wilding_paper

Josquin Daron

9 juillet 2021

1. Samples

- **AG1000g phase 2: 1142 genomes and 16 populations**

GNcol (4), GQgam (9), GHgam (12), FRgam (24), GNgam (40), KE (48), GHcol (55), GM (65), GAgam (69), Icol (71), BFcol (75), AOcol (78), GW (91), BFgam (92), UGgam (112), CMgam (297)

- **Wilding genomes: 96 genomes and 3 populations**

32 LVBdom (Libreville, Gabon domestic)
32 LPdom (La lope, Gabon domestic)
32 LPfor (La lope, Gabon forest)

- **Urbano genomes: 88 genomes and 3 populations**

10 BZV (Brazzaville, Congo)
36 DLA (Douala, Cameroon)
42 LBV (Libreville, Gabon)

---

2. Dataset creation: reads mapping, SNP calling and filtering

Reads mapping

- FASTQC report:
- Urbano data: fastqc report git hub link.
- cutadapt:

```
cutadapt -a AGATCGGAAGAGCACACGTCTGAA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT  -q
```

- bwa mem:

```
header=$(zcat $sampleId.R1.fastq.gz | head -n 1)
id=$(echo $header | head -n 1 | cut -f 1-4 -d":" | sed 's/@//' | sed 's/:/_/g')
sm=$(echo $header | head -n 1 | grep -Eo "[ATGCN]+$")
echo "Read Group @RG\tID:$id\tSM:$id"_"$sm\tLB:$id"_"$sm\tPL:ILLUMINA"

bwa mem -t 1 Anopheles_gambiae.AgamP4.dna.chr.fna $sampleId.R1.fastq.gz $sampleId.R2.fastq.gz -R $(echo
m\tPL:ILLUMINA") | samtools view -F 4 -b - | samtools sort - -o $sampleId.map.sort.bam
```

- gatk realigner:

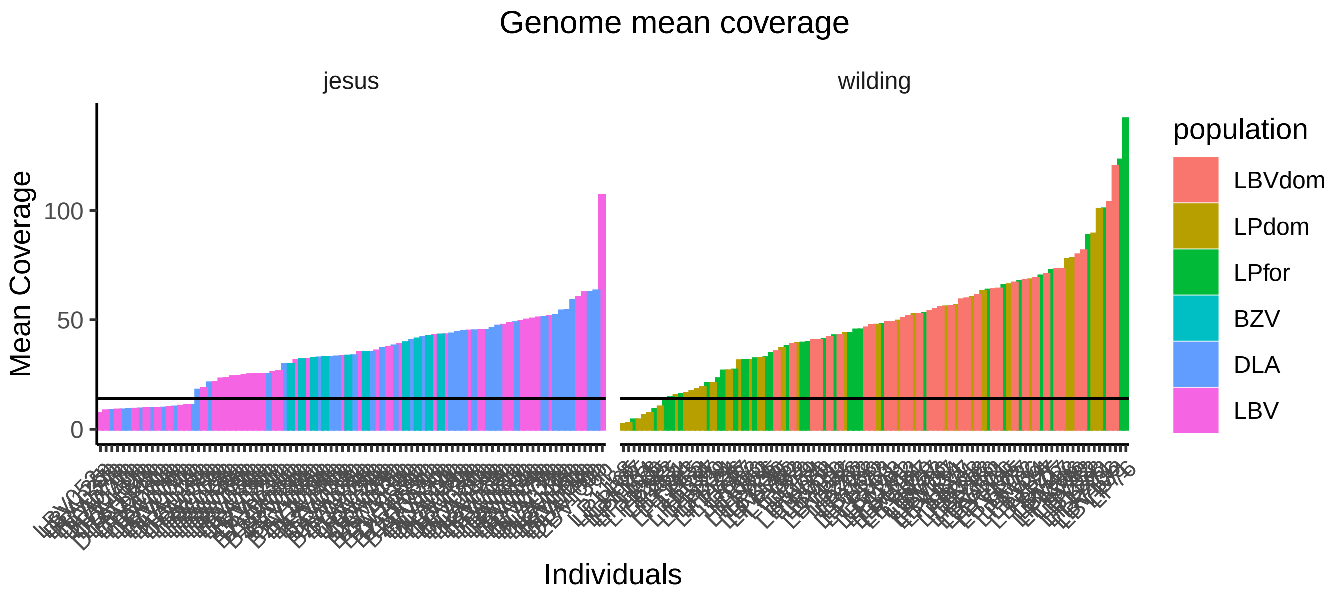
```
java -jar ~/bioInf/bin/GenomeAnalysisTK-3.8-0-ge9d806836/GenomeAnalysisTK.jar -T RealignerTargetCreator
 -I $sampleId.map.sort.bam -o $sampleId.realignertargetcreator.intervals

java -Xmx8G -Djava.io.tmpdir=/tmp -jar ~/bioInf/bin/GenomeAnalysisTK-3.8-0-ge9d806836/GenomeAnalysisTK.
ae.AgamP4.dna.chr.fna -targetIntervals $sampleId.realignertargetcreator.intervals -I $sampleId.map.sort
```

- bam file report (qualimap):

```
qualimap bamqc -bam /scratch/daron_anopheles/bam/$inputFile.indelrealigner.bam -c --java-mem-size=8G -ou
es/fastqBamInfo/bamInfo/jesus_qualimap/$inputFile.outqualimap -nt 2 -outformat HTML
```

