GENETIC DIVERSITY

POPULATION GENETICS 2022

GENÍS GARCIA ERILL
ADAPTED FROM SLIDES BY **PATRÍCIA CHRZANOVÁ PEČNEROVÁ**

PHOTO CREDIT : MOGENS TROLLE

PROGRAM

- Examine the PLINK-format, and use PLINK to do some basic data manipulation and summary statistics
- 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

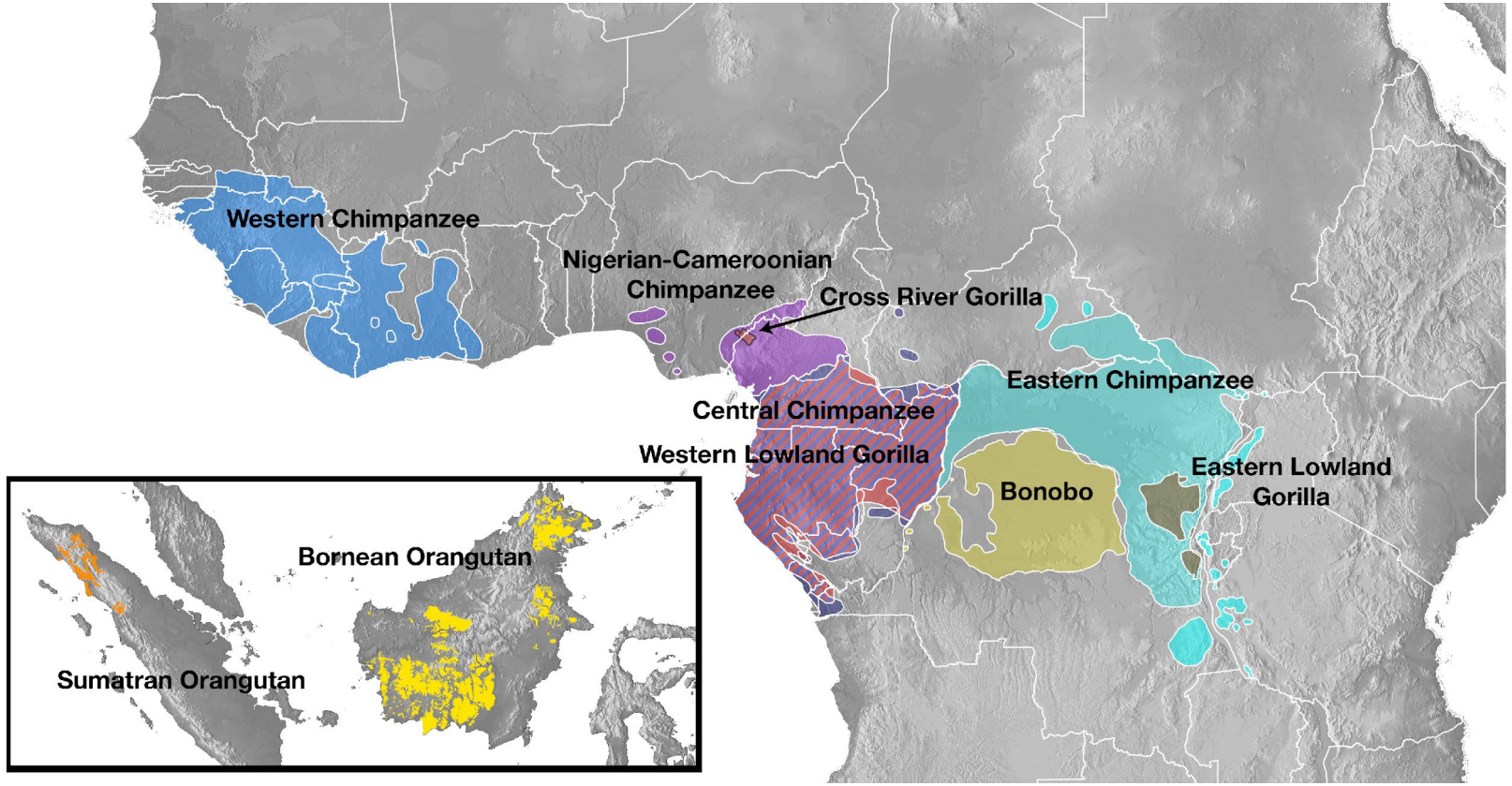
- Get familiar with the PLINK format and usage
- Get familiar with manipulating data, extraction of summary statistics and plotting in R
- Be able to estimate and interpret genetic diversity measures in populations

