GENETIC DIVERSITY

POPULATION GENETICS 2020

PATRÍCIA CHRZANOVÁ PEČNEROVÁ

PHOTO CREDIT : MOGENS TROLLE

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 FOR THE EXERCISE

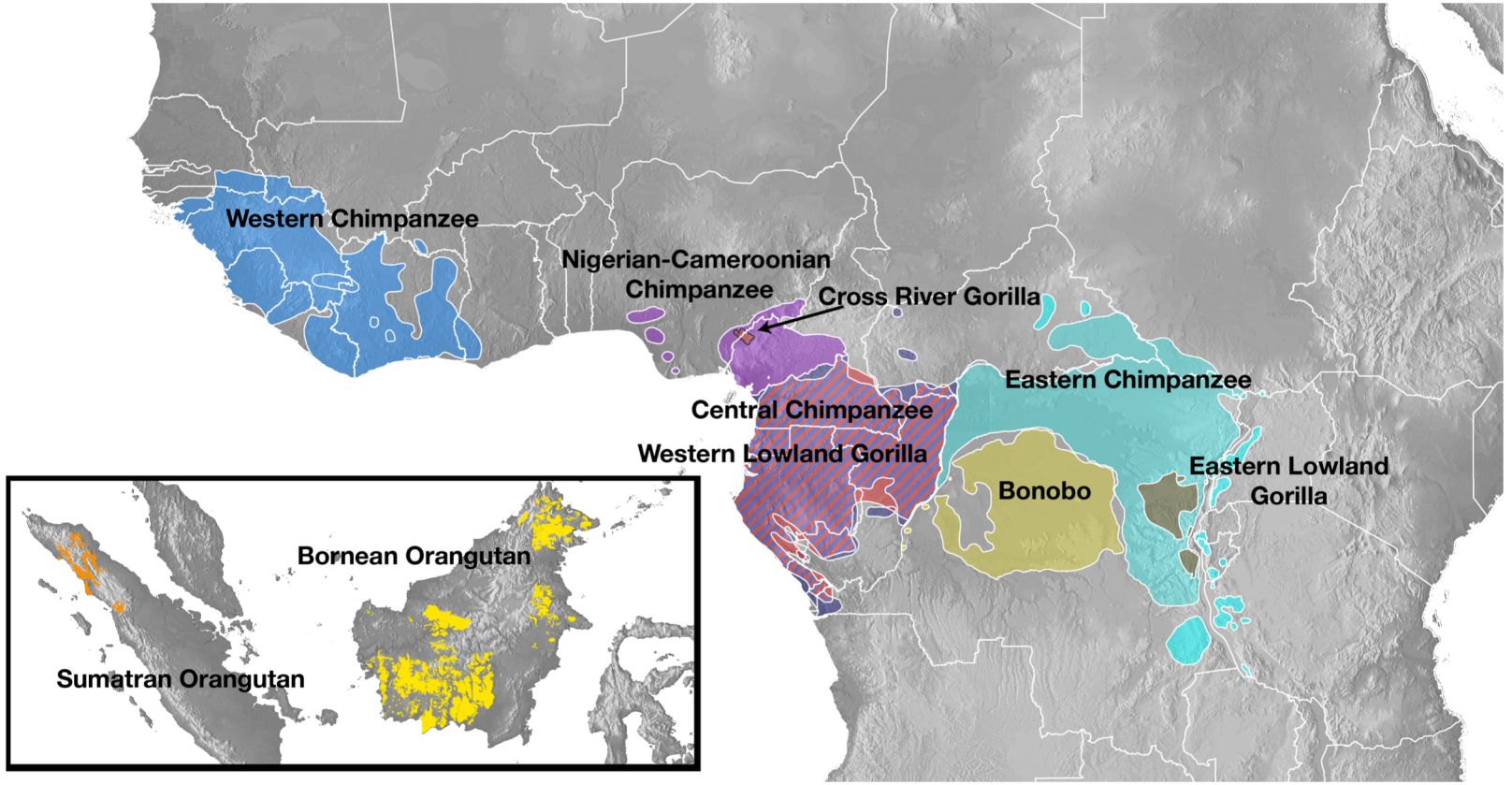
- Get familiar with the commonly used PLINK-format
- Get familiar with extraction of simple summary statistics of data in R
- Get familiar with representation of results
- Be able to interpret diversity measures in populations