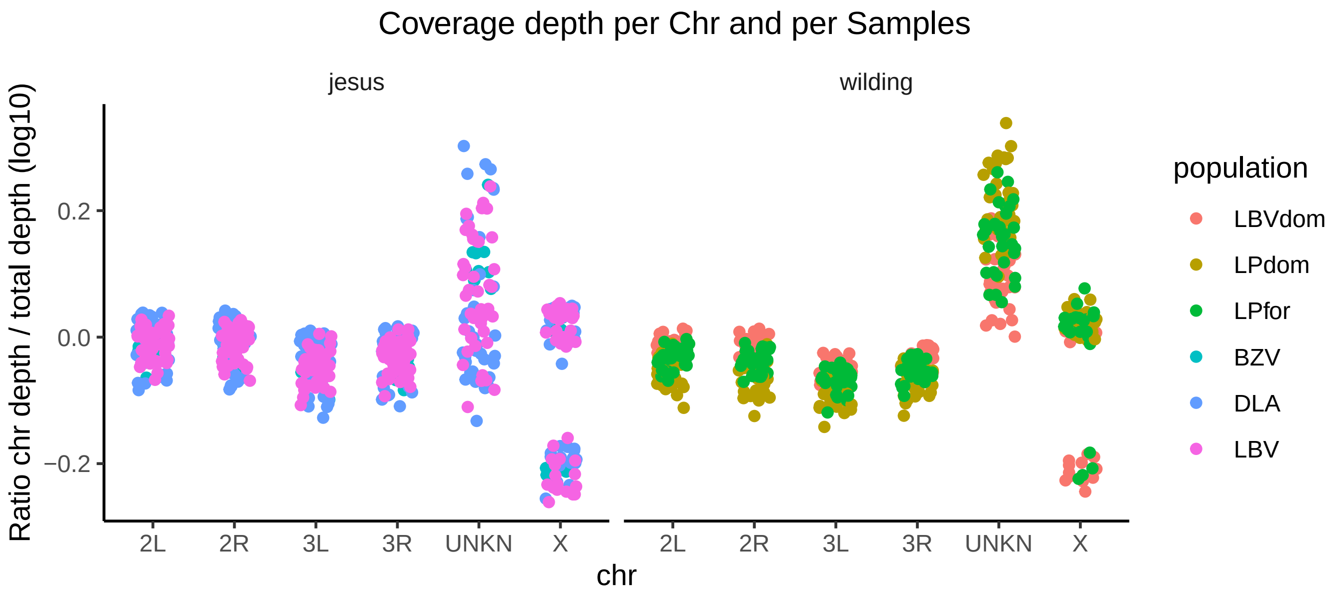
1. wilding report: qualimap report git hub link.
2. urbano report: qualimap report git hub link.
3. Genome mean coverage:



–> Filtering individuals: Remove individuals with mean coverage lower than 14x

- 9 individuals from Wilding: LP69, LP243, LP697, LP1118, LP1125, LP1164, LP1165, LP1168, LP1285
- 17 individuals from Urbano: DLA037p, DLA076p, DLA077p, DLA102p, DLA105p, DLA130p, DLA132p, DLA155Bp, LBV001p, LBV007p, LBV009p, LBV052p, LBV125p, LBV127p, LBV137p, LBV140p, LBV142p

4. Determining sex of each samples:



- wilding: 16/96 Males
- urbano: 36/88 Males

SNPs calling and filtering

- gatk unifiedGenotyper:

```
java -jar ~/bin/GenomeAnalysisTK-3.8-0-ge9d806836/GenomeAnalysisTK.jar -T UnifiedGenotyper -R Anopheles_
ist -L $interval --genotyping_mode DISCOVERY --downsampling_type BY_SAMPLE -dcov 250 --output_mode EMIT_
17 --genotype_likelihoods_model BOTH --heterozygosity 0.01 --indel_heterozygosity 0.001 -stand_call_con
rose.$out.unifiedGenotyper.vcf
```

- SNP filtering:

```
launch_ipynb.py -i vcfStats_slurm.ipynb -o wilding.chr3R.vcfStats_slurm.html
```

1. Wilding samples only SNPs stat report: git hub link.

- Final number of SNPs per chr 2L 1,228,916
  2R 1,605,477
  3L 1,159,765
  3R 1,610,164
  X 295,618
- 9 inds removed because imiss > 10%: LP1120 LP1134 LP1283 LP255 LP51 LP53 LP65 LP934 LP937

2. Wilding and Urbano samples SNPs stat report: git hub link.

- Final number of SNPs per chr 2L 1,228,916
  2R 1,605,477
  3L 1,159,765
  3R 1,610,164
  X 295,618
- 4 inds removed because imiss > 10%: LP1124 LP1145 LP47 LP63
- /! for X chr 10 samples are removed: BZV093bu DLA136u DLA137u LBV066u LBV072u LBV131u LP1124w LP1145w LP47w LP63w

- Summary of filtering step:

1. Remove individual with mean coverage lower than 14x.
2. Discard SNPs present in none accessible area (defined in ag1000g), QD < 5.00, FS > 60.000 and ReadPosRankSum < -8.000
3. Replace by NA genotype with low confidence (GQ<20)
4. Remove SNP with >5% lmiss
5. Remove Ind with >10% imiss ***

3. Structure of genetic variation

Global genetic structure

PCA

Admixture

Stat descriptives

Deomgraphic history

Investigate recent changes in population size over time.

SNP phasing

SNP polarization

Polarize alleles using as outgroup Anophele Merus and Anophele

IBDne

Infer recent population history (200-500 last generation). Determine change from La Lope village vs forest
Dataset: use data from IBDseq

Stairway plot

Datase: Use polarize alleles

MSMC

Dataset: Use phased SNPs