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

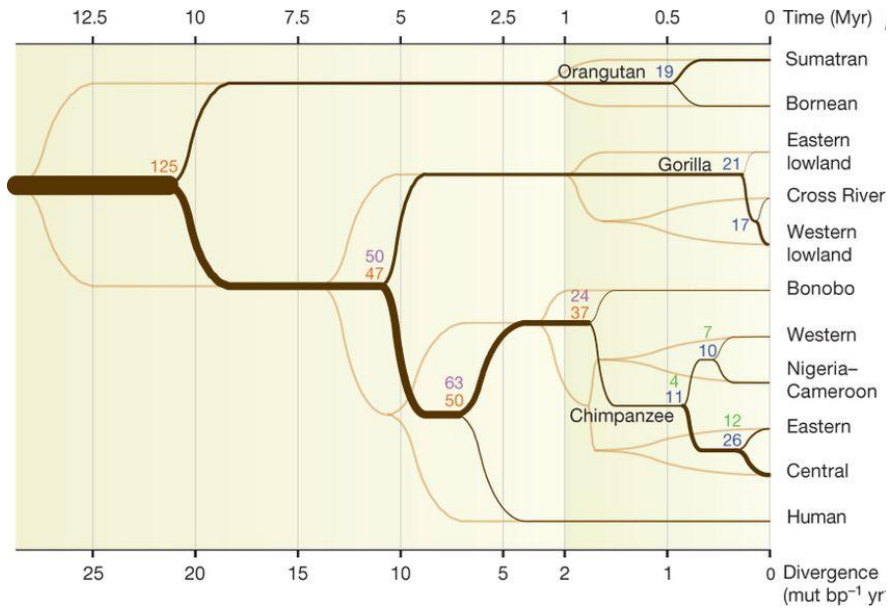
Prado-Martinez et al. 2013, Nautre

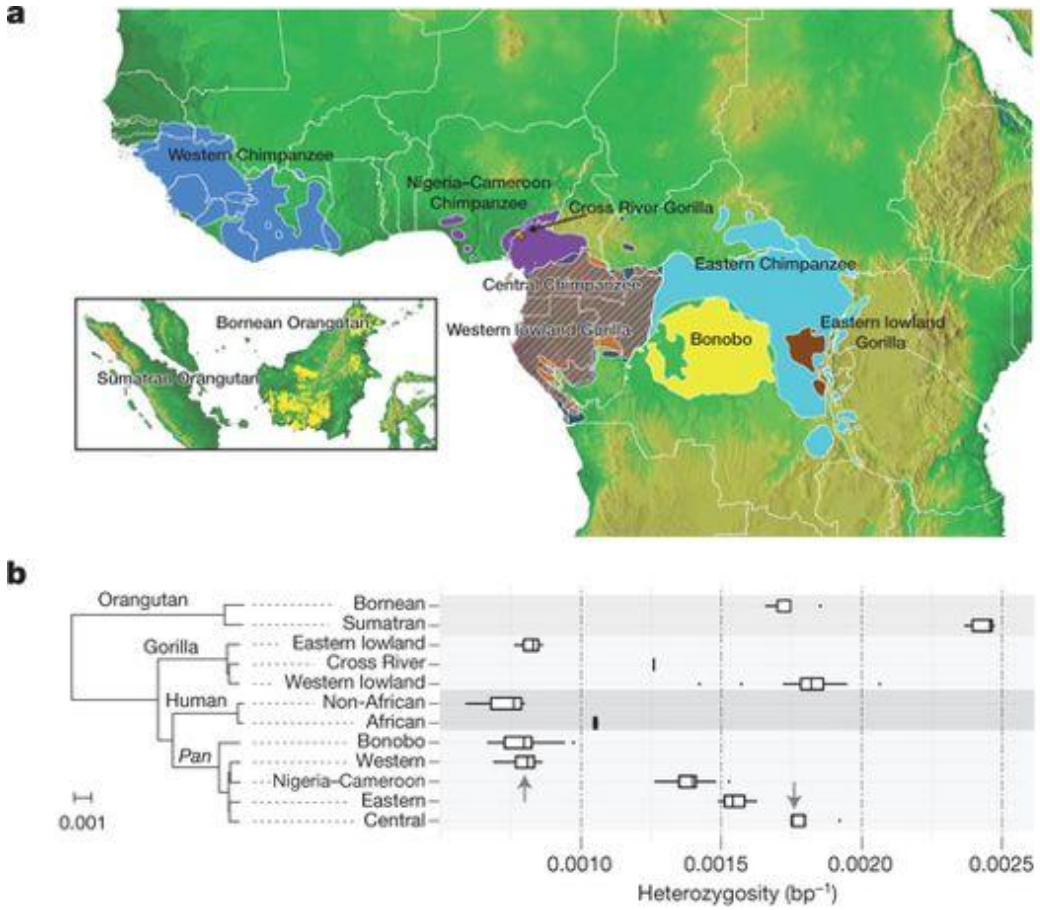


Credit: Peter Sudmant

POPULATION GENOMICS OF THE COMMON CHIMPANZEE

FROM COMPARATIVE GENOMICS TO POPULATION GENOMICS

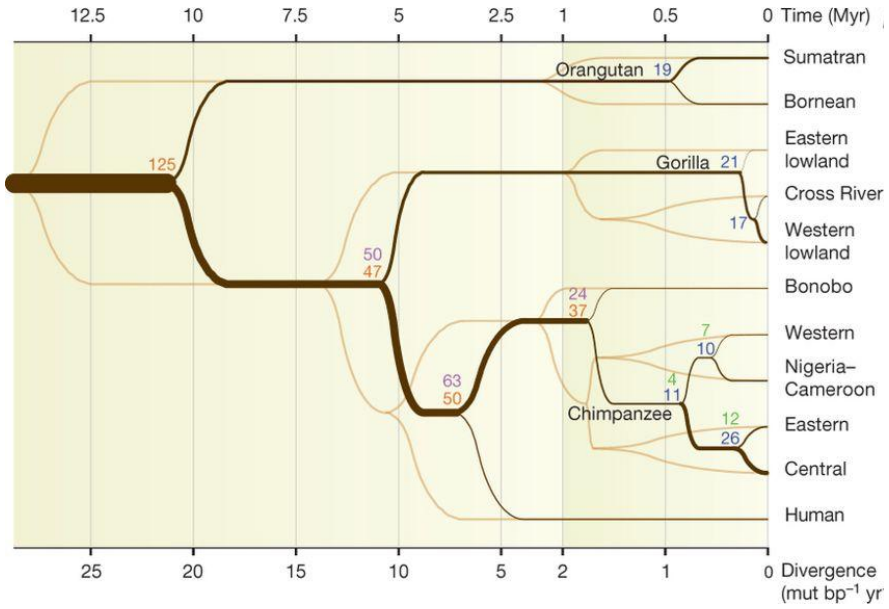


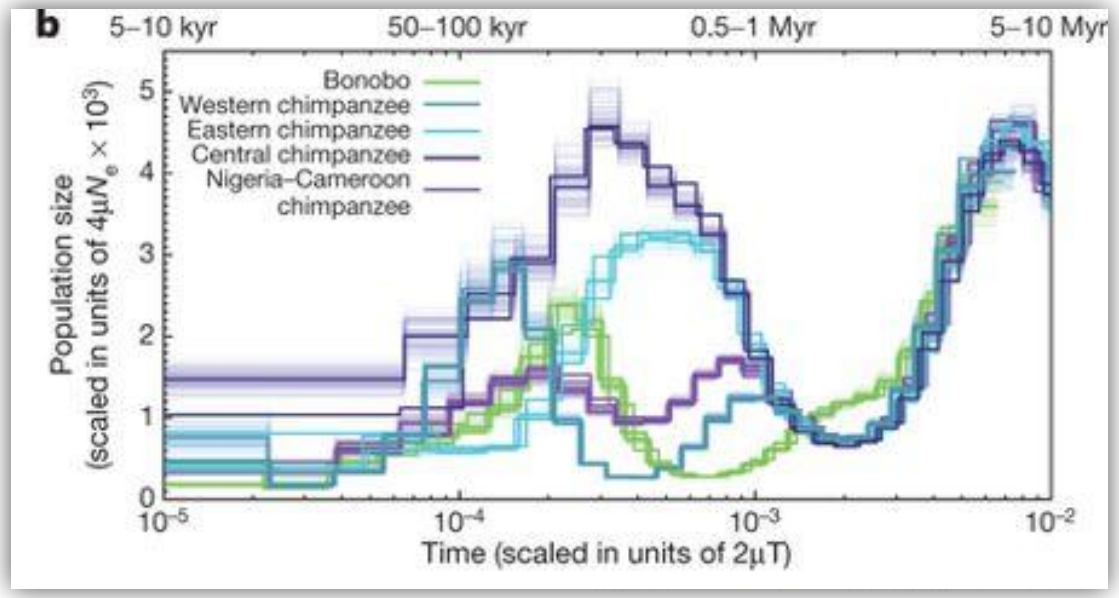


