Exercise in estimating nucleotide diversity

Casper-Emil Pedersen, Peter Frandsen and Hans R. Siegismund

Program

- Examine the PLINK-format
- Read SNP data into R and extract information about the data
- Estimate nucleotide diversity (here as the expected heterozygosity) in different populations
- Estimate the inbreeding coefficient for each individual in the different populations
- Plot your results to graphically present the diversity in different population and in different regions along the chromosome

Aim

- Get familiar with the commonly used PLINK-format
- Get familiar with extraction of simple summary statistics of data in R
- Get familiar with representation of results
- Be able to interpret diversity measures in populations

Recommended background reading:

See Prado-Martinez et al. 2013 for a comprehensive great ape paper

http://www.nature.com/nature/journal/v499/n7459/full/nature12228.html

Genetic Diversity in Chimpanzees

During this exercise you will be introduced to population genetic analysis of SNP data. The datasets used here consist of the variable sites found on chromosome 22 in chimpanzees. The data set contains genotypes from all four subspecies of chimpanzee (*Pan troglodytes*, see Figure 1), two human populations, one with European ancestry (CEU) and one with African ancestry (YRI). The data has been filtered to reduce the size of your working data set and includes only SNP's with exactly two different bases (bi-allelic).

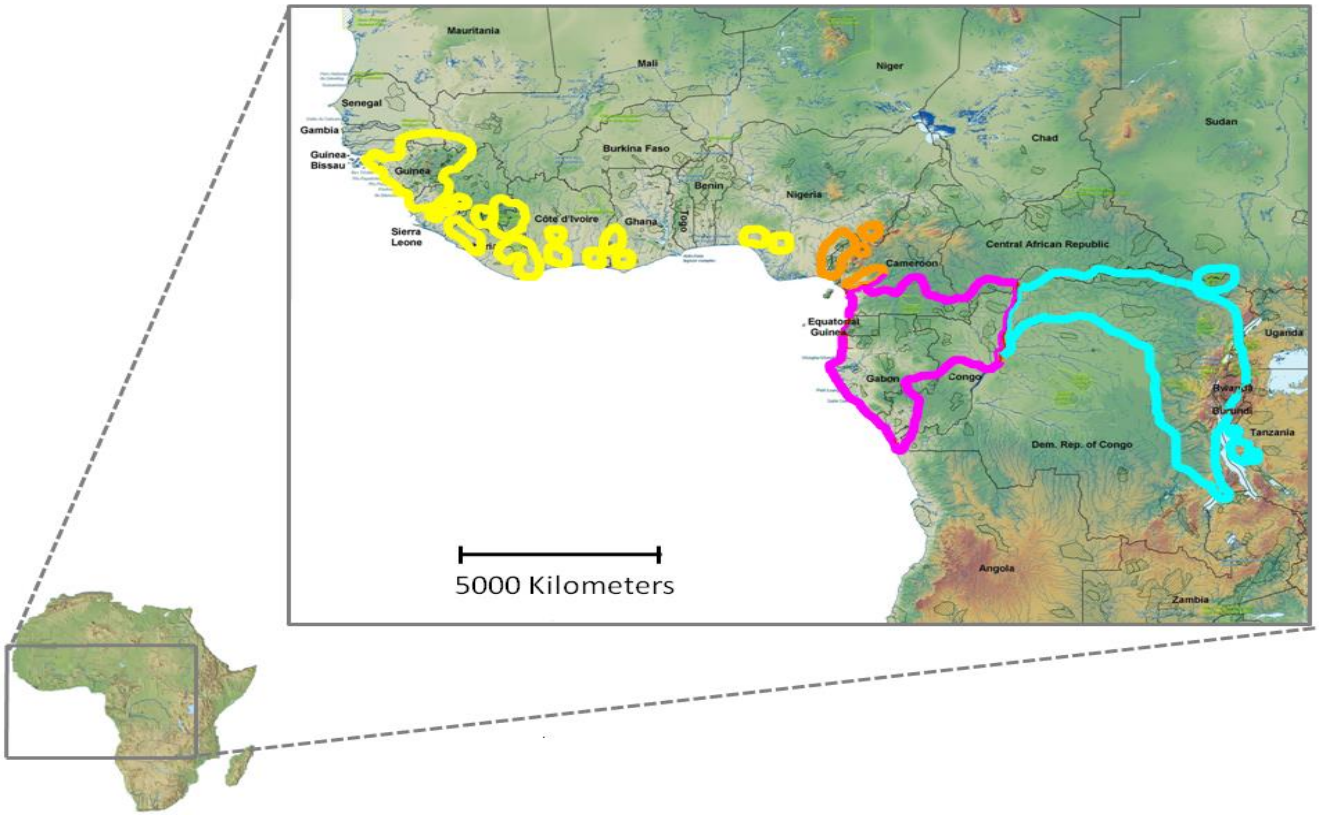


Figure 1 | Geographical distribution ranges for the *Pan troglodytes* subspecies. The range borders of each subspecies; yellow for *Pan troglodytes verus* (western chimpanzee), orange for *Pan troglodytes ellioti* (Nigerian-Cameroon chimpanzee), pink for *Pan troglodytes troglodytes* (central chimpanzee), and blue for *Pan troglodytes schweinfurthii* (eastern chimpanzee). Reprint from (http://www.unep.org/grasp/) modified by C. Hvilsom.

Q1: Before we get started, why do you think we chose chromosome 22?

> Chromosome 22 was chosen because it is one of the smallest somatic chromosomes. We would not be able to complete this exercise if we had chosen whole genomes or even one of the largest chromosomes.

Getting started

Change the directory to the exercise directory and copy your data:

```
cd ~/exercises
cp ~/groupdirs/SCIENCE-BIO-Popgen_Course/exercises/apeDiversity/apeGenDiv.tar.gz .
tar -zxvf apeGenDiv.tar.gz
rm apeGenDiv.tar.gz
cd apeDiversity
```

Now you have all the data in the correct folder so you can proceed to the exercise but first, have a look at the different files and the file format.

