The data is from the 1000 genomes project which included the populations:

Population	Explanation
CEU	Europeans (mostly of British ancestry)
JPT	East Asian - Japanese individuals
YRI	West African - Nigerian Yoruba individuals

Due to computation We will use a very reduced data set:

- Input data: bam files
- 10 individuals from each population
- a very reduced genome 30 x 100k random regions across the autosomes
    - a non-random region
- Each individual is sequenced at 2-6X

Aims:

- To reconstruct the SFS (1D and 2D)

- To estimate Fst between pairs of popualtions
- To perform a scan statistics using PBS to detect signs of positive selection

First set some paths

Set some environment variables and paths

```
# NB this must be done every time you open a new terminal


# Set path to ANGSD program
ANGSD=angsd/angsd

#realSFS
REALSFS=angsd/misc/realSFS

#ancestral fasta file (chimp)
ANC=hg19ancNoChr.fa.gz

#reference genome for human
REF=hg19.fa.gz

# a bam filelist for a several bam files
BAMFOLDER=dt/smallerbams
BAMFOLDERchr5=dt/chr5_33M_v2

#copy R plot function to folder
cp dt/plot2dSFS.R .
```

Let us first validate that we setup our variables correctly

```
ls ${ANGSD} ${REALSFS} ${ANC} ${REF} ${BAMFOLDER} ${BAMFOLDERchr5}
plot2dSFS.R
```

Make some file lists of bam files

```
#a African population
find $BAMFOLDER | grep bam$ | grep YRI > YRI.filelist
#a Asian population
find $BAMFOLDER | grep bam$ | grep JPT > JPT.filelist
```

```
#a European population
find $BAMFOLDER | grep bam$ | grep CEU > CEU.filelist
```

Let us see how many samples we have in each population

```
wc -l *.filelist
```

Reconstructing the site frequency spectrum

First lets set some filter to remove the worst reads (minMapQ), remove the worst of the bases (minQ).

```
FILTERS="-minMapQ 30 -minQ 20"
```

Lets set some options that means we will calculate genotype likelihoods using the GATK model (gl) and calculate the site allele frequency likelihoods (saf)

```
OPT=" -dosaf 1 -gl 2"
```

Generate site frequency likelihoods using ANGSD

```
$ANGSD -b  YRI.filelist  -anc $ANC -out yri $FILTERS $OPT -ref $REF &
$ANGSD -b  JPT.filelist  -anc $ANC -out jpt $FILTERS $OPT -ref $REF &
$ANGSD -b  CEU.filelist  -anc $ANC -out ceu $FILTERS $OPT -ref $REF
```

The run time is a couple of minutes

If it talks to long then you can copy the results using this command:

```
cp dt/yri.saf* .
cp dt/ceu.saf* .
cp dt/jpt.saf* .
```

Estimate the site frequency spectrum for each of the 3 populations without having to call genotypes or variable sites directly from the site frequency likelihoods

```
#calculate the 1 dimensional SFS
$REALSFS yri.saf.idx > yri.sfs
$REALSFS jpt.saf.idx > jpt.sfs
$REALSFS ceu.saf.idx > ceu.sfs
```

In order to plot the results open R and make a barplot

```
 ##run in R
#plot the results
nnorm <- function(x) x/sum(x)
#expected number of sites with 1:20 derived alleles
res <- rbind(
  YRI=scan("yri.sfs")[-1],
  JPI=scan("jpt.sfs")[-1],
  CEU=scan("ceu.sfs")[-1]
)
colnames(res) <- 1:20

# density instead of expected counts
res <- t(apply(res,1,nnorm))

#plot the none ancestral sites
barplot(res,beside=T,legend=c("YRI","JPT","CEU"),names=1:20,main="realSFS
non ancestral sites")

#plot the polymorphic sites.
resPoly <- t(apply(res[,-20],1,nnorm))
barplot(resPoly,beside=T,legend=c("YRI","JPT","CEU"),names=1:19,main="reals
polymorphic sites")

#due the very limited amount of sites
#downsample to 5 individuals (10 chromosome) and exclude fixed derived
downsampleSFS <- function(x,chr){ #x 1:2n , chr < 2n
    n<-length(x)
    mat <- sapply(1:chr,function(i) choose(1:n,i)*choose(n- (1:n),chr-
i)/choose(n,chr))
    nnorm( as.vector(t(mat) %*% x)[-chr] )
}
resDown <- t(apply(res,1,downsampleSFS,chr=10))
barplot(resDown,beside=T,legend=c("YRI","JPT","CEU"),names=1:9,main="realSF
downsampled polymorphic sites")
```

If you had problems with the above commands the plots can also be found her

- Which population has the largest population size?

