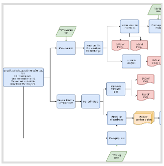


Sep 21, 2020

VCF2PCP

Judith Ballesteros Villascan[1], Israel Aguilar Ordoñez[2], Fernando Pérez-Villatoro[2]

[1]Centro de Investigación y de Estudios Avanzados del IPN (Cinvestav); [2]Instituto Nacional de Medicina Genómica

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

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 that runs and plots admixture and smartpca from a compressed VCF.

EXTERNAL LINK

https://github.com/jbv2/VCF2PCP

DOI

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

PROTOCOL CITATION

Judith Ballesteros Villascan, Israel Aguilar Ordoñez, Fernando Pérez-Villatoro 2020. VCF2PCP. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.bkwbkxan

EXTERNAL LINK

https://github.com/jbv2/VCF2PCP

LICENSE

CREATED

Sep 05, 2020

LAST MODIFIED

Sep 21, 2020

PROTOCOL INTEGER ID

41635

GUIDELINES

Installation

Download VCF2PCP from Github repository:

```
git clone https://github.com/jbv2/VCF2PCP.git
```

Compatible OS*:

- Ubuntu 18.04.03 LTS

* VCF2PCP may run in other UNIX based OS and versions, but testing is required.

Software Requirements:

bcftools 1.9

plink 2

Eigensoft 6.1.4

Admixture 1.3

Nextflow 19.04

Plan9
source

R 3.4.4

MATERIALS TEXT

Pipeline Inputs

- A compressed VCF file with extension '.vcf.gz'.

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

- A file that contains the name of samples and the group that belongs to, separated by " ".(samples.txt)

Example line(s):

```
sample1 Zoque
sample2 PEL
sample3 PEL
sample4 CHB
...
```

- A file that contains fields: sample, pop, and region separated by tabs. (tag_data.tsv). It helps for regions like north, central, and south.

Example line(s):

```
sample  pop region
sample2 PEL PEL
sample3 PEL PEL
sample4 CHB CHB
...
```

BEFORE STARTING

Test

To test VCF2PCP 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:

```
======
vcf2pcp: Basic pipeline TEST SUCCESSFUL
======
```

VCF2PCP results for test data should be in the following file:

```
VCF2PCP/test/results/VCF2PCP-results
```

Usage

To run VCF2PCP go to the pipeline directory and execute:

```
nextflow run vcf2pcp.nf --vcffile <path to input 1> [--output_dir path to results ]
```

For information about options and parameters, run:

```
nextflow run vcf2pcp.nf --help
```

Before Nextflow

1 Format and select samples

Removes unused contigs in the *header and keeps given samples.*

Dependencies:

bcftools 1.9

Pre-processing

2 **Split chromosomes**

Split chromosomes from a compressed VCF file.

Dependencies:

bcftools 1.9

3 **Simplify and remove LD**

SimplifyVCCF to keep only INFO/AF and GT and removes LD variants with bcftools +prune. Please, consider window for LD pruning *is given in bp.*

a) Remove variants in LD.
b) Simplify VCF to keep only INFO/AF and GT

Dependencies:

bcftools 1.9

4 **Rejoin VCF**

Concatenate multiple VCF of different chromosomes.

Dependencies:

bcftools 1.9

5 **VCF to PLINK**

Convert VCF to plink and filters MAF.

a) Convert VCF to PLINK file.
Filter MAF with PLINK.

Citation: Judith Ballesteros Villascan, Israel Aguilar OrdoÃÂ±ez, Fernando PÃÂ©rez-Villatoro (09/21/2020). VCF2PCP.
https://dx.doi.org/10.17504/protocols.io.bkwbkxan

Dependencies:

plink 2

6 **Make pedind**

Make pedind file for running smartpca by using tagger.R

- tagger.R is a tool that takes columns of fam file and the groups of samples and makes pedind file.

Dependencies:

- tagger.R

7 **Make pop info**

Make popinfo file for plotting admixture results by using make_popinfo.R

- make_popinfo.R is a tool that takes columns of fam file and the groups of samples and makes popinfo file.

Dependencies:

- make_popinfo.R

Core-processing

8 **Make par file for smartpca**

Make par file to run smartpca, runs it and take best snps a-nd Tracy-Widom statistics from stdout.

a) Write par file
b) Run smartpca
c) Get best snps

Dependencies:

Eigensoft 6.1.4

9 **Keep autosomes**

Keep only autosomal chromosomes for running admixture, as it is said in its documentation.

Dependencies:

plink 2

10 **Run admixture**

Run admixture with K 2:9 by default and gathers all logs.