The **PLINK format** was originally designed for genotype/phenotype data analyses in association studies but also has a range of features applicable to other disciplines within population genetics. The two main **PLINK files** are the *MAP* and the *PED* files, which often serves as the starting point for any analysis in PLINK. These two files are the standard output from a file conversion from another widely used file format, the Variant Call Format (VCF). Many large-scale genome studies, like the 1000 genome project, use the *VCF* format when they publish their data. This format holds all the information about the variant call (*e.g.* 'read-depth', 'quality'). A large range of analyses can be handled with tools designed for the *VCF* format but most often, this toolset only serves to apply a number of standard filters, while downstream analysis are performed in other formats, like PLINK.

In concert, the *MAP* and *PED* files contains a bi-allelic extraction of the genotype information from the *VCF*. The two files are structured slightly different but together, they hold information about each called variant or SNP in each genotyped individual.

PLINK Flat files (MAP/PED)

(Copy from "Genome Wide Association Study pipeline [GWASpi]**" [http://www.gwaspi.org/?page_id=145])**

PLINK is a very widely used application for analyzing genotypic data. It can be considered the "de-facto" standard of the field, although newer formats are starting to be widespread as well.
The standard PLINK fromat provides sufficient information for a straight-forward association study. You may use the sex and affection fields for GWASpi to perform GWS studies.

MAP files
The fields in a MAP file are:

- Chromosome
- Marker ID
- Genetic distance
- Physical position

Example of a MAP file of the standard PLINK format:			
21	rs11511647	0	26765
X	rs3883674	0	32380
X	rs12218882	0	48172
9	rs10904045	0	48426
9	rs10751931	0	49949
8	rs11252127	0	52087
10	rs12775203	0	52277
8	rs12255619	0	52481

(Note: We do not have information about the physical map (Genetic distance – in centiMorgans) for the data used in this exercise)

PED files

The fields in a PED file are

- Family ID
- Sample ID
- Paternal ID
- Maternal ID
- Sex (1=male; 2=female; 0=unknown)
- Affection (0=unknown; 1=unaffected; 2=affected)
- Genotypes (space or tab separated, 2 for each marker. 0=missing)

Example of a PED file of the standard PLINK format:																					
FAM1	NA06985	0	0	1	1	A	T	T	T	G	G	C	C	A	T	T	T	G	G	C	C
FAM1	NA06991	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G	C	C
0	NA06993	0	0	1	1	C	T	T	T	G	G	C	T	C	T	T	T	G	G	C	T
0	NA06994	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G	C	C
0	NA07000	0	0	2	1	C	T	T	T	G	G	C	T	C	T	T	T	G	G	C	T
0	NA07019	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G	C	C
0	NA07022	0	0	2	1	C	T	T	T	G	G	0	0	C	T	T	T	G	G	0	0
0	NA07029	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G	C	C
FAM2	NA07056	0	0	0	2	C	T	T	T	A	G	C	T	C	T	T	T	A	G	C	T
FAM2	NA07345	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G	C	C