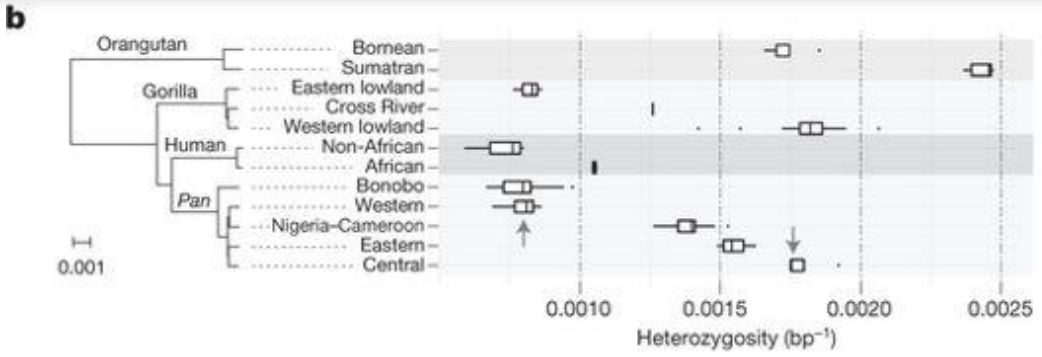
Prado-Martinez et al. 2013, Nautre

POPULATION GENOMICS OF THE COMMON CHIMPANZEE

FROM COMPARATIVE GENOMICS TO POPULATION GENOMICS

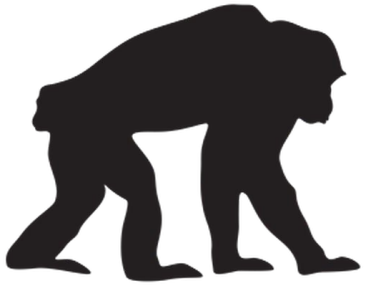




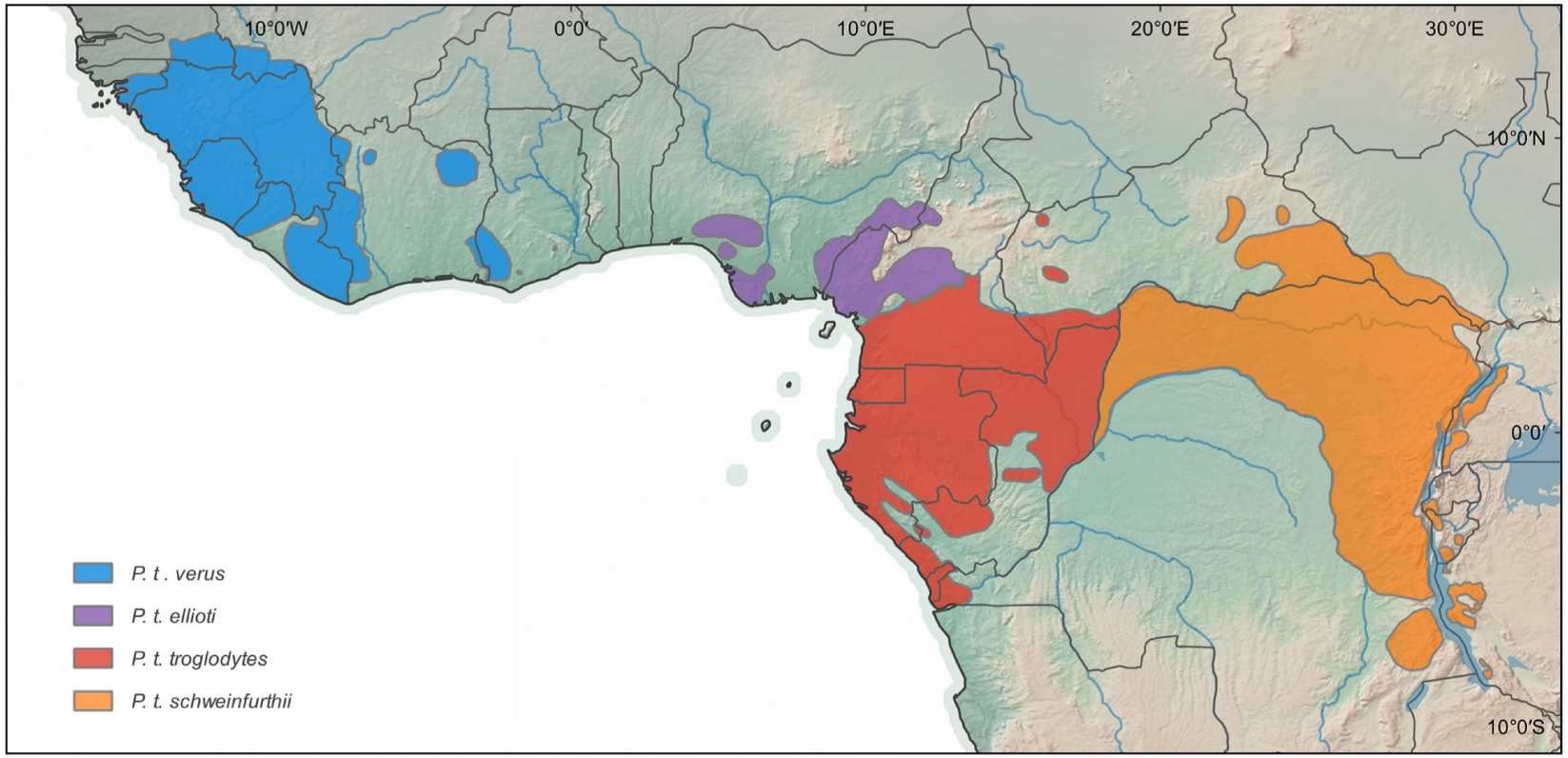


Prado-Martinez et al. 2013, Nautre

POPULATION GENOMICS OF THE COMMON CHIMPANZEE
EXERCISE



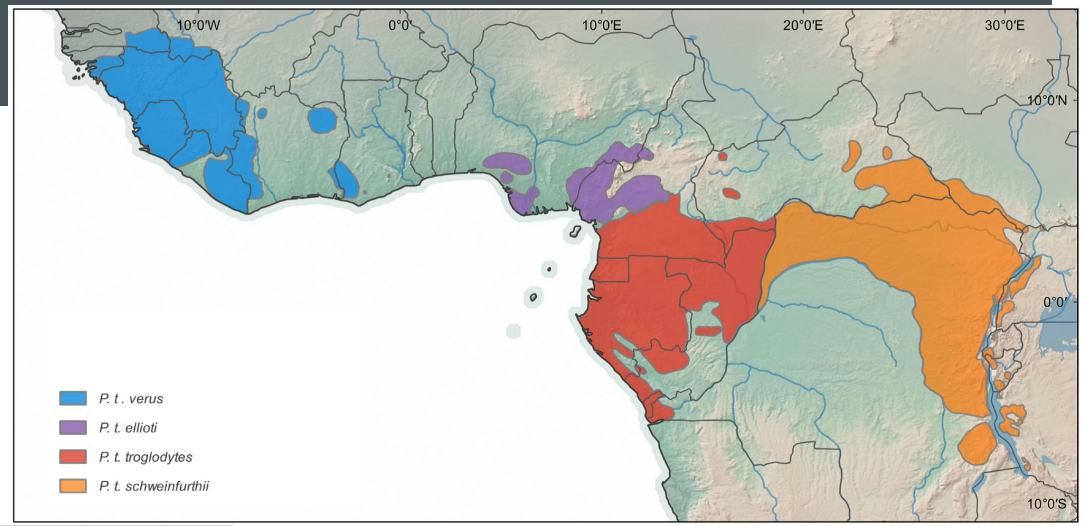
59 complete genomes

