README.md

Stacks_documentation

By Enora Geslain on 25/02/21

Reference: Rochette & Catchen (2017), Deriving genotypes from RAD-seq short-read data using Stacks (refered as "the protocol" below)

You can find the script and data examples at this link:
https://github.com/Enorya/LBEG_documentation/tree/main/Stacks

De novo analysis

Preparing the working directory and the data

1. Do the steps 1 and 2 of the protocol. In addition to the requested directories, create one named `scripts` and another called `populations`. At the end you should have the same tree structure as the one of the `Example/` folder.

You can remove the folders `genome`, `alignments`, `stacks.ref` and `tests.ref` from the tree structure, they only concern the reference-based analysis (except if you want to do both type of analysis).

Your raw data should look like the ones in the `Example/raw/` folder if they are not demultiplexed, otherwise you should have 2 files per sample (one for forward reads R1 and one for reverse reads R2).

2. Do the step 3 of the protocol. Depending on the technique you used in order to sequenced your samples you can have indexes in addition to your barcodes. If it is the case you should have barcodes files looking like this (without the first line):

```
Barcode Index   Sample_name
GCATG   ATCACG  T_eul_PS82_313
AACCA   ATCACG  T_eul_PS82_314
CGATC   ATCACG  T_eul_PS82_315
TCGAT   ATCACG  T_eul_PS82_317
TGCAT   ATCACG  T_sco_PS96_229
CAACC   ATCACG  T_loe_PS96_213
```

You can see a complete example in the `Example/info/` repertory

3. Do the step 4 of the protocol. You will find an example of a popmap file in `Example/info/`

Demultiplexing and filtering (trimming) the reads

4. Do the step 7 and 8 of the protocol. In order to run the *process_radtags* command on the VSC you can use a script looking like `Example/scripts/process_radtags_trem_lib1.pbs` and you can launch it by writing the following command in your working directory:

```
qsub ./scripts/process_radtags_trem_lib1.pbs
```

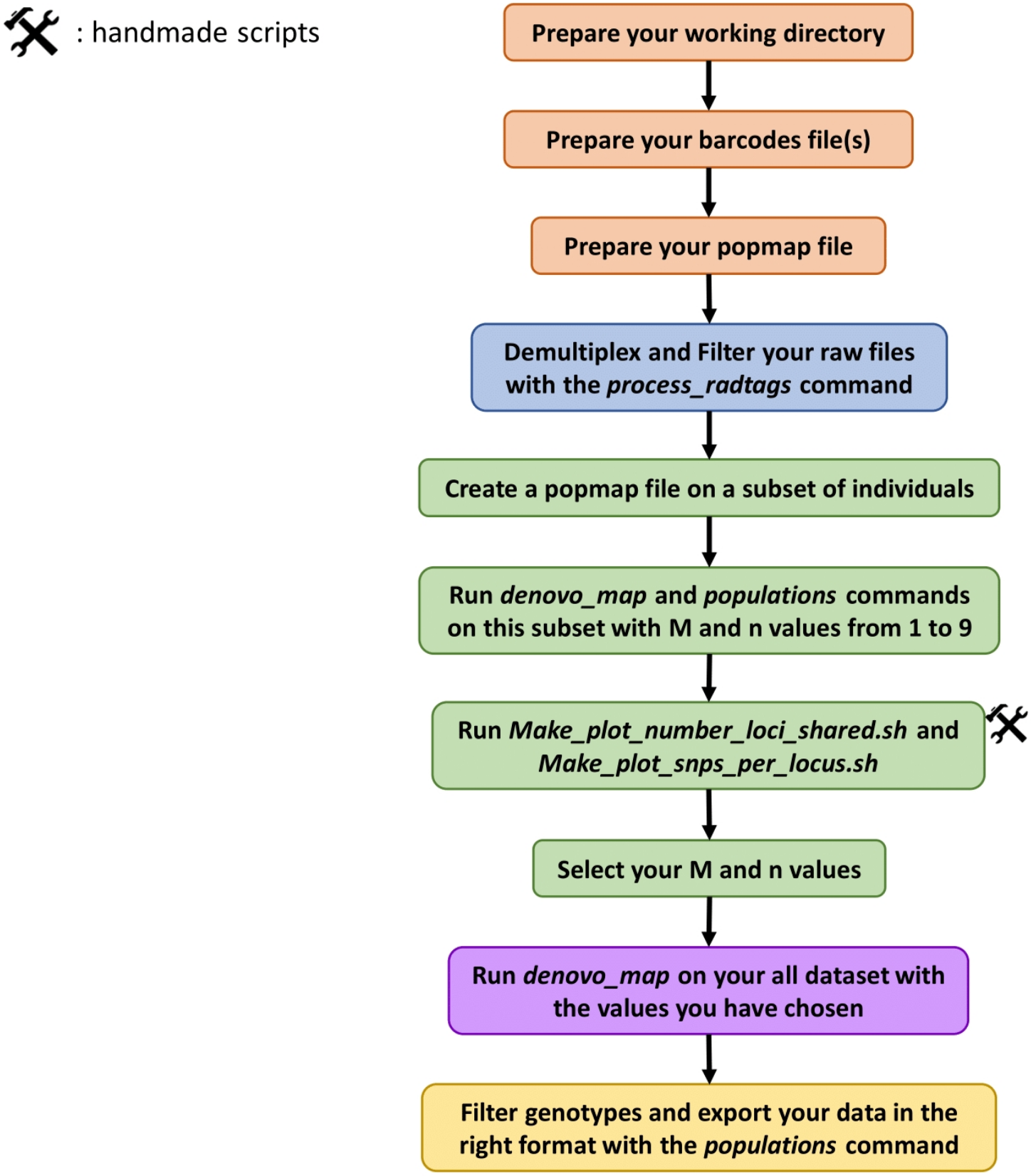
(you will need to run this command as many times as you have libraries)