Dependencies:

Admixture 1.3

Pos-processing

11 **Parallel coordinate plot**

Get number of snps for PCA, and the number of statistically significant PCs and plots it by using parallel_plotter.R

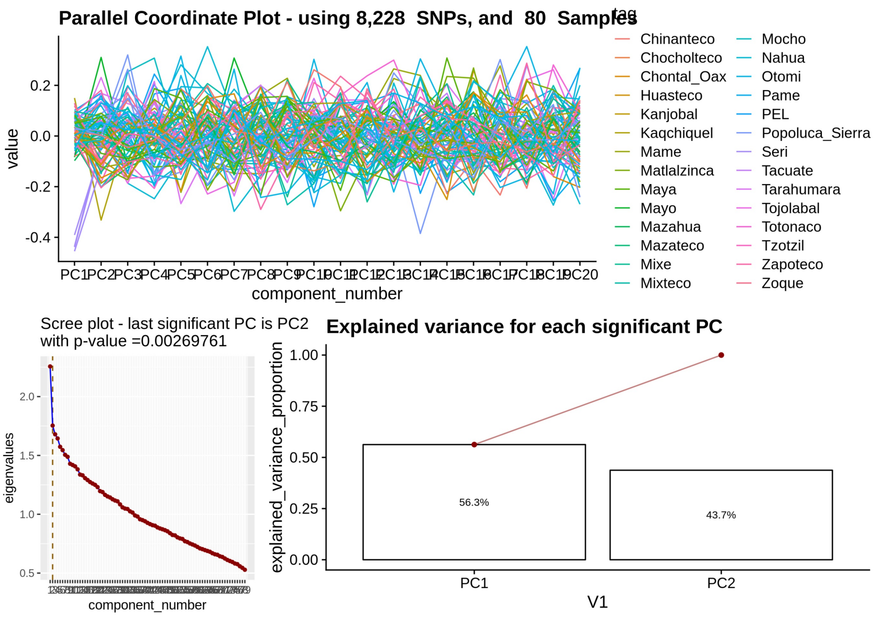
- *parallel_plotter.R is a tool for making parallel coordinates plots.*

a) Get the number of snps for PCA, and number of statisttically significnt PCs.
b) Reformat the evec file to replace spaces.
c) Run Rscript

Dependencies:

- parallel_plotter.R

Final Output:



12 **Regional PCA**

Plot PCA of PC1 vs all PCs and makes PCP by region by using plotter.R

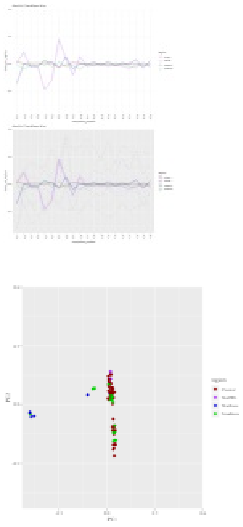
- plotter.R is a tool for making parallel coordinates plot by region.

Dependencies:

- plotter.R

Final Output:

Citation: Judith Ballesteros Villascan, Israel Aguilar OrdoÃÂ±ez, Fernando PÃÂ©rez-Villatoro (09/21/2020). VCF2PCP. https://dx.doi.org/10.17504/protocols.io.bkwbkxan



13 Plot Admixture

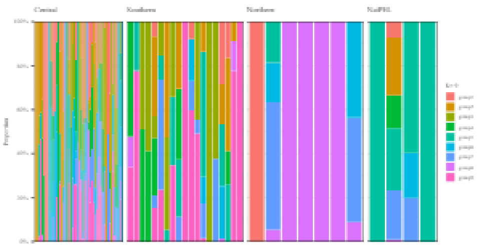
Plot all admixture results by using admixture_plotter.R

- admixture_plotter.R is a tool for plotting each admixture result.

Dependencies:

- admixture_plotter.R

Final Output:



14 Plot CVS

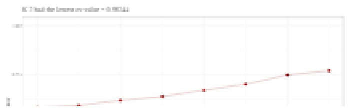
Plot CVS from admixture by using plotter.R

- plotter.R is a tool for plotting each CV from admixture results.

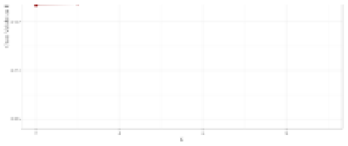
Dependencies:

- plotter.R

Final Output:



Citation: Judith Ballesteros Villascan, Israel Aguilar OrdoÃÂ±ez, Fernando PÃÂ©rez-Villatoro (09/21/2020). VCF2PCP. https://dx.doi.org/10.17504/protocols.io.bkwbkxan



15 **Gather admixture plots**

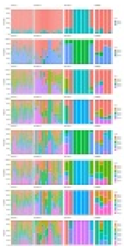
Plot all admixture results in one file by using plotter.R

- plotter.R is a tool for plotting k 2:9 from admixture results.

Dependencies:

- plotter.R

Final Output:



16 **Kmeans**

Get k means from significant PCs using kmean.R

- kmean.R is a tool for making groups (k) fro significant PCs.

a) Get the number of snps for PCA, and the number of statistically significant PCs

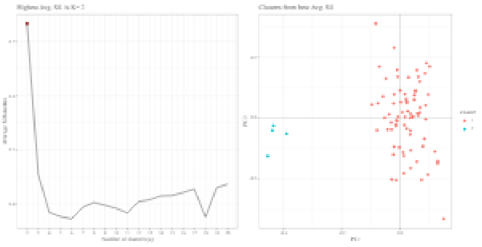
b) Reformat the evec file to replace spaces

c) Run Rscript

Dependencies:

kmean.R

Final Output:



Citation: Judith Ballesteros Villascan, Israel Aguilar OrdoÃÂ±ez, Fernando PÃÂ©rez-Villatoro (09/21/2020). VCF2PCP. https://dx.doi.org/10.17504/protocols.io.bkwbkxan

Citation: Judith Ballesteros Villascan, Israel Aguilar OrdoÃÂ±ez, Fernando PÃÂ©rez-Villatoro (09/21/2020). VCF2PCP. https://dx.doi.org/10.17504/protocols.io.bkwbkxan