POPULATION GENOMICS OF THE COMMON CHIMPANZEE

FROM COMPARATIVE GENOMICS TO POPULATION GENOMICS

6 great ape species
79 complete genomes
~25X average coverage
88.8 million segregating sites

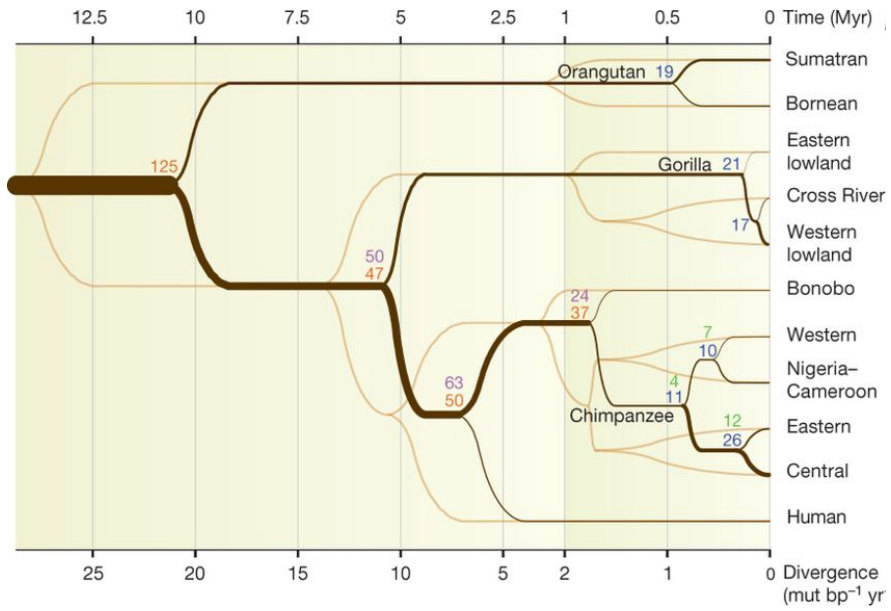
Prado-Martinez et al. 2013, Nautre

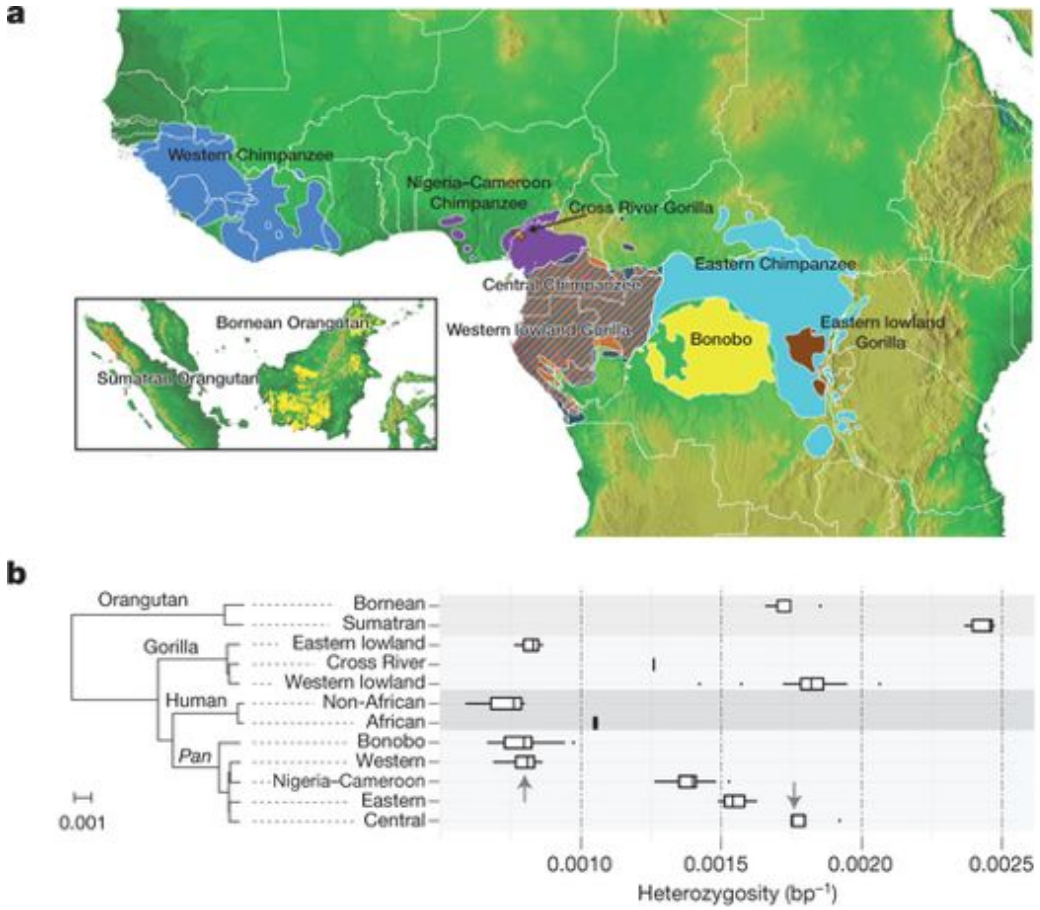


