Tutorial on Variability Analysis of NGS data

Valencia, 7th March 2017

Paths relative to the folder ./AdaptNet_DATA/Scripts

Define paths for the sorce data and for the analyes:

```
PATH_REF=../Reference_seq
PATH_BAM=../BAM
PATH_ANALYSIS=../Analysis
PATH_SCRIPTS=.
```

Download and compile GHcaller, fastaconvtr and mstatspop:

First define the directory:

```
mkdir ../Software
cd ../Software
```

Download GHcaller:

https://bioinformatics.cragenomica.es/projects/ghcaller/index.php

```
curl -O https://bioinformatics.cragenomica.es/projects/ghcaller/binaries/ghcaller-serial_0.1.0_src.tgz
tar -xvf ghcaller-serial_0.1.0_src.tgz
cd ghcaller
g++ -Wall -g -std=c++0x -O3  ./sources/*.cpp -o ./GHcaller #Linux
#clang++ -std=c++11 -stdlib=libc++ -O3 -Wall ./sources/*.cpp -o ./GHcaller #for MacOS
cd ..
```

Download fastaconvtr:

https://bioinformatics.cragenomica.es/numgenomics/people/sebas/software

```
curl -O https://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/files/page3_5.zip
unzip page3_5.zip
cd ./fastaconvtr_pack20170215
sh ./compile_fastaconvtr.sh
cd ..
```

Download mstatspop:

https://bioinformatics.cragenomica.es/numgenomics/people/sebas/software

```
curl -O https://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/files/page3_4.zip
unzip page3_4.zip
cd ./mstatspop_pack_version20170224
sh ./compile_mstatspop.sh
cd ..
```

Finally come back to the folder 'Scripts'

```
cd ../Scripts
```

Include the paths where the executable programs are:

```
PATH_GHCALLER=../Software/ghcaller
PATH_FASTACNVTR=../Software/fastaconvtr_pack20170215/bin
PATH_MSTATSPOP=../Software/mstatspop_pack_version20170224/bin
```

and define the sequence length:

```
len=250000
```

Analysis of Variability for two sampled populations:

Once we have the packages installed, we want to obtain the levels of patterns of variability for two populations of 10 diploid individuals each. We will calculate the levels of variability for all the available positions (includingt hose with missing data) and also we will calculate the level of differentiation.

```
mkdir ../Analysis
```

Indexing the ref fasta

```
REF_SEQ=data_reference_seq.fa
samtools faidx $PATH_REF/$REF_SEQ
```

Loop to transform BAM -> mpileup -> SNPcaller -> fasta files:

#First calculate the mean depth for individual

```
MIN=""; MAX=""; LISTBAM=""
for bamfile in $( ls $PATH_BAM | grep \.bam$); do
  #indexing the bamfiles
  samtools index $PATH_BAM/$bamfile

  #list  of all bam files
  LISTBAM=$(echo "$LISTBAM $PATH_BAM/$bamfile")

  #calculate the mean read depth:
  samtools depth $PATH_BAM/$bamfile > $PATH_ANALYSIS/${bamfile}_read.depth.txt
  mrd=$(cat $PATH_ANALYSIS/${bamfile}_read.depth.txt | awk '{sumrd +=$3} END {print sumrd/'$len'}')
  echo $mrd  > $PATH_ANALYSIS/${bamfile}_MEAN_read.depth.txt
  MINi=`echo $mrd / 2 | bc -l`
  MAXi=`echo $mrd + $mrd | bc -l`
  MIN=$(echo "$MIN,$MINi")
  MAX=$(echo "$MAX,$MAXi")
```

```
done
MIN=$(echo $MIN | sed 's/^,\(.*\)/\1/') #erase the first comma
MAX=$(echo $MAX | sed 's/^,\(.*\)/\1/') #erase the first comma
```

