**SNP calling**

**## Quality control**

**The data is single end 100bp. To understand the structure, the adaptor content and the barcodes I subset the big \*.gz\* files to 250'000 reads.**

```

#!/bin/sh

# in source\_files

zcat SQ1645\_S1\_R1\_001.fastq.gz | head -n 1000000 > SQ1645\_sample.fq

```

```

#!/bin/sh

fastqc \*fq

```

**Then checked the quality of the sequencing run using fastqc, output in [metadata/laneSQ1645\_sample\_fastqc.html](metadata/laneSQ1645\_sample\_fastqc.html)**

**Removed the last bases in each lane as there is a lot of adaptor contamination. Balanced number of reads vs number of bases trimming all reads to a common 70bp. Shorter reads are removed.**

```

#!/bin/sh

module load cutadapt

cutadapt -a CTGCAAGATCGGAAGAGC -m 60 -o trimmed\_SQ1645\_sample.fq SQ1645\_sample.fq

fastqc trimmed\_SQ1645\_sample.fq

cutadapt -a GCAAGATCGGAAGAGCACACGTCTGAACTCCAGTCACGAACTGAGCGATC -m 60 -o trimmedALLSQ1645\_sample.fq trimmed\_SQ1645\_sample.fq

```

**Ran it on the whole data:**

```

#!/bin/sh

module load cutadapt

zcat SQ1645\_S1\_R1\_001.fastq.gz > SQ1645\_S1\_R1\_001.fastq

cutadapt -j 16 -a CTGCAAGATCGGAAGAGC -o temptrimmed\_SQ1645\_S1\_R1\_001.fastq SQ1645\_S1\_R1\_001.fastq

cutadapt -j 16 -a GCAAGATCGGAAGAGCACACGTCTGAACTCCAGTCACGAACTGAGCGATC -m 50 -o trimmed\_SQ1645\_S1\_R1\_001.fastq temptrimmed\_SQ1645\_S1\_R1\_001.fastq

rm SQ1645\_S1\_R1\_001.fastq temptrimmed\_SQ1645\_S1\_R1\_001.fastq

```

```

22,637,008,603 bp in

13,701,692,270 bp out

```

**### Demultiplexing**

**First extracted barcodes from the .key file of the sequencing platform. See section 4.1.2 of the Stacks manual (**[**https://catchenlab.life.illinois.edu/stacks/manual/index.php#clean**](https://catchenlab.life.illinois.edu/stacks/manual/index.php#clean)**).**

```

#!/bin/sh

##Key provided in folder

less -S SQ1645\_key\_correct.txt | grep -v GBSNEG | cut -f 3,24 | grep -v barcode > barcodes.txt # first l

# cp barcodes.txt ~/repos/scripts/stoneflies\_gracie/metadata/

```

```sh

#/home/ludovic.dutoit/projects/stoneflies\_gracie

#!/bin/sh

mkdir raw samples

cd raw

ln -s ../source\_files/Stonefly\_GBS\_Data\_August21/trimmed\_SQ1645\_S1\_R1\_001.fastq

cd ..

```

```sh

#!/bin/sh

module load Stacks/2.58-gimkl-2020a

process\_radtags -p raw/ -o ./samples/ -b barcodes.txt -e PstI -r -c -q --inline-null

```

```

178044058 total sequences

1528067 barcode not found drops (0.9%)

17991 low quality read drops (0.0%)

101427 RAD cutsite not found drops (0.1%)

176396573 retained reads (99.1%)

```

**### Alignment and variant calling**

**Indexed the genome:**

```bash

#!/bin/sh

mkdir alignment

cd alignment

bwa index stoneflygenomeassemblyv1.fasta

```

**First alignment for every sample using BWA.**

```bash

#!/bin/sh

bwa mem -t 8 stoneflygenomeassemblyv1.fasta $src/${sample}.fq.gz | samtools view -b | samtools sort --threads 4 > ${sample}.bam

```

The complete list of commands is in [align.sh](align.sh)

**Then ran refmap from Stacks to identify low quality individuals:**

```

#!/bin/sh

module load Stacks

mkdir output\_refmap

ref\_map.pl --samples alignment --popmap popmap.txt -T 8 -o output\_refmap

```

**Ran populations again to obtain a VCF:**

```

populations -P output\_refmap/ -M popmap.txt --vcf --structure --plink --treemix --max-obs-het 0.65 -r 0.75 -O output\_refmap

```

```

73241 variants remained

```

**Got some stats per inds:**

```

vcftools --vcf populations.snps.vcf --missing-indv

```

**Here are some low-quality individuals:**

```

STY06\_FW 26853 0.366639

STY02\_FW 26923 0.367595

STY09\_FW 27031 0.369069

AWA05\_FW 27542 0.376046

AWA04\_VW 27809 0.379692

AWA03\_VW 27844 0.38017

AWA08\_FW 28754 0.392594

SH09\_FW 30442 0.415642

```

**and the worst 4:**

```

MW04\_FW 44311 0.605003

STY05\_FW 63764 0.870605

SIX09\_VW 69980 0.955476

BOU02\_VW 72609 0.991371

```

**Only removed the worst 4 as the other ones still have plenty of data.**

```

cat popmap.txt | grep -v MW04\_FW | grep -v STY05\_FW | grep -v SIX09\_VW | grep -v BOU02\_VW > popmap\_nooutliers06missing.txt

```

**After some preliminary analyses, some individuals turned out to be hybrids or from another species. They are excluded here.**

**Also combined this dataset with some blackjacks and lug/six mile samples.**

Those files were aligned after trimming and demultiplexing as done in [www.github.com/ldutoit/stoneflies\_blackjacks](www.github.com/ldutoit/stoneflies\_blackjacks) and [www.github.com/ldutoit/stoneflies\_GBS1](www.github.com/ldutoit/stoneflies\_GBS1) and then aligned with [align\_blackjacks.sh](align\_blackjacks.sh) and [align\_lugsixmile.sh](align\_lugsixmile.sh)

**Then created the file** [metadata/combined\_popmap.txt](metadata/combined\_popmap.txt) and ran ref\_map.sh

```

ref\_map.pl --samples alignment --popmap popmap\_combined.txt -T 8 -o output\_refmap

populations -P output\_refmap/ -M popmap\_combined.txt --vcf --structure --plink --treemix --max-obs-het 0.65 -r 0.75 -O output\_refmap

```

**A bit of quality checking:**

```

cd output\_refmap

vcftools --vcf populations.snps.vcf --missing-indv

```

**We are left with one additional sample that has more than 50% missing data: `SH09\_FW`**

**After removing it in the file** **[metadata/popmap\_combined\_nooutliers.txt](metadata/popmap\_combined\_nooutliers.txt), we re-ran populations tolerating a maximum of 20% missing data at any given site:**

```

populations -P output\_refmap/ -M popmap\_combined.txt --vcf --structure --plink --treemix --max-obs-het 0.65 -r 0.80 -O output\_refmap

```

**\*\*47319 for 258 inds\*\***