- The data is a small subset of the genome (2Mb). If you had analysed 6Mb it sould have looked like this
- The analysed whole chromosome for the 1000G individual look like this

lets use the sfs to calculate some statistics for the population

```
 ##run in R
## read sfs
yri<-scan("yri.sfs");
jpt<-scan("jpt.sfs");
ceu<-scan("ceu.sfs");

x<-ceu #change this one to try one of the other populations

nSites<-sum(x)   #Number of sites where we have data
nSeg<-sum(x[c(-1,-21)])    #Number of segregating sites
an <- function(n) sum(1/1:(n-1))
thetaW <- nSeg/an(20) # Wattersons Theta
thetaW / 2.5e-8 / nSites / 4 # effective population size
```

The above example is for the African population. Try to run it for all three populations.

- which has the largest populations size
- which has the largest variability (fraction of polymorphic/segregating sites)

Fst and PBS In order to estimate Fst between two population we will need to estimate the 2-dimensional frequency spectrum from the site allele frequency likelihoods

```
#calculate the 2D SFS
$REALSFS yri.saf.idx ceu.saf.idx >yri.ceu.ml &
$REALSFS yri.saf.idx jpt.saf.idx >yri.jpt.ml &
$REALSFS jpt.saf.idx ceu.saf.idx >jpt.ceu.ml
```

Plot the results in R

```
 ##run in R
```

```
yc<-scan("yri.ceu.ml")
yj<-scan("yri.jpt.ml")
jc<-scan("jpt.ceu.ml")
    source("plot2dSFS.R")
plot2<-function(s,...){
    dim(s)<-c(21,21)
    s[1]<-NA
    s[21,21]<-NA
s<-s/sum(s,na.rm=T)

    pal <-
color.palette(c("darkgreen","#00A600FF","yellow","#E9BD3AFF","orange","red
space="rgb")
    pplot(s/sum(s,na.rm=T),pal=pal,...)
}

plot2(yc,ylab="YRI",xlab="CEU")
x11()
plot2(yj,ylab="YRI",xlab="JPT")
x11()
plot2(jc,ylab="JPT",xlab="CEU")
```

If you had problems running the above commands the plots can be found here

Due to the very limited amount of data the plots are very noisy. However they are still informative.The colors indicate the density. High density means many sites will look like this and low density (green) means that few sites looks like this.

Based on the plots try to guess

- Which populations has most private SNPs (sites that are only polymorphic in this population)
- Which two populatons are most closely related?

close R

In order to get a measure of this populations are most closely related we willl estimate the pairwise Fst

```
#first will will index the sample so the same sites are analysed for
each population
$REALSFS fst index jpt.saf.idx ceu.saf.idx -sfs jpt.ceu.ml -fstout
jpt.ceu
$REALSFS fst index yri.saf.idx ceu.saf.idx -sfs yri.ceu.ml -fstout
yri.ceu
```

```
$REALSFS fst index yri.saf.idx jpt.saf.idx -sfs yri.jpt.ml -fstout
yri.jpt

#get the global estimate
$REALSFS fst stats jpt.ceu.fst.idx
$REALSFS fst stats yri.jpt.fst.idx
$REALSFS fst stats yri.ceu.fst.idx
```

look at the weigthed Fst (Fst.Weight).

- which two populations are most closely related?
- which two populations are most distantly related?