/!\ **Depending of the number of samples in your libraries this step can be long and heavy thus you might need to change the number of nodes and cores requested as well as the walltime and the memory indicated in the lines 3, 4 and 5 of the script**

5. As suggested in the step 9 of the protocol you can check the proportion of retained reads (and the number of retained reads per sample) in the log file in order to see if some samples should be discarded (because of too low or too high retained reads). You will find it in the `cleaned/` directory along with the fastq files of your cleaned reads (= demultiplexed + filtered) like in the `Example/cleaned/` folder.

6. **Warning**: If you are working with paired-end sequencing data **don't do the step 11** of the protocol. It is no longer necessary because, since the writing of the protocol, an option has been added to specify that the data are paired.

Also, don't do the steps 12 and 13 of the protocol as you are working on *de novo* analysis.



Working on a subset of samples for parameter testing

7. Do the step 14 of the protocol. I suggest you choose a subset of individuals representative of your entire dataset (like individuals from different locations) and take the ones with a high number of retained reads. You can find an example here: `Example/info/popmap.test_samples.tsv`

8. Do the steps (i) to (iv) from the step 15 of the protocol (warning: **A** is for **de novo** analysis and **B** for **reference-based** analysis). You will find the scripts in order to run the denovo_map command and the populations command (for M=n=1 and 2) in `Example/scripts/` (You can launch them with the same command as the step 4 with `qsub` before the name of the script).

(steps (v) and (vi) are not mandatory)

9. Do the steps (vii) and (viii) from the step 15 of the protocol.

/!\ **Warning**: the explainations in the protocol are longer working.

You need to follow these steps instead:

- copy and paste the scripts present in the folder `/staging/leuven/stg_00026/Useful_scripts/Stacks` in your own `scripts/` folder
- open the script `Make_plot_number_loci_shared.sh` and follow the instructions at the beginning of the script (changing the path for the input data and the path to the script `plot_R_graphs_number_loci_shared.r`)
- open the script `Make_plot_snps_per_locus.sh` and do the same manipulation as above
- launch each `Make_plot_...` script with the following command:

```
bash script_name
```

- look at the 2 graphs generated and choose the better value for the parameters of your dataset

At the end you should have 4 files:

- 2 PDF files corresponding to the graphs
- 2 text files containing the tables used to generate the graphs

Running Stacks on the full dataset

10. Do the steps 16 and 17-A-(i) of the protocol.

11. For the rest of the steps (17-A-(ii) to (vii)) you can choose to follow the protocol or to run only one command for the integrality of the steps. Indeed, in the protocol the authors use the commands in a decomposed way: ustacks, cstacks and sstacks but it's the same as running the denovo_map command (it's just easier to parallelize the jobs). If you want to run directly the denovo_map command you can do a script as the one called `denovo_map_all_trem.pbs` in the `Example/scripts/` folder.

At the end you will have:

- 5 files called `catalog....`
- 2 files called `gstacks....`
- 1 file called `denovo_map.log`
- 1 file called `tsv2bam.log`
- 6 files called `populations....`
- 5 files for each sample

You can see some of the results files in the `Example/stacks.denovo/` folder

(You can do steps 18 to 20 of the protocol but they are not necessary)

Filtering genotypes and exporting the data

12. Do the step 21 of the protocol. You can add a lot of different options in order to filter the genotypes or to have different output formats. You will find all these options and their descriptions at this link: https://catchenlab.life.illinois.edu/stacks/comp/populations.php

