SeqApiPop analyses: RFMix

Phasing with shapeit

Select SNPs

- MAF > 1
- Max alleles = 2
- Chromosomes as numbers

```
#!/bin/bash

#makeVCF.bash

#select from the vcf with chromosomes indicated as numbers, for plinkAnalyses
#filter on MAF > 0.01
#nb alleles max = 2

module load bioinfo/samtools-1.10
module load bioinfo/vcftools-0.1.15
module load bioinfo/tabix-0.2.5
module load bioinfo/bcftools-1.9

#Select the 629 samples and filter on MAF001

bcftools view  --samples-file ~/plinkAnalyses/WindowSNPs/RFMix/in/IndsAll629.list \
            --min-af 0.01:minor \
            --max-alleles 2 \
            --output-file ~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_629_MAF001_diAllelic_plink.vcf.gz \
            --output-type z \
            ~/plinkAnalyses/MetaGenotypesCalled870_raw_snps_allfilter_plink.vcf
bcftools index SeqApiPop_629_MAF001_diAllelic_plink.vcf.gz
```

VCFs per chromosome

- Shapeit requires one vcf per chromosome

```
#!/bin/bash

#separateChromosomes.bash

module load bioinfo/samtools-1.10
module load bioinfo/vcftools-0.1.15
module load bioinfo/tabix-0.2.5
module load bioinfo/bcftools-1.9

#Select the 629 samples and filter on MAF001

for i in  $(seq 1 16)

do

bcftools view  --regions ${i} \
            --output-file ~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_629_MAF001_diAllelic_plink_chr${i}.v
            ~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_629_MAF001_diAllelic_plink.vcf.gz
bcftools index SeqApiPop_629_MAF001_diAllelic_plink_chr${i}.vcf.gz
done

for i in  $(seq 1 16)
do
bcftools index SeqApiPop_629_MAF001_diAllelic_plink_chr${i}.vcf.gz
done
```

Phasing with Shapeit

Phasing

```
#!/bin/bash

#phasingShapeit.bash

module load bioinfo/shapeit.v2.904

for i in  $(seq 3 16)

do

sbatch --mem=50g --wrap="shapeit --input-vcf --force -O SeqApiPop_629_MAF001_diAllelic_phased_chr${i}.vcf"

done
```

convert phased genotypes back to vcf

```
#!/bin/bash

#convertToVcf.bash

module load bioinfo/shapeit.v2.904

for i in  $(seq 1 16)

do

sbatch --wrap="shapeit -convert --input-haps SeqApiPop_629_MAF001_diAllelic_phased_chr${i}.vcf \
               --output-vcf SeqApiPop_629_MAF001_diAllelic_phased_chr${i}.vcf"

done

for i in  $(seq 1 16)

do

sbatch --wrap="bgzip SeqApiPop_629_MAF001_diAllelic_phased_chr${i}.vcf"

done

for i in  $(seq 1 16)

do

sbatch --wrap="bcftools index SeqApiPop_629_MAF001_diAllelic_phased_chr${i}.vcf.gz"

done
```

Concatenate VCFs

- make list

```
ls ~/plinkAnalyses/WindowSNPs/RFMix/out/*phased*.vcf.gz > vcf.list

$ head vcf.list
~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_629_MAF001_diAllelic_phased_chr1.vcf.gz
~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_629_MAF001_diAllelic_phased_chr2.vcf.gz
~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_629_MAF001_diAllelic_phased_chr3.vcf.gz
```

The chromosomes were out in the correct order by cut and paste in nano

- Concatenate

```
#!/bin/bash

#concatenateVCFs.bash

module load -f  /home/gencel/vignal/save/000_ProgramModules/program_module

bcftools concat -f ~/plinkAnalyses/WindowSNPs/RFMix/out/vcf.list \
                -o ~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_629_MAF001_diAllelic_phased_All.vcf.gz \
                -O z

tabix SeqApiPop_629_MAF001_diAllelic_phased_All.vcf.gz
```

- Remove unnecessary Files

```
rm shapeit*
rm *.haps
rm *.sample
rm *chr*
rm SeqApiPop_629_MAF001_diAllelic_plink.vcf.gz*
```

RFMix

Select reference and query Samples

Make lists of samples: see Jupyter notebook "Treemix"

Select from an Admixture Q matrix with K = 3 the individuals with > 0.95 pure backgrounds as reference => IndsReference.list
The other samples => IndsQuery.list