Lets see how the Fst and PBS varies between different regions of the genome my using a sliding windows approach (windows site of 50kb)

```
$REALSFS fst index yri.saf.idx jpt.saf.idx ceu.saf.idx -fstout
yri.jpt.ceu -sfs yri.jpt.ml -sfs yri.ceu.ml -sfs jpt.ceu.ml
$REALSFS fst stats2 yri.jpt.ceu.fst.idx -win 50000 -step 10000
>slidingwindowBackground
```

read the data into R

```
 ##run in R
r<-read.delim("slidingwindowBackground",as.is=T,head=T)
names(r)[-c(1:4)] <-
c("wFst_YRI_JPT","wFst_YRI_CEU","wFst_JPT_CEU","PBS_YRI","PBS_JPT","PBS_CEU


head(r) #print the results to the screen

#plot the distribution of Fst
mmax<-max(c(r$wFst_YRI_JPT,r$wFst_YRI_CEU,r$wFst_JPT_CEU),na.rm=T)
par(mfcol=c(3,2))
hist(r$wFst_YRI_JPT,col="lavender",xlim=c(0,mmax),br=20)
hist(r$wFst_YRI_CEU,col="mistyrose",xlim=c(0,mmax),br=20)
hist(r$wFst_JPT_CEU,col="hotpink",xlim=c(0,mmax),br=20)

mmax<-max(c(r$PBS_CEU,r$PBS_YRI,r$PBS_JPT),na.rm=T)

#plot the distribution of PBS
mmax<-max(c(r$PBS_CEU,r$PBS_YRI,r$PBS_JPT),na.rm=T)
hist(r$PBS_YRI,col="lavender",xlim=c(0,mmax),br=20)
hist(r$PBS_CEU,col="mistyrose",xlim=c(0,mmax),br=20)
```

```
hist(r$PBS_JPT,col="hotpink",xlim=c(0,mmax),br=20)
```

If you had problems running the above commands you can find the result here

note the maximum observed values for both the pairwise fst and the PBS

Lets do the same for not so randomly selection 1Mb region of on chr 5. Remember to close R

```
#a African population for a region on chr 5
find $BAMFOLDERchr5 | grep bam$ | grep YRI > YRIchr5.filelist
#a Asian population for a region on chr 5
find $BAMFOLDERchr5 | grep bam$ | grep JPT > JPTchr5.filelist
#a European population for a region on chr 5
find $BAMFOLDERchr5 |  grep bam$ | grep CEU > CEUchr5.filelist

#use the same filters and options as before
FILTERS="-minMapQ 30 -minQ 20 -baq 1 -C 50 -minInd 8"
OPT=" -dosaf 1 -gl 2"

#get site frequency likelihoods
$ANGSD -b  YRIchr5.filelist  -anc $ANC -out yriChr5 $FILTERS $OPT -ref
$REF
$ANGSD -b  JPTchr5.filelist  -anc $ANC -out jptChr5 $FILTERS $OPT -ref
$REF
$ANGSD -b  CEUchr5.filelist  -anc $ANC -out ceuChr5 $FILTERS $OPT -ref
$REF

#estimate the 1D SFS
$REALSFS yriChr5.saf.idx ceuChr5.saf.idx >yri.ceuChr5.ml
$REALSFS yriChr5.saf.idx jptChr5.saf.idx >yri.jptChr5.ml
$REALSFS jptChr5.saf.idx ceuChr5.saf.idx >jpt.ceuChr5.ml

#get FST and PBS in sliding window
$REALSFS fst index yriChr5.saf.idx jptChr5.saf.idx ceuChr5.saf.idx
-fstout yri.jpt.ceuChr5 -sfs yri.jptChr5.ml -sfs yri.ceuChr5.ml -sfs
jpt.ceuChr5.ml
$REALSFS fst stats2 yri.jpt.ceuChr5.fst.idx -win 50000 -step 10000
>slidingwindowChr5
```