(Note: For our data, the Family ID and the Sample ID are the same. Furthermore, the Paternal ID and Maternal ID is not included. Neither is the Sex or Affection)

For a complete breakdown of the structure of the file formats see:
https://www.cog-genomics.org/plink2/formats#ped

With a starting point in these two file format, the PLINK toolset offers a long list of different ways to analyze your data, for a complete list see https://www.cog-genomics.org/plink2/basic_stats

The PLINK files in this exercise

In your 'apeDiversity' directory, you will find twenty different PLINK files arranged to include variable sites from chromosome 22 in each of the four subspecies of chimpanzee (separately and combined), and the two human populations, for an overview, see Table 1.

Table 1 | Sample overview

Population	File name prefix	n
Chimpanzee (*Pan troglodytes)*	Pan_troglodytes	59
Central chimpanzee (*Pan troglodytes troglodytes*)	Pt_troglo	18
Eastern chimpanzee (*Pan troglodytes schweinfurtii*)	Pt_scwein	19
Western chimpanzee (*Pan troglodytes verus*)	Pt_verus	12
Nigerian-Cameroon chimpanzee (*Pan troglodytes ellioti*)	Pt_ellioti	10
Utah residents with ancestry in Europe (CEU)	CEU	7
Yoruba ethnic group in North and Central Nigeria (YRI)	YRI	7

For each population there is a corresponding MAP and PED file in the 'apeDiversity' directory

To look at the PLINK-files, in turn type (for the CEU or the YRI samples)

less -S FILENAME.map # type **q** to quit

less -S FILENAME.ped

Now, try to look at the *MAP* file containing all chimpanzees ("Pan_troglodytes.map").

Q2: What is the position of the first SNP? (Confer with link above about file format)

22:14436989 (see line one of map file)

Q3: What information is in the two file formats (*MAP* and *PED*)? (Again, look at the file and confer with the link above.)

Position and chromosome in the .map file. Sample information and genotypes in .ped file

Q4: How many SNPs are there in total for the chimpanzees and the human populations? Remember the command **wc -l filename** gives the total number of lines in the file. Now should you count the lines in the *MAP* or the *PED* file?

Chimpanzees: 509051

Human: 494328

Q5: In this format, we will have no information about the certainty of SNP calls. Is it reasonable to assume that *e.g.* read depth might influence the identified number of SNPs? Why / why not ? And would you expect more or less SNPs to be identified, than the true number of SNPs, when using low depth data?

Read depth influences the certainty of the SNP calls, often you exclude SNPs based on sites where you only have a limited read depth.

Using PLINK to find the nucleotide diversity in chimpanzees and humans

Now, having an overview of the different data sets along with a brief idea of what the PLINK format looks like, we can start to analyze our data.

Estimate the genetic diversity

Since we only included bi-allelic SNPs (i.e. SNPs with two bases), we can estimate the diversity as the expected heterozygosity, $H_e = 2p(1 - p)$, assuming HWP. To do this, we will use PLINK and R to calculate the allele frequency of the minor allele for each variable site.

Paste in the following command but make sure to look through each command option and try to understand each called operation.

```
plink --noweb --file Pan_troglodytes --freq --out Pan_troglodytes
plink --noweb --file Pt_ellioti --freq --out Pt_ellioti
plink --noweb --file Pt_schwein --freq --out Pt_schwein
plink --noweb --file Pt_troglo --freq --out Pt_troglo
plink --noweb --file Pt_verus --freq --out Pt_verus
plink --noweb --file CEU --freq --out CEU
plink --noweb --file YRI --freq --out YRI

cat Pan_troglodytes.frq |grep -v NA > Pan_troglodytes_noNA.frq
cat Pt_ellioti.frq |grep -v NA > Pt_ellioti_noNA.frq
cat Pt_schwein.frq |grep -v NA > Pt_schwein_noNA.frq
cat Pt_troglo.frq |grep -v NA > Pt_troglo_noNA.frq
cat Pt_verus.frq |grep -v NA > Pt_verus_noNA.frq
cat CEU.frq |grep -v NA > CEU_noNA.frq
cat YRI.frq |grep -v NA > YRI_noNA.frq
```

Q6: Try and look in the Pan_troglodytes.frq file, what information do you get?