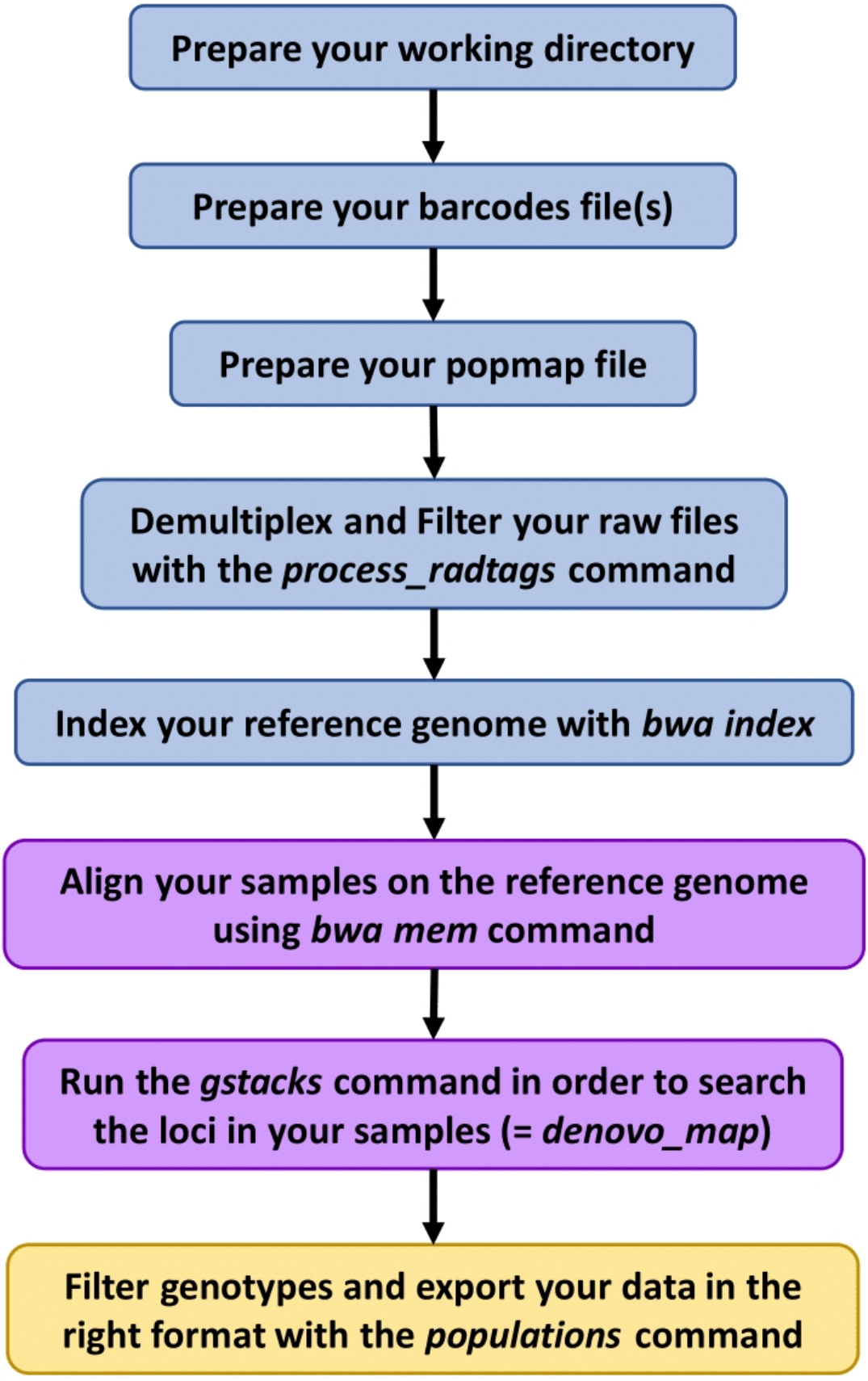
Some interesting options are:

- --min-mac

- --min-maf
- --write-single-snp

And, of course, choose carefully the file output options in order to have the files you need for your downstream analysis. You can find a script example in the `Example/scripts/` folder, it is named `populations_all_trem.pbs`

Reference-based analysis



Preparing, demultiplexing and filtering the data

1. Do the steps 1 to 9 of the protocol as explained above (points 1 to 5 of the *De novo* analysis part of this document).

You can remove the folders `stacks.denovo` and `tests.denovo` from the tree structure they only concern the *De novo* analysis.

2. Do the steps 12 and 13 of the protocol. In order to run the *bwa index* command on the VSC you can use a script looking like `Example/scripts/index_bwa_trem.pbs`

Be careful to the prefix you are choosing for the output database, you will need it after (`-p bwa/tre_ber` in the script)

At the end you should have 5 different files in the `Example/genome/bwa/` folder with the following suffixes: *.amb, .ann, .bwt, .pac, .sa*

You will find only the first 2 as example, the others were to heavy.

Working on a subset of samples for parameter testing

3. It was the step 15-B of the protocol but the *pstacks* command that is needed no longer works. In any case in the reference-based analysis this step is not used to choose parameters, it just allows to check the quality of the data further.

Running Stacks on the full dataset

4. Do the steps 16 and 17-B-(i). In order to align your samples on your reference genome with the *bwa mem* command on the VSC you can use a script looking like `Example/scripts/alignment_bwa_trem.pbs`

After the alignment you should have 1 bam file per sample. In the `Example/alignments/` folder the file is empty because it was to heavy.

5. Do the steps 17-B-(ii) and (iii) of the protocol.

6. As I explained above the *pstacks* command no longer exist. Hence, you cannot do the steps 17-B-(iv) to (vii) of the protocol. Instead you need to use the *gstacks* command like in the script `Example/scripts/stacks_bwa_trem.pbs` .

Filtering genotypes and exporting the data

7. Finally, as the *De novo* analysis, you can do the step 21 of the protocol (point 12 of the *De novo* analysis part of this document).