Lets view how it looks in this region

```
#run in R
```

```
r<-read.delim("slidingwindowChr5",as.is=T,head=T)
names(r)[-c(1:4)] <-
c("wFst_YRI_JPT","wFst_YRI_CEU","wFst_JPT_CEU","PBS_YRI","PBS_JPT","PBS_CEU


par(mfrow=1:2)
plot(r$midPos,r$wFst_YRI_CEU,ylim=c(0,max(r$wFst_YRI_CEU)),type="b",pch=18,
on Chr 5")
points(r$midPos,r$wFst_YRI_JPT,col=2,type="b",pch=18)
points(r$midPos,r$wFst_JPT_CEU,col=3,type="b",pch=18)
legend("topleft",fill=1:3,c("YRI vs. CEU","YRI vs. JPT","JPT vs CEU"))

plot(r$midPos,r$PBS_YRI,ylim=c(0,max(r$PBS_CEU)),type="b",pch=18,ylab="PBS'
on Chr 5")
points(r$midPos,r$PBS_JPT,col=2,type="b",pch=18)
points(r$midPos,r$PBS_CEU,col=3,type="b",pch=18)
legend("topleft",fill=1:3,c("YRI","JPT","CEU"))
```

If you problems running the above commands the results can be found here

- Compare the values you observed on this part of the genome with the random pars of the genome you looked at earlier). Is this region extreme?
- Why is there two peak for the Fst and only one for the PBS?
- In which of the populations are this loci under selection?

Find out what genes is in this region by going to the UCSC browser. Choose Genome browser. Choose human GRCh37/hg19 and find the region. Read about this gene on wikipedia and see if this fits PBS results.

What happens if we try to call genotypes?

We can compare with what happens if we try to call genotypes by calling SNPs and genotypes like GATK. If you are running out of time then skip this part

```
FILTERS2="-minMapQ 30 -minQ 20 -minInd 10"

OPT2="-gl 2 -doGeno 2 -doPost 2 -doMajorMinor 4 -doMaf 1 -SNP_pval 1e-6
-postCutoff 0.95"
$ANGSD -b  YRI.filelist  -out yri $FILTERS2 $OPT2 -ref $REF &
$ANGSD -b  JPT.filelist  -out jpt $FILTERS2 $OPT2 -ref $REF &
$ANGSD -b  CEU.filelist  -out ceu $FILTERS2 $OPT2 -ref $REF
```

While it runs you can look at the options we choose:

- *minInd 10*: minimum individuals with data (in this case it means with called genotypers). Why do we need this?
- *-doGeno 2* Print only the count (0,1,2) and not the based e.g. AA,AT,TT
- *-doPost 2* Use uniform prior for genotype i.e. call the genotype with the highest likelihood
- *-DoMajorMinor 4* Use the Ancestral allele from the chimp
- *-doMaf 1 -SNP_pval 1e-6* Use this p-value cutoff to call SNPs. What would happend to the SFS if you change this threshold?
- *-PostCutoff 0.95* only call genotype with a propability above 0.95

Plot the results in R

```
 ##run in R
#plot the results
nnorm <- function(x) x/sum(x)
getSFS<-function(x) table(factor(rowSums(read.table(x)[,-
c(1:2)]),levels=1:20))

res <- rbind(
  YRI=getSFS("yri.geno.gz"),
  JPI=getSFS("jpt.geno.gz"),
  CEU=getSFS("ceu.geno.gz")
)
colnames(res) <- 1:20

# density instead of expected counts
res <- t(apply(res,1,nnorm))

#plot the none ancestral sites
barplot(res,beside=T,legend=c("YRI","JPT","CEU"),names=1:20,main="SFS
from called genotypes")

#plot the polymorphic sites.
resPoly <- t(apply(res[,-20],1,nnorm))
barplot(resPoly,beside=T,legend=c("YRI","JPT","CEU"),names=1:19,main="SFS
from call\
ed genotypes")
```

```
#down sample to 5 individuals (10 chromosome) and exclude fixed derived
downsampleSFS <- function(x,chr){ #x 1:2n , chr < 2n
    n<-length(x)
    mat <- sapply(1:chr,function(i) choose(1:n,i)*choose(n- (1:n),chr-
i)/choose(n,chr))
    nnorm( as.vector(t(mat) %*% x)[-chr] )
}
resDown <- t(apply(res,1,downsampleSFS,chr=10))
barplot(resDown,beside=T,legend=c("YRI","JPT","CEU"),names=1:9)
```

If you had problems running the above commands the result can be found here

- How does this compare to the likelhood based estimates (pdf)