Sep 21, 2020

nf-100GMX-variant-summarizer

Israel Aguilar Ordoñez[1]
[1]Instituto Nacional de Medicina Genómica (INMEGEN)

1 Works for me dx.doi.org/10.17504/protocols.io.bkv6kw9e

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 count variants for the 100GMX project
'nf-100GMX-variant-summarizer' is a pipeline tool that counts variants in a VEPextended annotated VCF file.
This pipeline generates 3 outputs:
1) a TSV file with the total number of SNV and indels
2) a TSV file with per sample counts for variants of type SNV, indel, novel, worldwide singletons, clinvar, gwascat and pharmgkb
3) a PDF file with the number of discernible variants in sample groups of interest.

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

EXTERNAL LINK

https://github.com/Iaguilaror/nf-100GMX-variant-summarizer

DOI

dx.doi.org/10.17504/protocols.io.bkv6kw9e

PROTOCOL CITATION

Israel Aguilar Ordoñez 2020. nf-100GMX-variant-summarizer. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.bkv6kw9e

EXTERNAL LINK

https://github.com/Iaguilaror/nf-100GMX-variant-summarizer

CREATED

Sep 05, 2020

LAST MODIFIED

Sep 21, 2020

PROTOCOL INTEGER ID

41630

GUIDELINES

Installation

Download nf-100GMX-variant-summarizer from Github repository:

```
git clone https://github.com/Iaguilaror/nf-100GMX-variant-summarizer
```

Compatible OS*:

- Ubuntu 18.04.03 LTS
- Ubuntu 16.04 LTS

* nf-100GMX-variant-summarizer may run in other UNIX based OS and versions, but testing is required.

Software Requirements:

bcftools 1.9

htslib 1.9

filter_vep 96 & 97

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||||||||||||||
```


Citation: Israel Aguilar OrdoÃÂ±ez (09/21/2020). nf-100GMX-variant-summarizer. https://dx.doi.org/10.17504/protocols.io.bkv6kw9e

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

- *.tsv: A metadata file, relating every sample ID (as registered in the VCF file) and a sample group in column format.

Example line(s):

```
sample          group
SM-3MG5L Chinanteco
SM-3MG5F Chocholteco
SM-3MG46 Kanjobal
```

BEFORE STARTING

Test

To test nf-100GMX-variant-summarizer'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:

```
======
VCF summarizer: Basic pipeline TEST SUCCESSFUL
======
```

nf-100GMX-variant-summarizer results for test data should be in the following file:

```
nf-100GMX-variant-summarizer/test/results/VCFsummarizer-results
```

Usage

To run nf-100GMX-variant-summarizer go to the pipeline directory and execute:

```
nextflow run summarize-vcf.nf --vcffile <path to input 1> --metadata <path to input 2>
--nsamples <integer> --group_minaf <numeric> --outgroup_maxaf <numeric> [--output_dir
path to results ]
```

For information about options and parameters, run:

```
nextflow run summarize-vcf.nf --help
```

Branch A

1 **Project Counts**

a) Count samples and raw stats for all samples.
b) Give the total counts data.

Branch B

2 **No filter counts**

Citation: Israel Aguilar OrdoÃÂ±ez (09/21/2020). nf-100GMX-variant-summarizer. https://dx.doi.org/10.17504/protocols.io.bkv6kw9e

a) Filter variants that are not in ClinVar, GWAS Catalog or PGKB.
b) Give the total counts data.

Dependencies:

- final-counter.R

bcftools 1.9

3 **Novel counts**

a) Filter variants without a rsID.
b) Give the total counts data.

Dependencies:

- final-counter.R

bcftools 1.9

4 **Worldwide singletons counts**

a) Filter variants that are singletons.
b) Keep variants that have not frequencies in another population.

Dependencies:

- final-counter.R

bcftools 1.9

filter_vep 96 & 97

5 **ClinVar counts**

a) Filter variants harbored in CinVar.
b) Give the total counts data.

Citation: Israel Aguilar OrdoÃÂ±ez (09/21/2020). nf-100GMX-variant-summarizer. https://dx.doi.org/10.17504/protocols.io.bkv6kw9e

Dependencies:

- final-counter.R

filter_vep 96 & 97

bcftools 1.9

6 **GWASCatalog counts**

a) Filter variants harbored in GWAS Catalog.
b) Give the total counts data.

Dependencies:

- final-counter.R

bcftools 1.9

filter_vep 96 & 97

7 **PGKB counts**

a) Filter variants harbored in Pharm GKB.
b) Give the total counts data.

Dependencies:

- final-counter.R

bcftools 1.9

Citation: Israel Aguilar OrdoÃÂ±ez (09/21/2020). nf-100GMX-variant-summarizer. https://dx.doi.org/10.17504/protocols.io.bkv6kw9e

filter_vep 96 & 97

8 **Merge tables**

Merge tables of counted data.

Dependencies:

- merger.R

Branc C

9 **Define groups**

a) Count samples per group.
b) Join in a file samples per group.

Dependencies:

NONE

10 **Select world rare**

a) Filter variants according AF in gnomAD populations and NatMex.

Dependencies:

filter_vep 96 & 97

11 **Extract discernible VCF**

a) Extract variants in other samples.
b) Extract variants in local samples.
c) Extract exclusive variants.

Dependencies:

bcftools 1.9

12 **Count and Plot**

Citation: Israel Aguilar OrdoÃÂ±ez (09/21/2020). nf-100GMX-variant-summarizer. https://dx.doi.org/10.17504/protocols.io.bkv6kw9e

a) Count the number of discernible variants per group and type in a TSV file.
b) Plot number of discernible variants per group and type.

Dependencies:

bcftools 1.9

- plotter.R

Citation: Israel Aguilar OrdoÃÂ±ez (09/21/2020). nf-100GMX-variant-summarizer. https://dx.doi.org/10.17504/protocols.io.bkv6kw9e