Credit: Peter Sudmant

POPULATION GENOMICS OF THE COMMON CHIMPANZEE

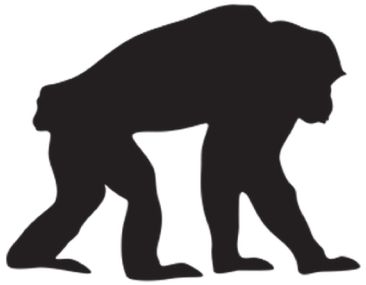
FROM COMPARATIVE GENOMICS TO POPULATION GENOMICS



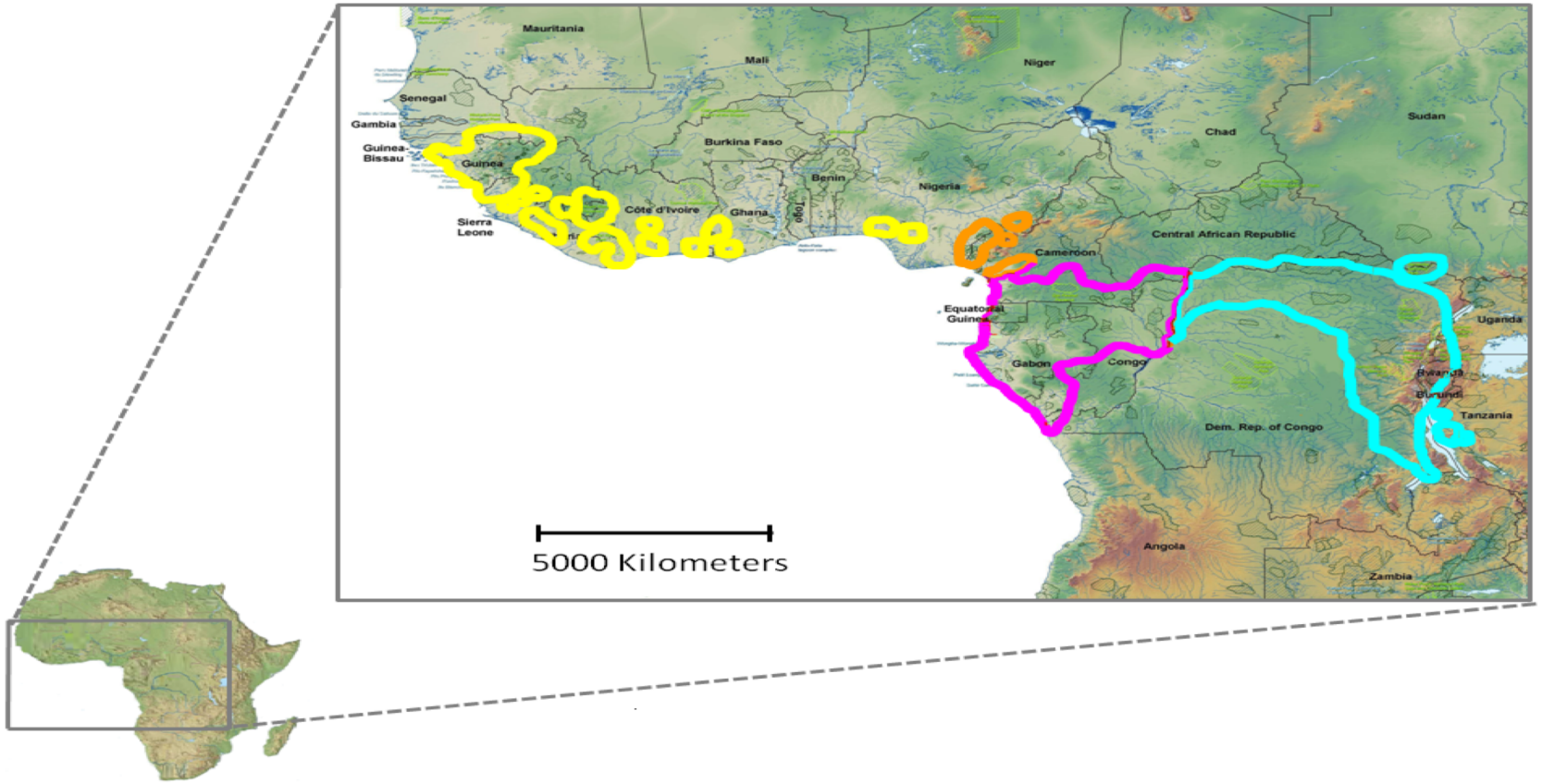


Prado-Martinez et al. 2013, Nautre

POPULATION GENOMICS OF THE COMMON CHIMPANZEE
EXERCISE