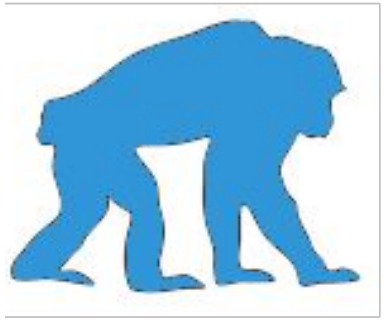


POPULATION GENOMICS OF THE COMMON CHIMPANZEE

4 SUBSPECIES

- Subspecies within the common chimpanzee



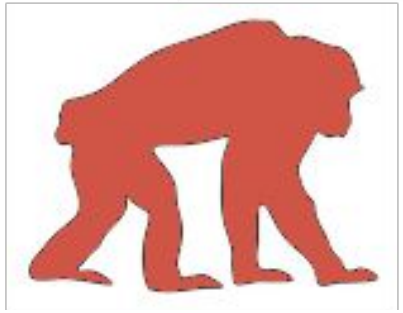


Pan troglodytes verus
Western chimpanzee

12

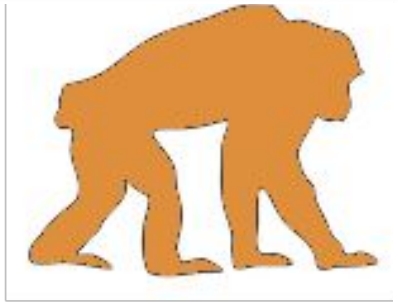
Pan troglodytes ellioti
Nigerian-Cameroon chimpanzee

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).

