository:

git clone https://github.com/Iaguilaror/nf-vcf-cataloguer.git

Compatible OS*:

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

Software Requirements:

protocols.io
1
09/21/2020

 $\textbf{Citation:} \ \ \text{Israel Aguilar Ordo} \\ \tilde{\text{A}} \\ \text{\pmez (09/21/2020)}. \ \ \text{nf-vcf-cataloguer.} \\ \\ \underline{\text{https://dx.doi.org/10.17504/protocols.io.bkmzku76}} \\ \text{\pmex (09/21/2020)}. \ \ \text{of-vcf-cataloguer.} \\ \underline{\text{https://dx.doi.org/10.17504/protocols.io.bkmzku76}} \\ \text{\pmex (09/21/2020)}. \\ \text{\pmex (09/21/2020)}.$



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.

The header names the eight mandatory columns: CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO. INFO must contain "AN", which is the target for filtering of this module.

For more information about the VCF format, please go to the next link: $\underline{\text{Variant Call Format}}$

Example line(s):

 $\textbf{Citation:} \ \ \text{Israel Aguilar Ordo} \\ \tilde{\mathbb{A}} \\ \text{\pm ez (09/21/2020)}. \ \ \text{nf-vcf-cataloguer.} \\ \\ \underline{\text{https://dx.doi.org/10.17504/protocols.io.bkmzku76}} \\$

- A '.txt' selection signals file chich lists rsIDs.
- A reference file to extract certain fields of vcf and transform it to tsv format.



fields_to extract.txt

BEFORE STARTING

Test

To test nf-vcf-cataloguer'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:

=====

nf-vcf-cataloguer: Basic pipeline TEST SUCCESSFUL

nf-vcf-cataloguer results for test data should be in the following file:

nf-vcf-cataloguer/test/results/catgorizeVCF-results

Usage

To run nf-vcf-cataloguer go to the pipeline directory and execute:

nextflow run categorize-vcf.nf --vcf <path to input 1> [--output_dir path to results]
[-resume]

For information about options and parameters, run:

nextflow run categorize-vcf.nf --help

Pre-processing

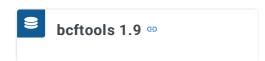
1 Custom filter

Remove the variants that have the AN (total number of alleles in called genotypes) value assigned.



a) Includes sites where the compressed VCF file '.vcf.gz' comply with the AN value.

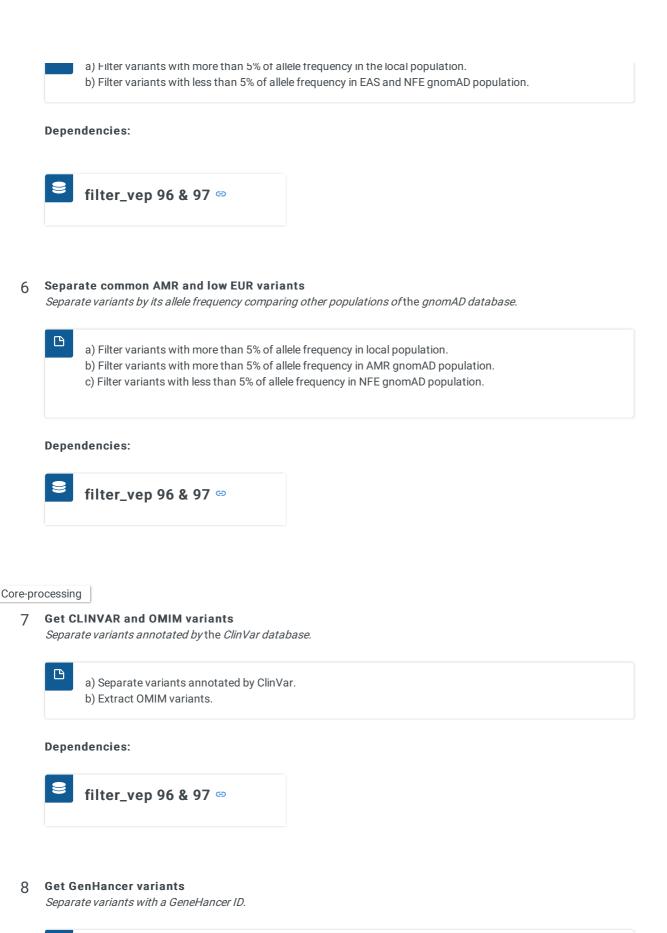
Dependencies:



 $\textbf{Citation:} \ \ \text{Israel Aguilar Ordo} \\ \tilde{\text{A}} \\ \text{\pmez (09/21/2020)}. \ \ \text{nf-vcf-cataloguer.} \\ \\ \underline{\text{https://dx.doi.org/10.17504/protocols.io.bkmzku76}} \\ \text{\pmex (09/21/2020)}. \ \ \text{of-vcf-cataloguer.} \\ \underline{\text{https://dx.doi.org/10.17504/protocols.io.bkmzku76}} \\ \text{\pmex (09/21/2020)}. \\ \text{\pmex (09/21/2020)}.$

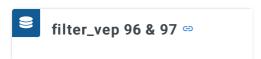
2	Separate SNVs and indels Keep only certain types of variants.
	a) Includes SNPs of a compressed VCF file '.vcf.gz'. b) Includes indels of a compressed VCF file '.vcf.gz'.
	Dependencies:
	bcftools 1.9 ©
3	Separate rare, low and common frequencies Keep only certain types of variants, set by its allele frequency.
	a) Separate variants by its allele frequency category. b) Separate variants in common frequency. c) Separate variants in low frequency. d) Separate variants in rare frequency.
	Dependencies:
	bcftools 1.9 ©
4	Separate selection signals Keep only variants with selection signals.
	a) Separate variants on selection, with a reference ID list of selection signals. b) Sort output file.
	Dependencies:
	bcftools 1.9 🖘
5	Separate low EAS and low EUR variants

B



a) Filter variants that match with annotations in the "GeneHancer type and genes" field.