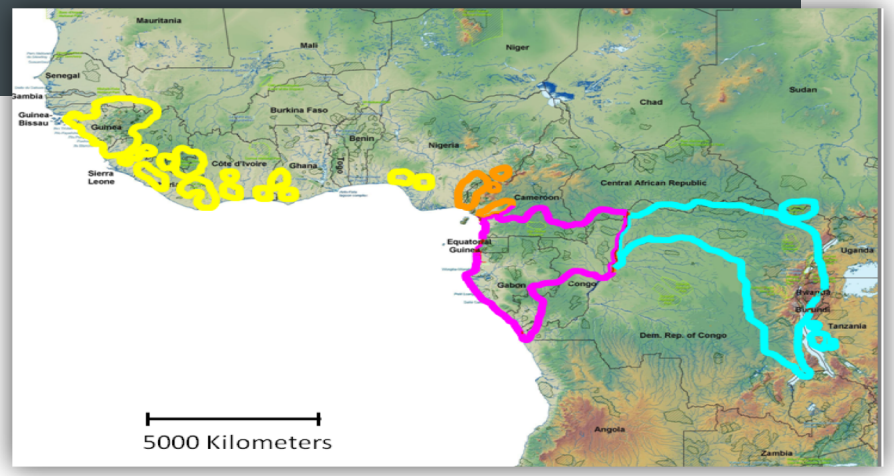
59 complete genomes

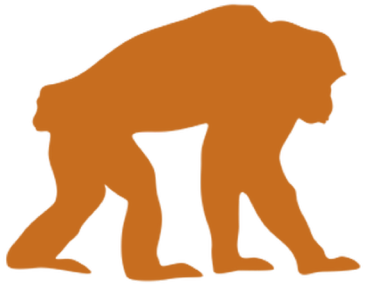


POPULATION GENOMICS OF THE COMMON CHIMPANZEE

4 SUBSPECIES

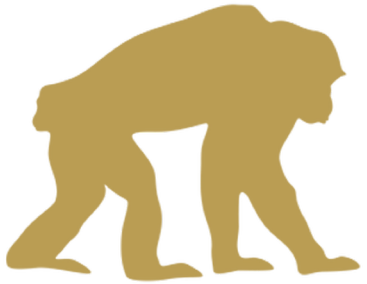
- Subspecies within the common chimpanzee





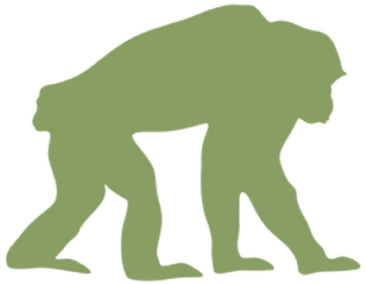
Pan troglodytes verus