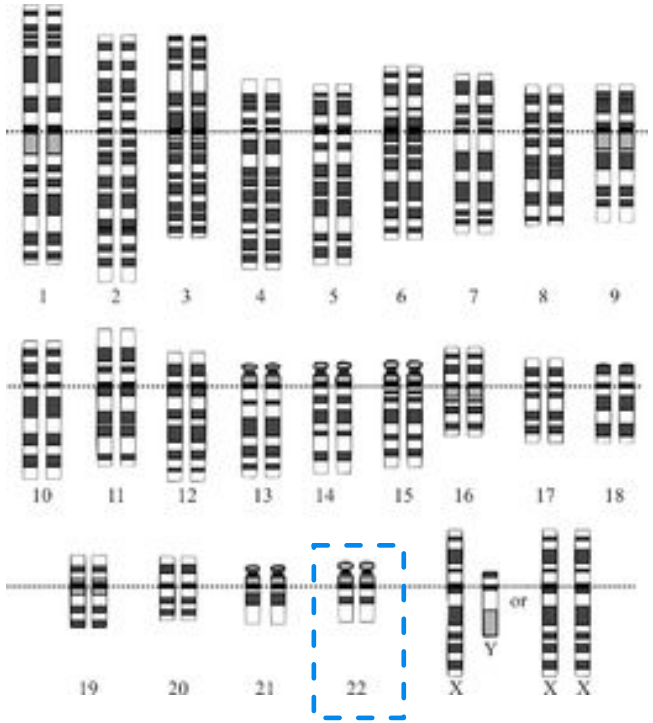
GENETIC DIVERSITY IN CHIMPANZEES

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

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 three main files in the binary PLINK format are the **.bed**, **.bim** and **.fam**
- Converted from VCF (Variant Call Format), which holds all the information about the variant call (e.g. 'read-depth', 'quality')
- The three files are complementary and, together, they hold information about each called variant in each genotyped individual.

PLINK FORMAT

- **.bed** file: it contains the genotype matrix in binary format; specifies what genotype each individual has at each variant.
- **.bim** file: it contains information for each of the variants (position in bp and in Morgans, ID, alleles…)
- **.fam** file: it contains information for each of the samples (individual and family ID, and it can also conain information about sex and phenotype)

PLINK FORMAT

- Change to the right directory
- Copy your data
- Look at the PLINK files (for example, the chimpanzee files with prefix pan_troglodytes):

```
less pan_troglodytes.bed
q
less pan_troglodytes.bim
q
less pan_troglodytes.fam
q
```

PLINK FORMAT

- **Q2**: What do you see when you open the .bed file? Why is that?
- **Q3**: What is the position of the first SNP? (Confer with link above about file format)
- **Q3** What information is in the .bim and .fam files)? (Again, look at the file and confer with the link above.)
- **Q5**: 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 .bim or the .fam file?
- **Q6**: 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?

PLINK FORMAT

- **Q2**: What do you see when you open the .bed file? Why is that?

  You should see some illegible characters, because it is a binary file made to be easily read by a computer but not so easily by humans.

- **Q3**: What is the position of the first SNP? (Confer with link above about file format)

  Chromosome 22 bp coordinate 14438985 (first line, 1st and 4th column of .bim file)

- **Q3** What information is in the .bim and .fam files)? (Again, look at the file and confer with the link above.)

  Variant information (chromosome, ID, position and alleles...) in .bim and sample information (ID, family ID, sex...) in .fam.

PLINK FORMAT

- **Q5**: 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 .bim or the .fam file?

  Chimpanzees: 369471

  Human: 154499

- **Q6**: 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 reduces the certainty of the SNP calls, often you exclude SNPs based on sites where you only have a limited read depth. In general low depth will lead to identifying less SNPs.

ESTIMATING GENETIC DIVERSITY

- We will estimate the genetic diversity as the **expected heterozygosity**,

$$He = 2p(1 - p)$$

- To do so, we need to assume the genotypes are in Hardy Weinberg proportions.

- **Q7**: Before starting, consider the two datasets we have. Is it a good idea to estimate the expected heterozygosity based on the combined datasets for each species? (Hint: look at the sample overview in Table 1 and at the Family ID column (first column) in the .fam files).

ESTIMATING GENETIC DIVERSITY

- We will estimate the genetic diversity as the **expected heterozygosity**,

$$He = 2p(1 - p)$$

- To do so, we need to assume the genotypes are in Hardy Weinberg proportions.
- **Q7**: Before starting, consider the two datasets we have. Is it a good idea to estimate the expected heterozygosity based on the combined datasets for each species? (Hint: look at the sample overview in Table 1 and at the Family ID column (first column) in the .fam files).

  It is not a good idea, to estimate expected heterozygosity we need to assume the genotypes are in Hardy Weinberg proportions, and when we combine genotype data from different populations we do not expect them to be in Hardy Weinberg proportions.

ESTIMATING GENETIC DIVERSITY

- We will first split each the two plink files by populations (notice that the first column of the .fam file indicates the population each sample belongs to).
- We will use the following plink commands (as example, this one splits the individuals from the verus subspecies from the pam_troglodytes file; keep.txt would a text file with "P.t.verus" as content)

```
plink --bfile pan_troglodytes --keep-fam keep.txt --mac 1 --geno 0.5 --make-bed --out verus
```

- **Q8**: Try to understand the plink command, what are the different options doing? (Plink has a very extensive documentation where you can find all the commands. You can ignore the bash specific syntax but feel free to ask an instructor if you are interested).

ESTIMATING GENETIC DIVERSITY

```
plink --bfile pan_troglodytes --keep-fam keep.txt --mac 1 --geno 0.5 --make-bed --out verus
```

- **Q8**: Try to understand the plink command, what are the different options doing? (Plink has a very extensive documentation where you can find all the commands. You can ignore the bash specific syntax but feel free to ask an instructor if you are interested).

  --bfile indicate the input file we want to load.

