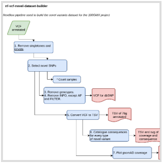

Sep 21, 2020

nf-vcf-novel-dataset-builder

Israel Aguilar Ordoñez[1]

[1]Instituto Nacional de Medicina Genómica (INMEGEN)

1 *Works for me* dx.doi.org/10.17504/protocols.io.bkh7kt9n

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

Nextflow pipeline used to build the novel variants dataset for the 100GMX project.

'nf-vcf-novel-dataset-builder' is a pipeline tool that builds a VCF file compiling only novel variants according to dbSNP and VEP, from a VEPextended annotated VCF file. This novel selection does not include singletons and private variants. The main output is in VCF format. Additional outputs include the dataset in TSV format, and a sequence coverage from gnomAD in these sites.

Important note: input file must be previously annotated byhttps://github.com/Iaguilaror/nf-VEPextended

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

EXTERNAL LINK

https://github.com/Iaguilaror/nf-vcf-novel-dataset-builder.git

DOI

dx.doi.org/10.17504/protocols.io.bkh7kt9n

PROTOCOL CITATION

Israel Aguilar Ordoñez 2020. nf-vcf-novel-dataset-builder. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.bkh7kt9n

EXTERNAL LINK

https://github.com/Iaguilaror/nf-vcf-novel-dataset-builder.git

LICENSE

CREATED

Aug 30, 2020

LAST MODIFIED

Sep 21, 2020

PROTOCOL INTEGER ID

41247

GUIDELINES

Instalation

Download nf-vcf-novel-dataset-builder from Github repository:

```
git clone https://github.com/Iaguilaror/nf-vcf-novel-dataset-builder.git
```

Compatible OS*:

- Ubuntu 18.04.03 LTS

* nf-vcf-novel-dataset-builder may run in other UNIX based OS and versions, but testing is required.

Software Requirements:

bcftools 1.9

htslib 1.9

Nextflow 19.04

Plan9
source

R 3.4.4

MATERIALS TEXT

Pipeline Inputs

- A compressed vcf file with extension '.vcf.gz'; the VCF must be previously annotated withhttps://github.com/Iaguilaror/nf-VEPextended

Example line(s):

```
##fileformat=VCFv4.2 #CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO chr21
5101724 . G A . PASS
AC=1;AF=0.00641;AN=152;DP=903;ANN=A|intron_variant|MODIFIER|GATD3B|ENSG00000280071|Tran
script|ENST00000624810.3|protein_coding||4/5|ENST00000624810.3:c.357+19987C>T|||||||||-
1|cds_start_NF&cds_end_NF|SNV|HGNC|HGNC:53816||5|||ENSP00000485439||A0A096LP73|UPI0004F
23660|||||||chr21:g.5101724G>A||||||||||||||||||||||||||||2.079|0.034663|||||||||||||||
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| chr21 5102165
rs1373489291 G T . PASS
AC=1;AF=0.00641;AN=140;DP=853;ANN=T|intron_variant|MODIFIER|GATD3B|ENSG00000280071|Tran
script|ENST00000624810.3|protein_coding||4/5|ENST00000624810.3:c.357+19546C>A|||||||rs1
373489291||-
```

Citation: Israel Aguilar OrdoÃÂ±ez (09/21/2020). nf-vcf-novel-dataset-builder. https://dx.doi.org/10.17504/protocols.io.bkh7kt9n

```
1|cds_start_NF&cds_end_NF|SNV|HGNC|HGNC:53816||5|||ENSP00000485439||A0A096LP73|UPI0004F
23660|||||||chr21:g.5102165G>T||||||||||||||||||||||||||||5.009|0.275409|||||||||||||||
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

BEFORE STARTING

Test

To test nf-vcf-novel-dataset-builder 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-novel-dataset-builder: Basic pipeline TEST SUCCESSFUL
======
```

nf-vcf-novel-dataset-builder results for test data should be in the following file:

```
nf-vcf-novel-dataset-builder/test/results/VCFnovelbuilder-results
```

Usage

To run nf-vcf-novel-dataset-builder go to the pipeline directory and execute:

```
nextflow run vcf-novel-finder.nf --vcffile <path to input 1> [--output_dir path to
results ]o results ] [-resume]
```

For information about options and parameters, run:

```
nextflow run vcf-novel-finder.nf --help
```

Pre-processing

1 Remove singletons and private

Remove singletons and private variants.

a) Filter positions where AC >= '3' to eliminate singletons and private.
* AC could be modified.

Dependencies:

bcftools 1.9

Core-processing

2 Select novel

Select novel SNPs, indels variants and concatenate both type variants.

a) Filter novel SNPs.
b) Filter novel indels.
c) Concatenate novel SNPs and indels.
d) Sort VCF.

Dependencies:

bcftools 1.9

Pos-processing

3 Count per sample

List and count samples to present block data in a column format.

- final-counter.R is a tool for transforming wide to long format.

a) List all samples.
b) Extract block of counted data only.
c) Transform to column format.

Dependencies:

bcftools 1.9

- final-counter.R

4 Simplify VCF for dbSNP upload

Remove genotypes, remove FORMAT and all INFO field except INFO/AF and FORMAT, also changes AF to AF_natmx.

a) Remove genotypes.
b) Remove fields except for INFO/AF and FILTER.
c) Rename local AF to AF_natmx, also in the header.

Dependencies:

bcftools 1.9

Final Output:

A compressed and simplified VCF file format.

Citation: Israel Aguilar OrdoÃÂ±ez (09/21/2020). nf-vcf-novel-dataset-builder. https://dx.doi.org/10.17504/protocols.io.bkh7kt9n

Example line(s):

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO chr21   5227536 .       C
CTCTCCTCTCT     .       .       AF_natmx=0.019 ...
```

5 **VCF2TSV**

Convert vcf to tsv format.

Dependencies:

bcftools 1.9

Final Output:

A compressed TSV file format.

Example line(s):

```
CHROM   POS     REF     ALT     AF_natmx        Allele  Consequence     IMPACT  SYMBOL
Gene    Feature_type    Feature BIOTYPE EXON    INTRON  HGVSc   HGVSp   cDNA_position
CDS_position    Protein_posi chr21   5227536 C       CTCTCCTCTCT     0.019   TCTCCTCTCT
intergenic_variant      MODIFIER        .       .       .       .       .       .       .
.       .       .       .       .       . ...
```

6 **Consequence cataloguer**

Catalogue consequences for each type of variant.

- cataloguer.R is a tool for cataloging the consequences of novel variants.

Dependencies:

- cataloguer.R

Final Output:

A compressed TSV file format by each category of variant and a SVG file.

Example line(s) of TSV:

```
Consequence     number_of_variants      Type    General_category
First_specific_consequence 3_prime_UTR_variant     2       noncoding       UTR     3 prime
UTR 3_prime_UTR_variant&NMD_transcript_variant     NA      noncoding       UTR     3 prime
UTR ...
```

Citation: Israel Aguilar OrdoÃÂ±ez (09/21/2020). nf-vcf-novel-dataset-builder. https://dx.doi.org/10.17504/protocols.io.bkh7kt9n

7 **Coverage gnomAD**

Plot gnomAD coverages.

- coverage-analyzer.R is a tool for plotting coverage of the gnomAD project.

Dependencies:

- coverage-analyzer.R

Final Output:

.tif file format by each variant category.

Citation: Israel Aguilar OrdoÃÂ±ez (09/21/2020). nf-vcf-novel-dataset-builder. https://dx.doi.org/10.17504/protocols.io.bkh7kt9n