Chromosome, SNP name, Allele 1, Allele 2, Minor Allele Frequency, Number of chromosomes

Q7: We use the command "grep -v NA" and write the output to a new file, what does this command do, and why do you think we do this? (to get an idea, try comparing the number of lines in the Pan_troglodytes.frq file with the number of lines in the Pan_troglodytes_noNA.frq file)

Grep can search for patterns in files. Grep –v does the opposite, excluding lines matching the pattern, here the pattern we DO NOT want to match is NA. This removes all sites with missing data.

Now open **R**. Paste in the following commands to read in the frequency output from PLINK (again, try to understand the code, do not hesitate to ask an instructor, or Google, if in doubt):

```
# Read in each of the frequency files
pt<-read.table("Pan_troglodytes_noNA.frq",h=T)
ellio<-read.table("Pt_ellioti_noNA.frq",h=T)
schwein<-read.table("Pt_schwein_noNA.frq",h=T)
troglo<-read.table("Pt_troglo_noNA.frq",h=T)
verus<-read.table("Pt_verus_noNA.frq",h=T)
ceu<-read.table("CEU_noNA.frq",h=T)
yri<-read.table("YRI_noNA.frq",h=T)
# Function for estimating the expected heterozygosity
het<-function(x){2*x*(1-x)}
# Remove all fixed alleles in each population
verus <- verus[verus[,"MAF"]>0,]
ellio <- ellio[ellio[,"MAF"]>0,]
schwein <- schwein[schwein[,"MAF"]>0,]
troglo <- troglo[troglo[,"MAF"]>0,]
pt <- pt[pt[,"MAF"]>0,]
yri <- yri[yri[,"MAF"]>0,]
ceu <- ceu[ceu[,"MAF"]>0,]
# Add columns with the position on the chromosome
# and the pi-values for each polymorphic SNP
verus <- cbind(verus,position= as.numeric(gsub("22:",'',verus[,"SNP"])))
verus <- cbind(verus, pi=het(verus$MAF)
*(length(verus$MAF)/(verus[length(verus[,"position"]),"position"] - verus
[1,"position"])))
# Pan troglodytes elliotti
ellio <- cbind(ellio,position= as.numeric(gsub("22:",'',ellio[,"SNP"])))
ellio <- cbind(ellio, pi=het(ellio$MAF) *(length(ellio$MAF)/(ellio
[length(ellio[,"position"]),"position"] - ellio [1,"position"])))
# Pan troglodytes schweinfurthii
schwein <- cbind(schwein,position= as.numeric(gsub("22:",'',schwein[,"SNP"])))
schwein <- cbind(schwein, pi=het(schwein$MAF) *(length(schwein$MAF)/(schwein
[length(schwein[,"position"]),"position"] - schwein [1,"position"])))
# Pan troglodytes troglodytes
troglo <- cbind(troglo,position= as.numeric(gsub("22:",'',troglo[,"SNP"])))
troglo <- cbind(troglo, pi=het(troglo$MAF) *(length(troglo$MAF)/(troglo
[length(troglo[,"position"]),"position"] - troglo [1,"position"])))
# Pan troglodytes
pt <- cbind(pt,position= as.numeric(gsub("22:",'',pt[,"SNP"])))
pt <- cbind(pt, pi=het(pt$MAF) *(length(pt$MAF)/(pt
[length(pt[,"position"]),"position"] - pt [1,"position"])))
# No obvious positions for yri and ceu plus a guestimate
# on the length of the included chromosome
yri <- cbind(yri, pi=het(yri$MAF)*(length(yri$MAF)/(35191058)))
ceu <- cbind(ceu, pi=het(ceu$MAF)*(length(ceu$MAF)/(35191950)))
```