  --keep-fam takes a filename with the family ID/s we want to keep

  --mac 1 mean minimum allele count 1, it will filter sites that are not variable in a given subspecies

  --geno 0.5 will remove variants where more than half of the individuals have missing data

  --make-bed will create a new plink binary file after filtering individuals and variants specified with the previous options

  --out will be the prefix for the new .bed .bim and .fam files

ESTIMATING GENETIC DIVERSITY

- **Q9**: Look at the output printed to the screen for the last plink command (should be the one that generate the verus plink file) and try to find out how many variant were removed due to missing data and minimum minor allele count, respectively (you can find the same information in the verus.log file).

ESTIMATING GENETIC DIVERSITY

- **Q9**: Look at the output printed to the screen for the last plink command (should be the one that generate the verus plink file) and try to find out how many variant where removed due to missing data and minimum minor allele count, respectively (you can find the same information in the verus.log file)

```
PLINK v1.90p 64-bit (9 Jan 2018)              www.cog-genomics.org/plink/1.9/
(C) 2005-2018 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to verus.log.
Options in effect:
  --bfile pan_troglodytes
  --geno 0.5
  --keep-fam keep.txt
  --mac 1
  --make-bed
  --out verus

48120 MB RAM detected; reserving 24060 MB for main workspace.
369471 variants loaded from .bim file.
59 people (0 males, 0 females, 59 ambiguous) loaded from .fam.
Ambiguous sex IDs written to verus.nosex .
--keep-fam: 12 people remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 12 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.927107.
12 variants removed due to missing genotype data (--geno).
301258 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
68201 variants and 12 people pass filters and QC.
Note: No phenotypes present.
--make-bed to verus.bed + verus.bim + verus.fam ... done.
```

ESTIMATING GENETIC DIVERSITY

- **Q9**: Look at the output printed to the screen for the last plink command (should be the one that generate the verus plink file) and try to find out how many variant where removed due to missing data and minimum minor allele count, respectively (you can find the same information in the verus.log file).

12 variants removed due to missing genotype data (--geno) and 301258 variants removed due to minor allele threshold.

```
PLINK v1.90p 64-bit (9 Jan 2018)              www.cog-genomics.org/plink/1.9/
(C) 2005-2018 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to verus.log.
Options in effect:
  --bfile pan_troglodytes
  --geno 0.5
  --keep-fam keep.txt
  --mac 1
  --make-bed
  --out verus

48120 MB RAM detected; reserving 24060 MB for main workspace.
369471 variants loaded from .bim file.
59 people (0 males, 0 females, 59 ambiguous) loaded from .fam.
Ambiguous sex IDs written to verus.nosex .
--keep-fam: 12 people remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 12 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.927107.
12 variants removed due to missing genotype data (--geno).
301258 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
68201 variants and 12 people pass filters and QC.
Note: No phenotypes present.
--make-bed to verus.bed + verus.bim + verus.fam ... done.
```

ESTIMATING GENETIC DIVERSITY

- Now we are finally ready to start estimating genetic diversity as expected heterozygosity, separately for each of the populations we have just generated files for.
- We will use the following plink command for each population (here, with the troglodytes subspecies as example):

```
plink --bfile troglodytes --freq --out troglodytes
```

- **Q10**: Try and look in the troglodytes.frq file, what information do you get?

ESTIMATING GENETIC DIVERSITY

- Now we are finally ready to start estimating genetic diversity as expected heterozygosity, separately for each of the populations we have just generated files for.

- We will use the following plink command for each population (here, with the troglodytes subspecies as example):

  plink --bfile troglodytes --freq --out troglodytes

- **Q10**: Try and look in the troglodytes.frq file, what information do you get?

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

ESTIMATING GENETIC DIVERSITY

- Open R and calculate expected heterozygosity for all populations
- For example *verus:*
- # Read in each of the frequency files
  ```
  verus <- read.table("verus.frq", h=T)
  ```
- # Function for estimating the expected heterozygosity
  ```
  het<-function(x){2*x*(1-x)}
  ```
- # estimate expected heterozygosity for each variant and add as column
  verus$het <- het(verus$MAF)

ESTIMATING GENETIC DIVERSITY

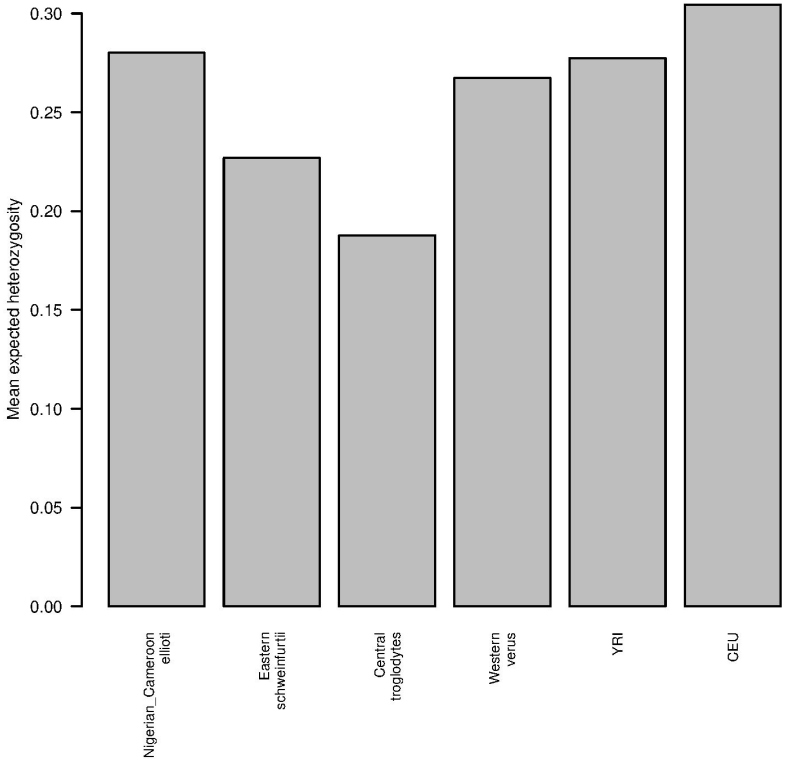
- # Get mean expected heterozygosity across all variants across populations
  mean_hets <- sapply(list(ellioti$het, schwein$het, troglo$het, verus$het, yrihet, ceuhet), mean)

