Processing ddRAD data from raw fastqs to vcf

Thom Nelson[1], Findley Finseth[2], Lila Fishman[3], Colette Sevier Berg[3]

[1]Embark Veterinary; [2]Claremont Colleges; [3]University of Montana

Jul 16, 2021

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

Colette Sevier Berg

ABSTRACT

This is a pipeline that the Fishman lab uses for processing paired-end ddRAD data, from fastq to vcf. These scripts were written and pipeline developed by Findley Finseth and Thom Nelson.

DOI

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

PROTOCOL CITATION

Thom Nelson, Findley Finseth, Lila Fishman, Colette Sevier Berg 2021. Processing ddRAD data from raw fastqs to vcf. **protocols.io** https://dx.doi.org/10.17504/protocols.io.bjnbkman

CREATED

Aug 12, 2020

LAST MODIFIED

Jul 16, 2021

PROTOCOL INTEGER ID

40355

Demultiplex plates using i7 barcode

1 This step requires 4 fastq files: forward reads, reverse reads, the i7 file, and the i5 file, and then a barcode file.

```
demultiplex_i7.py -f forwardFileName.fastq -r reverseFileName.fastq -i7
i7FileName.fastq -i5 i5FileName.fastq -b barcodeFile.txt
```

demultiplex_i7.py

Sample barcode file:
plate_name \t i7 barcode

The output files will be demultiplexed fastq files with the prefix from the first column of the barcode file. Each half-plate will have a forward and reverse read.

Remove PCR duplicates using molecular barcode

2 This step removes PCR duplicate using the i5 molecular barcode.

rmdup_molbarcodes.py

```
rmdup_molbarcodes.py -p {filePrefix} -s .fastq
```

The resulting files will have the same prefix, but will have the suffix .rmdup.1.fastq (forward reads) and .rmdup.2.fastq (reverse reads)

Flip the reads

3

flip2BeRAD.py

```
flip2BeRad.py -c {cutsite sequence, eg: TGCAG} -f {forward file}.rmdup.1.fastq -
r {reverse file}.rmdup.2.fastq -b {barcode file}.txt -m 1 -o 2
```

Use this command to flip the reads for each half-plate. The output file will be called filtered_forward.fastq and filtered_reverse.fastq, so now is a good time to make separate folders for each half-plate and put the resultng files in those folders.

The barcode file is a list of 48 RAD barcodes used.

Demultiplex by individual; join forward and reverse reads

4 This uses the process_shortreads bash command to demultiplex wach half-plate into one file for each individual. The input files required are filtered_forward.fastq, filtered_reverse.fastq, and a barcode file.

```
process_shortreads -1 {filtered_forward}.fastq -2 {filtered_reverse}.fastq -b
{barcode File}.txt -o ~/thatPlatesFolder -i fastq -y gzfastq -r -c -q -E phred33 --
inline_null
```

Barcode file format:
barcode \t sample_name

Double and triple check that the barcodes correspond to the correct sample name!

The outputs will be zipped fastqs for each individual, with the suffix .fq.gz

Make bams

5 This script uses the tools GATK and Trimmomatic, so those tools need to be installed beforehand.

The input files for this script are the sample fastq and the genome FASTA file. The genome file also needs to have the following index files: .dict, .fa.amb, .fa.bwt, .fa.fai, .fa.pac, .fa.sa with the same prefix as the genome file & stored in the same directory as the genome. **fastq2bam.sh**

```
fastq2bam.sh /file/path/prefix /path/to/genome.fa
```

This will output a .bam and a .bai for each individual. Now is a good time to to QC by visually examining the bams in IGV

Make genomic vcfs (gvcfs)

6 This step takes bams and makes a gvcf for each individual using the attached script, then combines them all into a vcf. Put all the bams into one directory first, and then edit the script so that there is a path to the directory with all the bams.

```
bash bams_to_vcf.sh
```

bams_to_vcf.sh