Dependencies:



9 Get GWASCatalog variants

Separate variants with a GeneHancer ID.

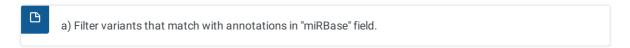


Dependencies:



1∩ Get miRNAs variants

Separate variants with miRNA data.



Dependencies:



11 Get novel and known variants

Separate known and unknown variants.



Dependencies:



12 Get coding variants

 Separate variants in coding regions.



- a) Separate exonic variants.
- b) Filter intronic variants.

Dependencies:



filter_vep 96 & 97 👄

13 Get PGKB variants

Separate variants found in PGKB database.



a) Filter variants that match with annotations in "PGKB" field.

Dependencies:



filter_vep 96 & 97 🖘

14 Get UTR variants

Separate variants found in 5' or 3' UTR regions.



- a) Filter variants that are in 5' UTR.
- b) Filter variants that are in 3' UTR.

Dependencies:



filter_vep 96 & 97 =

Pos-processing

15 VCF to TSV

Convert vcf files to tsv format.



- a) Search ANN header and separates it by tabs.
- b) Separate columns by tabs.
- c) Add a "." to blank spaces.

Dependencies:



bcftools 1.9 🖘

Final Output:



A '.tsv.gz' file with columns of the VEP annotations, by each vcf converted.

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 CADD PHRED CADD RAW 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_GWAScat_DISEASE_or_TRAIT gwascatalog_GWAScat_INITIAL_SAMPLE SIZE gwascatalog GWAScat REPLICATION SAMPLE SIZE gwascatalog GWAScat STRONGEST SNP and RISK ALLELE gwascatalog GWAScat PVALUE gwascatalog_GWAScat_STUDY_ACCESSION clinvar clinvar_CLNDN clinvar_CLNSIG clinvar_CLNDISDB miRBase pharmgkb_drug pharmgkb_drug_PGKB_Annotation_ID pharmgkb_drug_PGKB_Gene pharmgkb drug PGKB Chemical pharmgkb drug PGKB PMID pharmgkb drug PGKB Phenotype Category pharmgkb drug PGKB Sentence chr21 33241945 rs2229207 T C 27 152 2003 0.179 2 C missense variant MODERATE IFNAR2 ENSG00000159110 Transcript ENST00000342136.8 protein coding 2/9 . ENST00000342136.8:c.23T>C ENSP00000343957.4:p.Phe8Ser 349/2899 23/1548 8/515 F/S tTc/tCc rs2229207&CM066573 . 1 . SNV HGNC HGNC:5433 YES 1 P4 CCDS13621.1 ENSP00000343957 P48551 . UPI000012D69B . 1 tolerated benign hmmpanther:PTHR20859&hmmpanther:PTHR20859:SF53&Transmembrane helices:TMhelix . chr21:g.33241945T>C 0.1186 0.0809 0.147 0.1706 0.0736 0.1421 0.07558 0.07791 0.1033 0.07742 0.154 0.07741 0.1757 0.08546 0.08082 0.09088 0.1247 0.1757 gnomAD EAS risk factor . 1&1 16757563&19434718&23009887&28497593&18588853&27186094 2.171 0.043682 . rs2229207 3026 31374 0.0964493 657463 14 106 0.132075 0 3039 31416 0.0967341 168 736 8706 0.0845394 27 198 2136 0.0926966 10 122 848 0.143868 13 317 1552 0.204253 29 164 637 8592 0.0741387 31 607 4584 0.132417 35 1456 15418 0.0944351 76 277 3474 0.0797352 12 23 290 0.0793103 1 95 susceptibility to risk factor OMIM:610424

16 Count variants

Citation: Israel Aguilar Ordoñez (09/21/2020). nf-vcf-cataloguer. https://dx.doi.org/10.17504/protocols.io.bkmzku76



Summary_cleaner.R is a tool for counting variants from different types and subgroups.

Dependencies:

summary_cleaner.R

Final Output:



A '.tsv' the description of the counted variants by each subgroup of variants.

Example line(s):

row_nam common_freq commonAMR_lowEUR low_freq lowEAS_lowEUR No_filter rare_freq selection_signals miRNA 0 0 0 0 0 0 0 PGKB 0 0 0 0 0 0 0 GWAScatalog 3 0 0 0 3 0 0 0MIM 6 2 1 0 9 2 0 coding_region 7 0 1 1 19 11 0 clinvar 16 2 2 0 21 3 0 utr 90 5 26 7 161 45 0 dbSNPnovel 35 2 114 3 977 828 0 GeneHancer 599 44 187 47 1033 247 0 dbSNPknown 8995 746 2706 567 14586 2885 0 general (all indels) 9030 748 2820 570 15563 3713 0

17 QC VEP consequence plot

Plot consequences of each category of variants.

Dependencies:

plotter.R