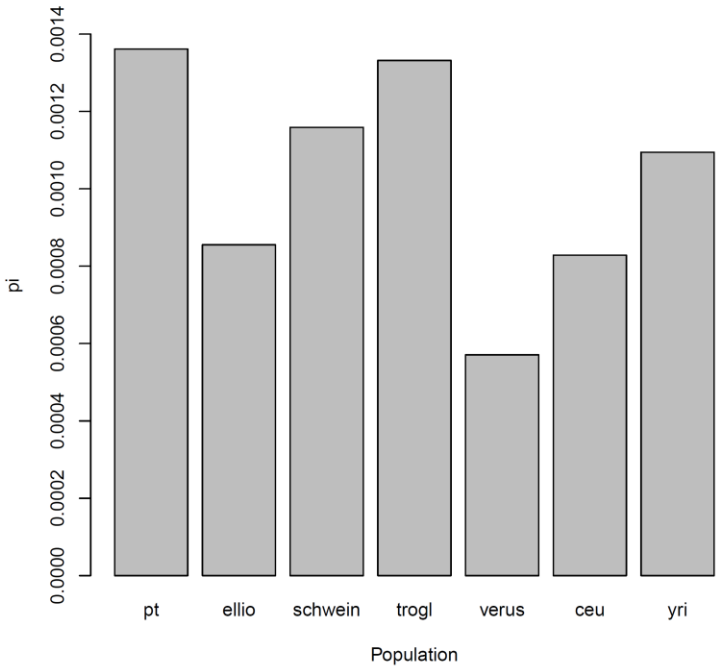
Plotting the results in R

```
# Making a barplot with the nucleotide diversity
par(mfrow=c(1,1))
val = c(mean(pt$pi), mean(ellio$pi), mean(schwein$pi), mean(troglo$pi),
mean(verus$pi), mean(ceu$pi), mean(yri$pi)) #
barplot(val,ylim=c(0.000,0.0015), ylab="pi",  xlab="Population",
names.arg=c("pt","ellio","schwein","trogl","verus","ceu","yri"))
# To shut down the plot window
dev.off()
```

Q8: In the R function (het), explain what 2*x*(1-x) calculates

Calculates heterozygosity based on allele frequencies

Q9: What is the average heterozygosity for the 7 populations? (read from plot or type **mean(**val**))**



Q10: Give a reason why the two human populations differ in heterozygosity?

Differences in demographic history. Europeans have gone through a population bottleneck, reducing their effective population size.

Q11: Why does the combined chimpanzee population ("Pan_troglodytes") have a higher average heterozygosity than each of the subspecies?

The mutations in the four populations are not necessarily unique, meaning that if you pool population it would appear as though the heterozygosity is higher.

Q12 Will rare mutations more often be in heterozygous individuals or homozygous individuals?

The risk that a mutation occurs on the same site in the same time is basically non-existent. It will happen in heterozygotes.

Q13: From your knowledge and from the amount of average heterozygosity, what population would you expect to have the highest N_e? And the lowest?

The central chimpanzee as this population has the highest effective population size based on heterozygosity. We would expect the verus chimpanzee population to have the lowest effective population size. Compared to the other chimpanzee population the verus population has been isolated for a longer time at a smaller census population size.

Estimating the nucleotide diversity along the chromosome

Above, we have estimated the mean heterozygosity for the whole chromosome in each of the populations. In this part of the exercise, we will take a closer look at different regions along the chromosome in order to see if this will tell us something different about the four chimpanzee populations.