Western chimpanzee

12



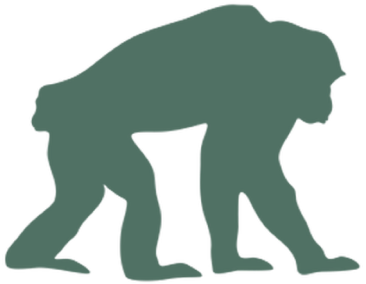
Pan troglodytes ellioti
Nigerian-Cameroon chimp.

10



Pan troglodytes troglodytes
Central chimpanzee

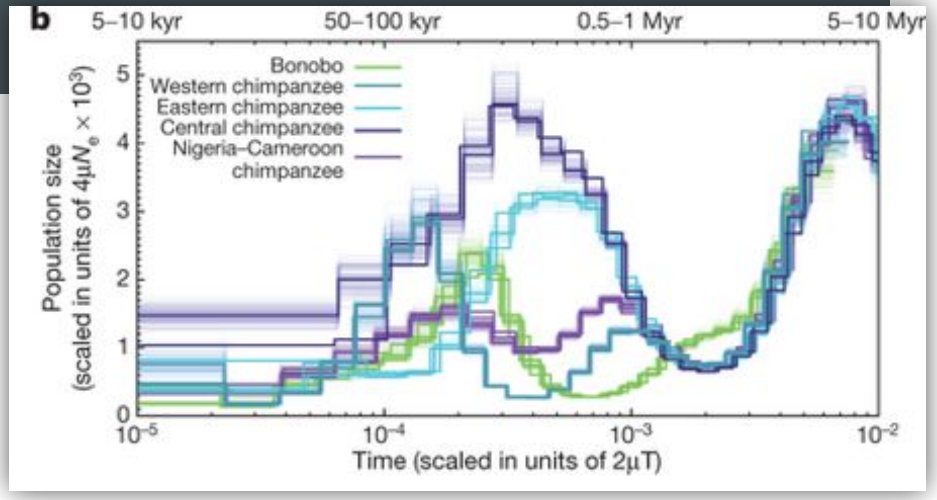
18



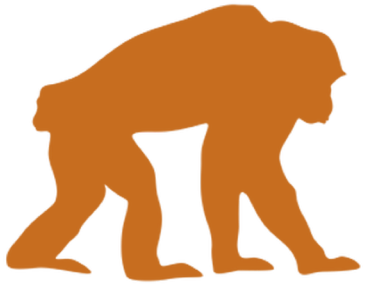
Pan troglodytes schweinfurthii
Eastern chimpanzee