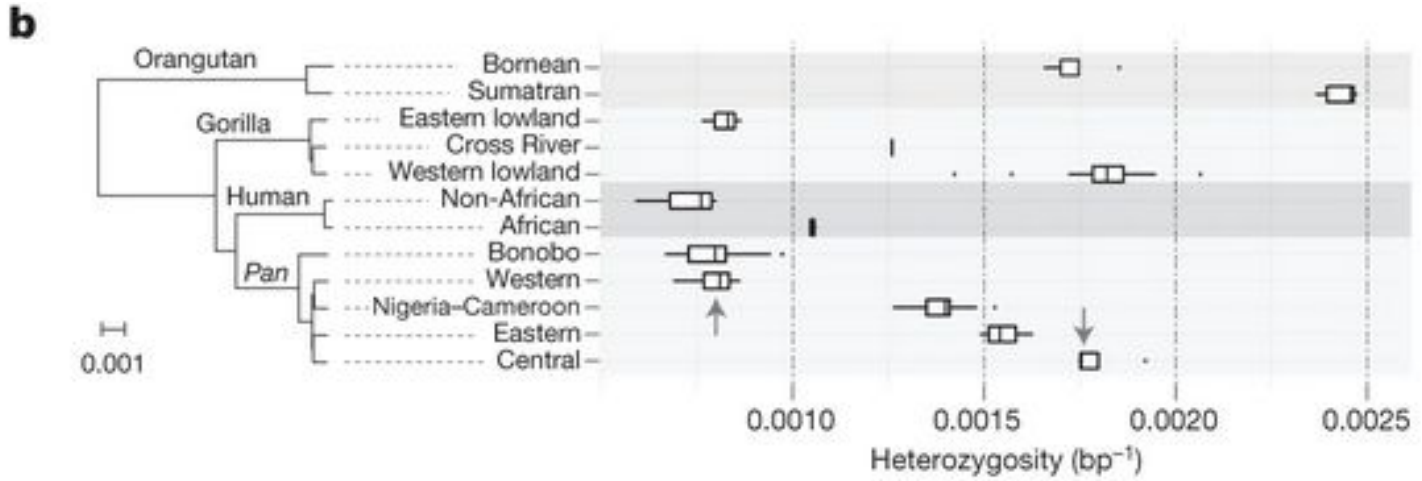
- # plot mean expected heterozygosity across variants for each population

```
par(mar=c(7,4,4,2))
barplot(mean_hets, names.arg=c("Nigerian_Camerron", "Eastern", "Central", "Western", "YRI", "CEU"),
las=2, cex.names=0.8)
```

ESTIMATING GENETIC DIVERSITY

- **Q11**: In the R function (het), explain what 2*x*(1-x) calculates
- **Q12**: Compare the expected heterozygosity we have estimated and plotted with the heterozygosity estimates from Figure 2. Can you explain why ours are much higher? (Hint: the chromosome 22 has around 55.000.000 bp; how many variants did we have in the plink files?)





ESTIMATING GENETIC DIVERSITY

- **Q11**: In the R function (het), explain what 2*x*(1-x) calculates

  Calculates expected heterozygosity based on population allele frequencies.

- **Q12**: Compare the expected heterozygosity we have estimated and plotted with the heterozygosity estimates from Figure 2. Can you explain why ours are much higher? (Hint: the chromosome 22 has around 55.000.000 bp; how many variants did we have in the plink files?)

  The expected heterozygosity we have estimated is based only in variable sites, which will increase a lot the proportion of heterozygous and will necessarily be representative of the genetic diversity in the population. The estimates in Prado-Martinez et al. are proportion of heterozygosity per bp, including positions that are equal across all samples which will always be homozygous and are the majority of the genome.

ESTIMATING GENETIC DIVERSITY

- We can get a proper estimate of the nucleotide diversity that takes into account fixed positions, by looking at the range of positions we have data for and assuming all positions we do not observe in our plink file are homozygous.

- # first we will read in the bim files, to know the position of each variant. we continue using ellioti as example
  elliotiBim <- read.table("ellioti.bim", h=F)

- # add the position to the tables with expected heterozygosity
  ellioti$pos <- elliotiBim$V4

