This exercise revolves around obtaining site frequency spectra from two different populations and see if we can find any sign of selection.

The input data are simulated bcffiles and should reflect chr20 in a european population.

ISIN fill in information regarding seqdepth and error rate

Sitefrequency spectrum and VCF files

First setup some paths and environment variables

```
DATA="WORKSHOP_DATA/data/bcf"
ANGSD="angsd/angsd"
REALSFS="angsd/misc/realSFS"
THETASTAT="angsd/misc/thetaStat"
```

Validate that we have setup our variables correctly

```
ls ${DATA} ${ANGSD} ${REALSFS} ${THETASTAT}
```

```
angsd/angsd  angsd/misc/realSFS  angsd/misc/thetaStat
```

```
WORKSHOP_DATA/data/bcf:
chr20.fa.gz  chr20.fa.gz.fai  chr20.fa.gz.gzi  POP1.bcf  POP1.bcf.csi
POP2.bcf  POP2.bcf.csi
```

###Understanding VCF/BCF

1. How many sites do we have in the BCFS
2. How many individuals do we have in the BCFS

Site allele frequencies

```
${ANGSD} -vcf-gl ${DATA}/POP1.bcf -doSaf 1 -anc ${DATA}/chr20.fa.gz -out
POP1
${ANGSD} -vcf-gl ${DATA}/POP2.bcf -doSaf 1 -anc ${DATA}/chr20.fa.gz -out
POP2


##if the above runs takes forever, (it took 5minutes on my desktop),
##we can limit the analyses to 20megabases in the central part of
chromosome20
${ANGSD} -vcf-gl ${DATA}/POP1.bcf -doSaf 1 -anc ${DATA}/chr20.fa.gz -out
POP1 -r chr20:20000000-40000000
${ANGSD} -vcf-gl ${DATA}/POP2.bcf -doSaf 1 -anc ${DATA}/chr20.fa.gz -out
POP2 -r chr20:20000000-40000000
```

Which files was generated?

Filetype	Explanation
arg	arguments used for the analysis
saf.gz	containing the sample allele frequencies for all sites
saf.pos.gz	containing the position
saf.idx	index file containing the binary offset

Site frequency spectrum

The data are the sample allele frequency loglikelihoods these can be viewed with:

```
${REALSFS} print FILE.saf.idx|head
```

The first two columns are the chromosome and position followed by the saf for each bin. We can obtain an estimate of the global site frequency spectrum for each popoulation using the following commands

```
${REALSFS} POP1.saf.idx > POP1.sfs
${REALSFS} POP2.saf.idx > POP2.sfs
```

Have a look at the .sfs files. If you had problems generating them, they can also be found here

```
cat POP1.sfs
cat POP2.sfs
```

We need to plot these, we will use R

```
p1 <- scan("POP1.sfs")
p2 <- scan("POP2.sfs")
barplot(p1)
barplot(p2)
barplot(p1[-1])
barplot(p2[-1])
```

See plot here

1. How many segregating(variable) sites do we have in each of the populations?
2. What is the probability of variability ?

```
sum(p1[-1])
sum(p2[-1])
sum(p1[-1])/sum(p1)
sum(p2[-1])/sum(p2)
```

Why are we discarding the first bin? Should we discard other bins?

Allele frequency posterior probabilities and associated statistics

We are interested in performing a sliding window analyses using various estimates of the population scaled mutation rate. We can precompute the persite theta estimates by using the following commands

```
${REALSFS} saf2theta POP1.saf.idx -sfs POP1.sfs -outname POP1
${REALSFS} saf2theta POP2.saf.idx -sfs POP2.sfs -outname POP2
```

Which files was generated?

Filetype	Explanation
.thetas.gz	containing the thetas persite
.thetas.idx	index file containing the binary offset

We can view the theta statistics using `${THETASTAT} print thetas.idx` . This file contains log scaled per site estimates of the thetas.

```
${THETASTAT} print POP1.thetas.idx
```

```
$thetaStat print testout.thetas.idx 2>/dev/null |head
#Chromo Pos     Watterson       Pairwise        thetaSingleton  thetaH
thetaL
chr20   1       -13.837903      -15.382814      -12.393384
-19.039749      -16.050478
chr20   2       -14.297906      -15.843701      -12.852455
-19.502541      -16.511412
chr20   3       -13.446123      -14.991596      -12.001015
-18.649746      -15.659290
chr20   4       -12.615373      -14.158298      -11.172954
-17.810963      -14.825852
chr20   5       -14.952734      -16.499620      -13.506134
-20.160820      -17.167391
chr20   6       -11.360343      -12.901918      -9.919370
-16.551733      -13.569401
chr20   7       -14.651113      -16.197880      -13.204640
-19.858820      -16.865644
chr20   8       -14.741365      -16.288082      -13.294944
-19.948916      -16.955843
chr20   9       -8.865955       -10.400315      -7.432686
-14.034883      -11.067410
```

Column index	1	2	3	4
Column ID	#Chromo	Pos	Watterson	Pairwise

Column index	1	2	3	4
Example data	chr20	1	-13.837903	-15.382814
Explanation	Contig name	Position	Watterson's theta	ThetaD Nucleotide diversity
Formula (if relevant)			$\sum_{i=1}^{n-1} \eta_i / a^{-1}, a = \sum_{i=1}^{n-1} i$	$\binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n -$

Sliding window

We can do a sliding window analysis using a window size of 50kb and a step size of 10kb:

```
${THETASTAT} do_stat POP1.thetas.idx -win 100000 -step 10000  -outnames
POP1
${THETASTAT} do_stat POP2.thetas.idx -win 100000 -step 10000  -outnames
POP2
```

`pestPG` contains the sum of the per site estimates for a region

```
#(indexStart,indexStop)(firstPos_withData,lastPos_withData)
(WinStart,WinStop)   Chr     WinCenter       tW      tP      tF      tH
tL    Tajima   fuf     fud     fayh    zeng    nSites
(0,63025519)(1,63025520)(0,63025520)    chr20   31512760
29084.489811    29094.351398    29120.408460    34251.072423
31672.711913  0.001278 -0.001687       -0.003197       -0.142269
0.072371        63025519
```

Let us try load the data into R and plot it

```
p1<-read.table("POP1.pestPG",header=F)
colnames(p1)
<-c("Index","Chr","WinCenter","tW","tP","tF","tH","tL","Tajima","fuf","fud
p2<-read.table("POP2.pestPG",header=F)
colnames(p2)
<-c("Index","Chr","WinCenter","tW","tP","tF","tH","tL","Tajima","fuf","fud
```

```
plot(p1$WinCenter,p1$Tajima)
plot(p2$WinCenter,p2$Tajima)

#or on same plot
plot(p2$WinCenter/1e6,p2$Tajima,col='blue',lwd=2,type='l',ylim=range(c(p1$
in MB",ylab="Tajimas D")
lines(p1$WinCenter/1e6,p1$Tajima,col='red',lwd=1)
legend("bottomright",c("POP1","POP2"),fill=c("red","blue"))
```

Plots can also be found here