19

POPULATION GENOMICS OF THE COMMON CHIMPANZEE
DEMOGRAPHY

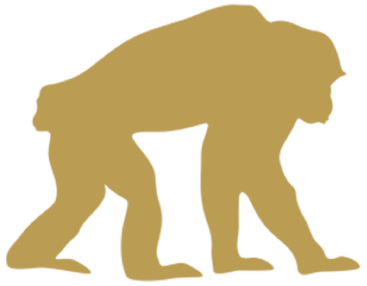


- Subspecies within the common chimpanzee



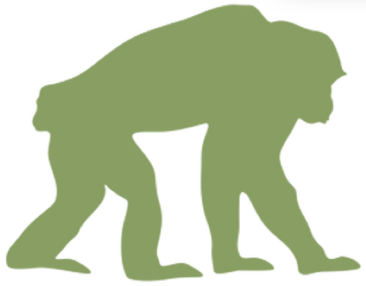
Pan troglodytes verus
Western chimpanzee

12



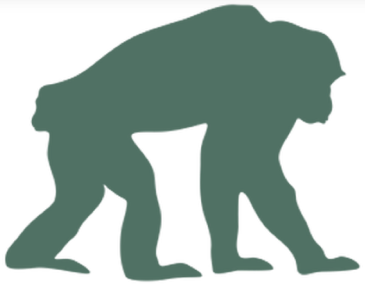
Pan troglodytes ellioti
Nigerian-Cameroon chimp.

10



Pan troglodytes troglodytes
Central chimpanzee

18



Pan troglodytes schweinfurthii
Eastern chimpanzee

19

GETTING STARTED

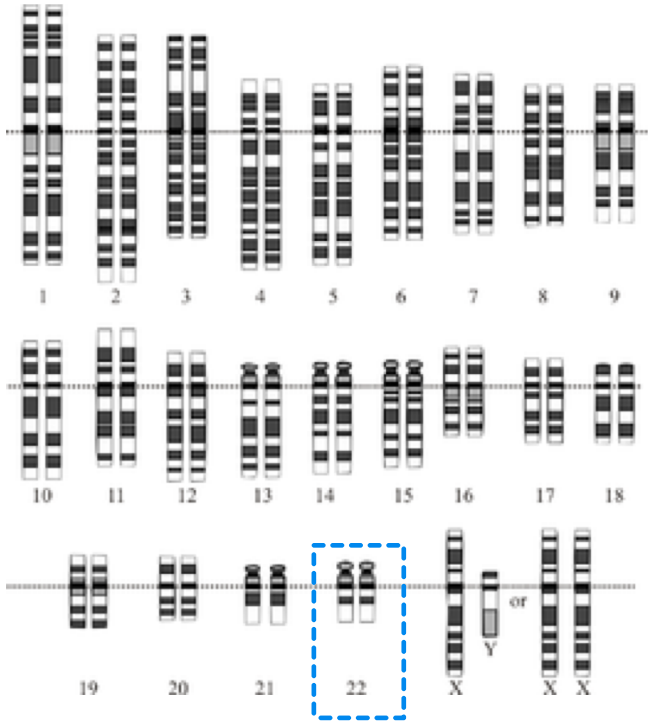
- Make sure to take some time to look at your data
- Get familiar with the file format
- Read through the code
- Interpret your results in a population genetic context

GENETIC DIVERSITY IN CHIMPANZEES

- The datasets
  - consist of the **variable sites found on chromosome 22 in chimpanzees**.
  - contains genotypes from all **four subspecies of chimpanzee and two human populations**.
  - has been filtered to reduce the size of your working data set and includes only SNP's with exactly two different bases (bi-allelic).

GENETIC DIVERSITY IN CHIMPANZEES

- Q1: Before we get started, why do you think we chose chromosome 22?
  - Small enough for us to finish on time.



PLINK FORMAT

- The two main PLINK files are the MAP and the PED files, which often serve as the starting point for any analysis in PLINK.
- Converted from VCF format, which holds all the information about the variant call (e.g. 'read-depth', 'quality').
- The two files are structured slightly different but together, they hold information about each called variant or SNP in each genotyped individual.