Make bcf files

- Reference

```
#!/bin/bash

#selectBcfRef.bash

module load bioinfo/bcftools-1.9

#Select the 629 samples and filter on MAF001

bcftools view  --samples-file ~/plinkAnalyses/WindowSNPs/RFMix/in/IndsReference.list \
            --output-file ~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_320_MAF001_diAllelic_phased_Ref.bcf
            --output-type b \
            ~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_629_MAF001_diAllelic_phased_All.vcf.gz
bcftools index SeqApiPop_320_MAF001_diAllelic_phased_Ref.bcf
```

- Query

```
#!/bin/bash

#selectBcfQuery.bash
```

```
module load bioinfo/bcftools-1.9

bcftools view  --samples-file ~/plinkAnalyses/WindowSNPs/RFMix/in/IndsQuery.list \
               --output-file ~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_309_MAF001_diAllelic_phased_Quer
               --output-type b \
               ~/plinkAnalyses/WindowSNPs/RFMix/out/SeqApiPop_629_MAF001_diAllelic_phased_All.vcf.gz
bcftools index SeqApiPop_309_MAF001_diAllelic_phased_Query.bcf
```

Run RFMix

Add population colums to IndsReference.list => IndsPopReference.list

```
head IndsPopReference.list
Ab-PacBio       Black
BER10   Yellow
BER11   Yellow
BER12   Yellow
BER13   Yellow
BER14   Yellow
BER15   Yellow
BER16   Yellow
BER18   Yellow
BER19   Yellow
```

```
#!/bin/bash

#LanceRunRFMix.bash

for i in  $(seq 1 16)
do
sbatch --mem=30g ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/scripts/runRFMixWithGenetMap.bash ${i}

done
```

```
#!/bin/bash

#runRFMixWithGenetMap.bash

module load bioinfo/rfmix-9505bfa

rfmix -f ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/in/SeqApiPop_Pure95_MAF001_diAllelic_phased_Query.bcf \
        -r ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/in/SeqApiPop_Pure95_MAF001_diAllelic_phased_Ref.bcf \
        --chromosome=${1} \
        -m ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/in/IndsPopReference.list \
        -g ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/in/GenetMap_march_2021_AV.txt \
        -o ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/out/SeqApiPopRfmixChr${1}
```

Couldn't get chromosome 1 to work

The generation of internal simulation samples for estimating the Conditional Random Field Weight went on for ever

For the other chromosomes, the CRF values used by the software after the simulation were as follow:

Loading genetic map for chromosome 2 … done
Maximum scoring weight is 53 (91.1)
Loading genetic map for chromosome 3 … done
Maximum scoring weight is 24 (94.3)
Loading genetic map for chromosome 4 … done

Maximum scoring weight is 31 (90.7)
Loading genetic map for chromosome 5 … done
Maximum scoring weight is 44 (94.5)
Loading genetic map for chromosome 6 … done
Maximum scoring weight is 78 (93.6)
Loading genetic map for chromosome 7 … done
Maximum scoring weight is 57 (92.6)
Loading genetic map for chromosome 8 … done
Maximum scoring weight is 33 (88.4)
Loading genetic map for chromosome 9 … done
Maximum scoring weight is 51 (89.9)
Loading genetic map for chromosome 10 … done
Maximum scoring weight is 26 (91.2)
Loading genetic map for chromosome 11 … done
Maximum scoring weight is 23 (88.4)
Loading genetic map for chromosome 12 … done
Maximum scoring weight is 29 (93.2)
Loading genetic map for chromosome 13 … done
Maximum scoring weight is 77 (93.8)
Loading genetic map for chromosome 14 … done
Maximum scoring weight is 94 (91.6)
Loading genetic map for chromosome 15 … done
Maximum scoring weight is 53 (93.5)
Loading genetic map for chromosome 16 … done
Maximum scoring weight is 49 (93.8)

These are not related to the chromosome size.
The mean value is 48, so chromosome 1 was run with this fixed value.

```
#!/bin/bash

#LanceRunRFMix.bash

for i in  $(seq 1 1)
do
sbatch --mem=30g ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/scripts/runRFMixWithGenetMap.bash ${i}
done
```

```
#!/bin/bash

#runRFMixWithGenetMap_Chr1_CRF_48.bash

module load bioinfo/rfmix-9505bfa

rfmix -f ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/in/SeqApiPop_Pure95_MAF001_diAllelic_phased_Query.bcf \
        -r ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/in/SeqApiPop_Pure95_MAF001_diAllelic_phased_Ref.bcf \
        --chromosome=1 \
        -m ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/in/IndsPopReference.list \
        -g ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/in/GenetMap_march_2021_AV.txt \
        -o ~/plinkAnalyses/WindowSNPs/RFMix/Pure95/out/SeqApiPopRfmixChr1_CRF48 \
        --crf-weight=48
```