Still in R, paste in the following commands:

```
## Function for generating sliding windows
slidingwindowplot <- function(mainv, xlabv, ylabv, ylimv, window.size,
step.size,input_x_data,input_y_data)
{
if (window.size > step.size)
       step.positions  <- seq(window.size/2 + 1, length(input_x_data)- window.size/2,
       by=step.size)
       else
       step.positions  <- seq(step.size/2 + 1, length(input_x_data)- step.size,
       by=step.size)
       n <- length(step.positions)
       means_x <- numeric(n)
       means_y <- numeric(n)
for (i in 1:n) {
       chunk_x <- input_x_data[(step.positions[i]-
       window.size/2):(step.positions[i]+window.size-1)]
              means_x[i] <-  mean(chunk_x,na.rem=TRUE)
       chunk_y <- input_y_data[(step.positions[i]-
       window.size/2):(step.positions[i]+window.size-1)]
              means_y[i] <-  mean(chunk_y,na.rem=TRUE)
       }


plot(means_x,means_y,type="b",main=mainv,xlab=xlabv,ylab=ylabv,ylim=ylimv,cex=0.25,
pch=20,cex.main=0.75)
vec <- c(0.025,0.5,0.975)
zz <- means_y[!is.na(means_y)]
abline(h=quantile(zz,0.025,na.rem=TRUE),col="blue")
abline(h=quantile(zz,0.925,na.rem=TRUE),col="blue")
abline(h=mean(input_y_data))
}
```

```
## Plotting the nucleotide diversity in sliding windows across the chromosome.
## R is doing strange things on the graphics window; therefore, we plot it on
## a pdf file. You can view it with evince afterwards
dev.off()
pdf ("nucleotide_diversity_in_4_subspecies.pdf")
par(mfrow=c(2,2))
windowsize<- 3000
steps<- 100
# Pan troglodytes verus
mainvv = paste("verus pi = ",format(mean(verus$pi,na.rem=TRUE), digits=3), "SNPs =",
length(verus$pi), "Win: ", windowsize, "Step: ", steps)
slidingwindowplot(mainv=mainvv, xlab=expression(paste("Position (x ", 10^6,")")),
ylab=expression(paste("pi")),ylimv=c(0.00,0.0016), window.size=windowsize/4,
step.size=steps, input_x_data=verus$position/1000000,input_y_data=verus$pi)
# Pan troglodytes ellioti
mainvv = paste("ellio pi = ",format(mean(ellio$pi,na.rem=TRUE), digits=3),"SNPs =",
length(ellio$pi),"Win: ", windowsize, "Step: ", steps )
slidingwindowplot(mainv=mainvv, xlab=expression(paste("Position (x ", 10^6,")")),
ylab=expression(paste("pi")),ylimv=c(0.000,0.0016), window.size=windowsize/3,
step.size=steps, input_x_data=ellio$position/1000000,input_y_data=ellio$pi)
# Pan troglodytes schweinfurthii
mainvv = paste("schwein pi = ",format(mean(schwein$pi,na.rem=TRUE), digits=3),"SNPs
=", length(schwein$pi),"Win: ", windowsize, "Step: ", steps )
slidingwindowplot(mainv=mainvv, xlab=expression(paste("Position (x ", 10^6,")")),
ylab=expression(paste("pi")),ylimv=c(0.000,0.0016), window.size=windowsize,
step.size=steps, input_x_data=schwein$position/1000000,input_y_data=schwein$pi)
# Pan troglodytes troglodytes
mainvv = paste("troglo pi =  ",format(mean(troglo$pi,na.rem=TRUE), digits=3),"SNPs
=", length(troglo$pi),"Win: ", windowsize, "Step: ", steps )
slidingwindowplot(mainv=mainvv, xlab=expression(paste("Position (x ", 10^6,")")),
ylab=expression(paste("pi")),ylimv=c(0.00,0.0016), window.size=windowsize,
step.size=steps, input_x_data=troglo$position/1000000,input_y_data=troglo$pi)

# To close the graphical window
dev.off()
```