PLINK FORMAT

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

PED files

The fields in a PED file are

- Family ID
- Sample ID
- Paternal ID
- Maternal ID
- Sex (1=male; 2=female; other=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

PLINK FORMAT

- Change to the right directory
- Copy your data
- Look at the PLINK files

```
less FILENAME.map
q
less FILENAME.ped
q
```

PLINK FORMAT

- 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.

ESTIMATING GENETIC DIVERSITY

- We can estimate the diversity as the **expected heterozygosity**, He = 2p(1 - p), assuming HWP.
- We will use PLINK and R to calculate the allele frequency of the minor allele for each variable site.
- For example, for Pan troglodytes:

```
plink --noweb --file Pan_troglodytes --freq --out Pan_troglodytes

cat Pan_troglodytes.frq |grep -v NA > Pan_troglodytes_noNA.frq
```

ESTIMATING GENETIC DIVERSITY

- 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

ESTIMATING GENETIC DIVERSITY

- Open R and calculate pi for all populations
- For example *verus:*
- # Read in each of the frequency files
```
verus<-read.table("Pt_verus_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,]
```
- # 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"])))
```

- # 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)))
```

ESTIMATING GENETIC DIVERSITY

- # Making a barplot with the nucleotide diversity

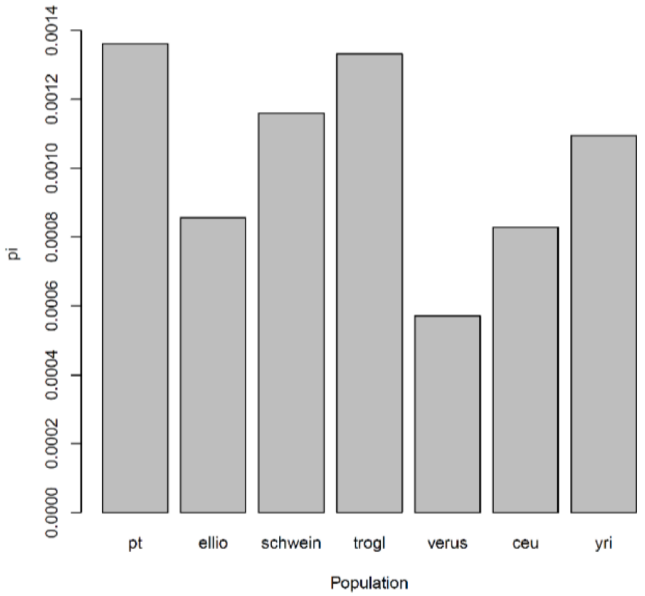
```
par(mfrow=c(1,1))
val = c(mean(pt$pi), mean(ellio$pi), mean(schwein$pi), mean(troglo$pi),mean(verus$pi),
mean(ceu$pi), mean(yri$pi)) #
barplot(val,ylim=c(0.000,0.0015), ylab="pi",
xlab="Population",names.arg=c("pt","ellio","schwein","trogl","verus","ceu","yri"))
```

- # To shut down the plot window

```
dev.off()
```

ESTIMATING GENETIC DIVERSITY

- 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))
  - > mean(val)
    [1] 0.001028742



- 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.

ESTIMATING GENETIC DIVERSITY

- 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 Ne? 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

- ## 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))
}
```

ESTIMATING THE NUCLEOTIDE DIVERSITY ALONG THE CHROMOSOME

- ## 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")
par(mfrow=c(2,2))
windowsize<- 3000
steps<- 100
```

- For example, verus:

```
# 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)
```