Creating the mpileup file

```
samtools mpileup -q 20 -Q 20 -B -f $PATH_REF/$REF_SEQ \
      $LISTBAM > $PATH_ANALYSIS/Adaptnet_multialign.mpileup
```

Run SNP/base-frequency caller

```
cat $PATH_ANALYSIS/Adaptnet_multialign.mpileup | $PATH_GHCALLER/GHcaller \
    -outfile $PATH_ANALYSIS/Adaptnet_multialign.fa -baseq 20 -mindep $MIN -maxdep $MAX \
    -platform 33 -outgroup $PATH_REF/$REF_SEQ \
    -names P0ind0,P1ind4,P0ind1,P0ind2,P0ind3,P0ind4,P1ind0,P1ind1,P1ind2,P1ind3
```

Convert 'fasta' into 'tfasta' for running sliding window analysis with mstatspop

Order the sequences per population with the option -O.

```
$PATH_FASTACNVTR/fastaconvtr -F f -i $PATH_ANALYSIS/Adaptnet_multialign.fa  -f t \
      -o $PATH_ANALYSIS/Adaptnet_multialign.tfa.gz \
      -O 21 0 1 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 2 3 20
```

Analysis using sliding windows with mstatspop

Here we include positions with missing data:

```
$PATH_MSTATSPOP/mstatspop -f tfa -i $PATH_ANALYSIS/Adaptnet_multialign.tfa.gz \
      -o 1 -u 1 -G 1 -N 3 10 10 1 -T $PATH_ANALYSIS/Adaptnet_multialign_statistics.txt \
      -w 5000 -z 5000
```

Select a number of statistics

Say thetaW, ThetaFL, thetaFW TajimaD, FuLiD, FayWuH and Fst:

```
perl $PATH_SCRIPTS/collect_data_columns.pl -in $PATH_ANALYSIS/Adaptnet_multialign_statistics.txt \
     -fc $PATH_SCRIPTS/choosen_cols.txt > $PATH_ANALYSIS/Adaptnet_multialign_statistics_selected.txt
```

Plot results for a number of statistics

```
R --vanilla < $PATH_SCRIPTS/plot_statistics.R
```

Contents of the script 'plot.statistics.R'

```r
data <- read.table(file="../Analysis/Adaptnet_multialign_statistics_selected.txt",header=T,nrows=50)

pdf(file="../Analysis/Adaptnet_multialign_statistics_selected.pdf",height=5,width=10)

plot(data[,1],data[,6]/data[,2],xlab="position",ylab="Theta Watt",
     col="blue",type="l",ylim=c(0,0.012),
     main="Watterson variability levels per position. (blue:pop1 red:pop2)")
lines(data[,1],data[,7]/data[,3],col="red",type="l")

plot(data[,1],data[,8]/data[,2],xlab="position",ylab="Theta Watt",
     col="blue",type="l",ylim=c(0,0.012),
     main="Tajima (Pi) variability levels per position. (blue:pop1 red:pop2)")
lines(data[,1],data[,9]/data[,3],col="red",type="l")

plot(data[,1],data[,10],xlab="position",ylab="Tajima's D",
     col="blue",type="l",ylim=c(-2.5,1.5),
     main="Tajima's D test. (blue:pop1 red:pop2)")
lines(data[,1],data[,11],col="red",type="l")

plot(data[,1],data[,12],xlab="position",ylab="Fu & Li's D*",
     col="blue",type="l",ylim=c(-2.5,1.5),
     main="Fu and Li's D test. (blue:pop1 red:pop2)")
lines(data[,1],data[,13],col="red",type="l")

plot(data[,1],data[,16],xlab="position",ylab="Fay & Wu's H",
     col="blue",type="l",ylim=c(-5,1.5),
     main="Fay and Wu's H test. (blue:pop1 red:pop2)")
lines(data[,1],data[,17],col="red",type="l")

plot(data[,1],data[,18],xlab="position",ylab="Fst",
     col="black",type="l",main="Fst values betwen two populations")
dev.off()
```