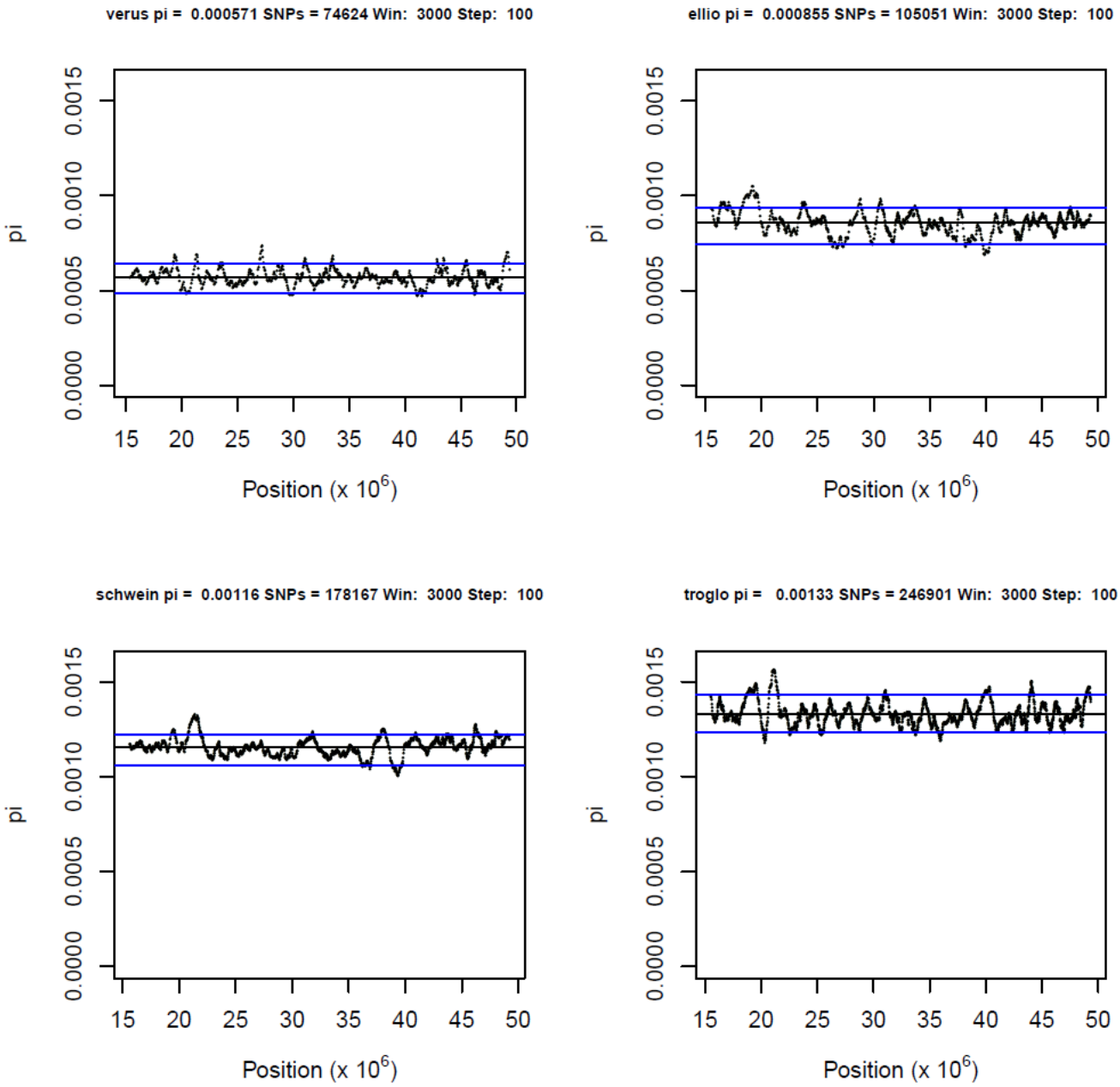
Leave R

```
q()
n
```

Take a look at the output

```
evince nucleotide_diversity_in_4_subspecies.pdf
```

Q14: Why is there a difference in pi along the chromosome?



Differences between coding and non-coding regions might reduce or increase the diversity depending on the constraint of selection.

Q15: Why is the pattern different among the populations?

Populations may be adapted to different environments or have different population sizes and experienced different levels of genetic drift.

Estimating inbreeding coefficient pr. individual

Now, instead of comparing diversity measures in different population, we will now look at the diversity within each individual in term of inbreeding. Given the large number of SNPs for each individual, we will estimate the individual inbreeding coefficient for all individuals in the different populations.

```
plink --file Pt_verus --het --out Pt_verus
plink --file Pt_ellioti --het --out Pt_ellioti
plink --file Pt_schwein --het --out Pt_schwein
plink --file Pt_troglo --het --out Pt_troglo
plink --file YRI --het --out YRI
plink --file CEU --het --out CEU
plink --file Pan_troglodytes --het --out Pan_troglodytes
```

This will produce output files with the extension ".het". Take a look at them. The inbreeding coefficient is found as the last column of this output. Start by looking at the four different chimpanzee subspecies and then the two human populations.

The headings of .het files are:

FID	Family ID
IID	Individual ID
O(HOM)	Observed number of homozygotes
E(HOM)	Expected number of homozygotes
N(NM)	Number of non-missing genotypes
F	F inbreeding coefficient estimate

Q16: Is there a sign of inbreeding in some of the humans?

No, all have an F value close to zero.

Q17: Do some of the chimpanzees show signs of inbreeding?

Verus: two inds with 0.11 and 0.062 F
Ellioti: Five with an F of 0.062 or more.
Troglodytes: Three with an F of 0.062 or more
Schwein: Five with an F of 0.062 or more.

Q18: If so, how related do they seem to be?

First cousin offspring has an F of around 0.0625, uncle-niece an F of 0.125 and offspring of brother sister around 0.25. Keep in mind that there is some variation around this number.

Now take a look at the total sample of the combined set of chimpanzees ("Pan_troglodytes.het").

Q19: What is going on here? Why are the inbreeding coefficients so high?

The Wahlund effect, where we see more homozygotes than what we expect from random mating, because we pool different populations into one.