- # To close the graphical window

```
dev.off()
```

ESTIMATING THE NUCLEOTIDE DIVERSITY ALONG THE CHROMOSOME

- Leave R

```
q()
n
```

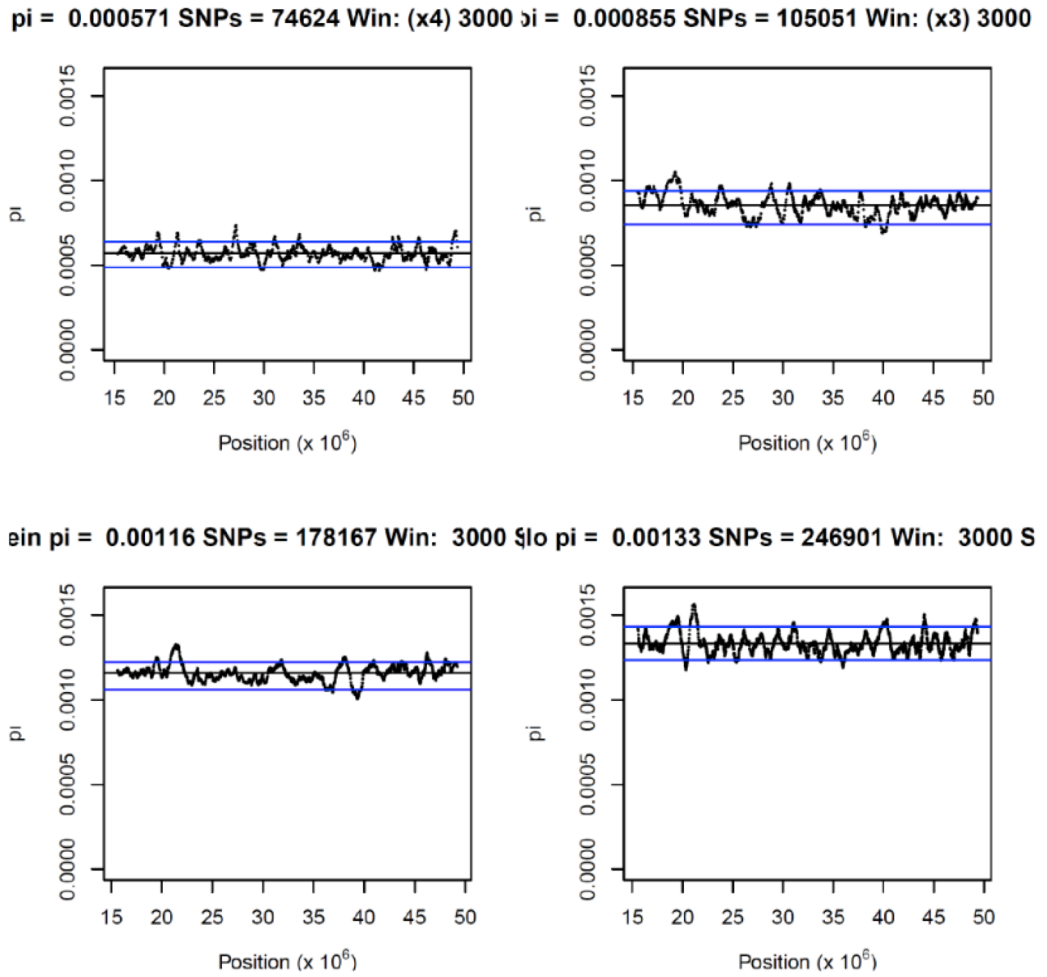
- Take a look at the output

```
evince nucleotide_diversity_in_4_subspecies.pdf
```

ESTIMATING THE NUCLEOTIDE DIVERSITY ALONG THE CHROMOSOME

- Take a look at the output

```
evince nucleotide_diversity_in_4_subspecies.pdf
```



ESTIMATING THE NUCLEOTIDE DIVERSITY ALONG THE CHROMOSOME

- 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

- Estimate the individual inbreeding coefficient for all individuals in the different populations, e.g.

```
plink --file Pt_verus --het --out Pt_verus
```

- 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.
- 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

ESTIMATING INBREEDING COEFFICIENT PR. INDIVIDUAL

- 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 F of 0.11 and 0.062
    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? Now take a look at the total sample of the combined set of chimpanzees ("Pan_troglodytes.het").
  - 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.
- 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.

SUMMARY

- PLINK format
- Summary statistics in R

- Practical application of plink formatted data
- Graphical representation of the data and biological interpretation of results