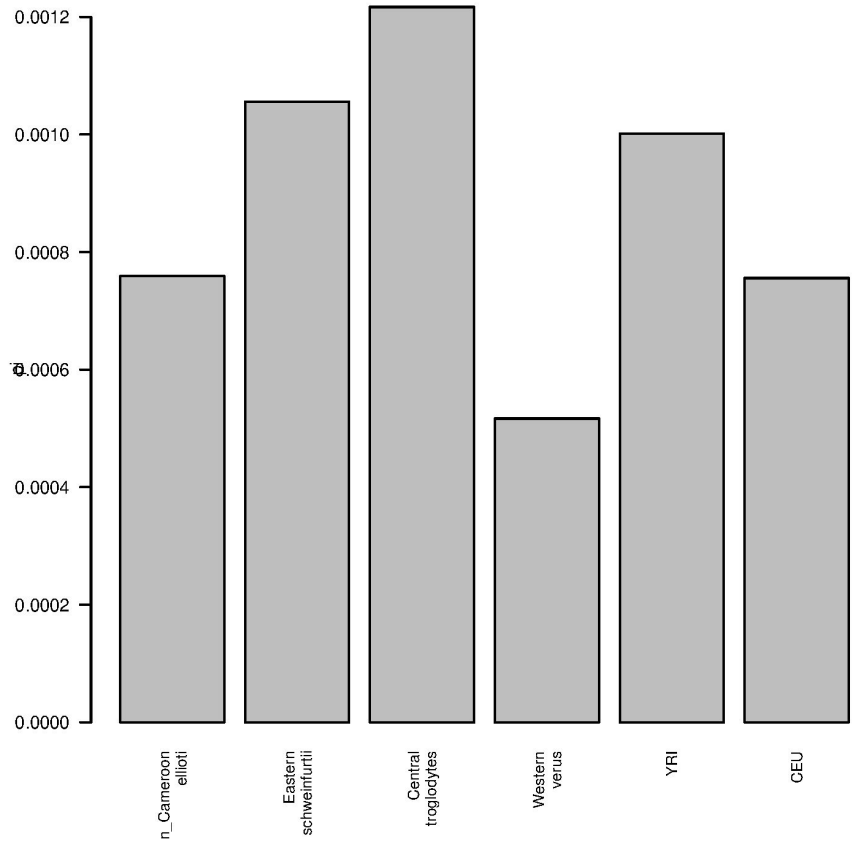
- # function to estimate pi using a rough estimate of number of base pairs
- # we have data from as difference between last and fist positon

  getPi <- function(x) sum(x$het) / (x$pos[nrow(x)] - x$pos[1])

- # once we have all populations, estimate pi for all and plot
  all_pi <- sapply(list(ellioti, schwein, troglo, verus, yri, ceu), getPi)
  names(all_pi) <- c("ellioti", "schwein", "troglo", "verus", "yri", "ceu")
  par(mar=c(7,5,4,2))
  barplot(all_pi, names.arg=c("Nigerian_Cameroon\nellioti", "Eastern\nschweinfurtii", "Central\ntroglodytes", "Western\nverus", "YRI", "CEU"), las=2, cex.names=0.8)

ESTIMATING GENETIC DIVERSITY

- **Q13**: Why do you think the two human populations differ in heterozygosity?
- **Q14** Will rare variants more often be found in heterozygous state or homozygous?
- **Q15**: From your knowledge and from the amount of average heterozygosity, what population would you expect to have the highest Ne? And the lowest?

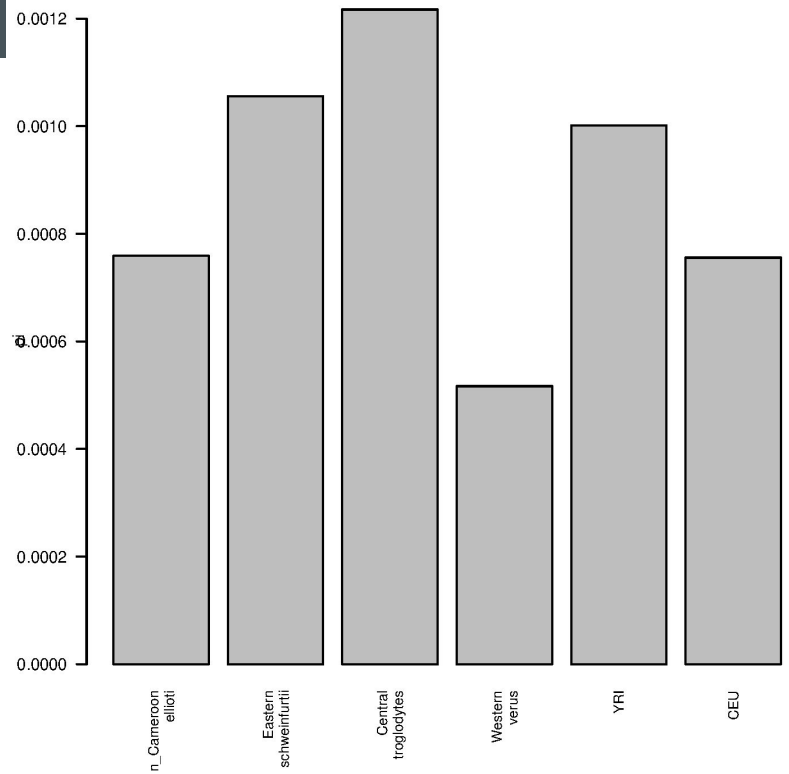


ESTIMATING GENETIC DIVERSITY

- **Q13**: Why do you think the two human populations differ in heterozygosity?
  Differences in demographic history. Europeans have gone through a population bottleneck, reducing their effective population size.

- **Q14** Will rare variants more often be found in heterozygous state or homozygous?
  Rare variants will mostly be found in heterozygous state, since it is very unlikely that a low frequency variants is paired together with itself. Based on HWE proportions, the probability of finding a variant with frequency 0.01 in heterozyougs state 2 * 0.01 * 0.99 = 0.0198 while in homozyougs is 0.01 * 0.01 = 0.0001.

- **Q15**: 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 estimating and plotting pi in sliding windows

```
slidingwindowPiplot <- function(mainv, xlabv, ylabv, ylimv=NULL, 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/2)]
                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/2)]
                means_y[i] <-  sum(chunk_y,na.rem=TRUE)/dist(range(chunk_x))
        }

        plot(means_x,means_y,type="b",main=mainv,xlab=xlabv,ylab=ylabv,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.
- # funciton to define window size as a funciton of the number of snps
- # so all populaitons have windows of equal size in bp
  winsize <- function(nsnp, nwin=75){round(nsnp/nwin/100) * 100}
  steps<- 100

- # do multipanel plot (6 plots arranged in 3 rows, 2 columns)
  par(mfrow=c(3,2))

- # Plot all populations, for example, verus:

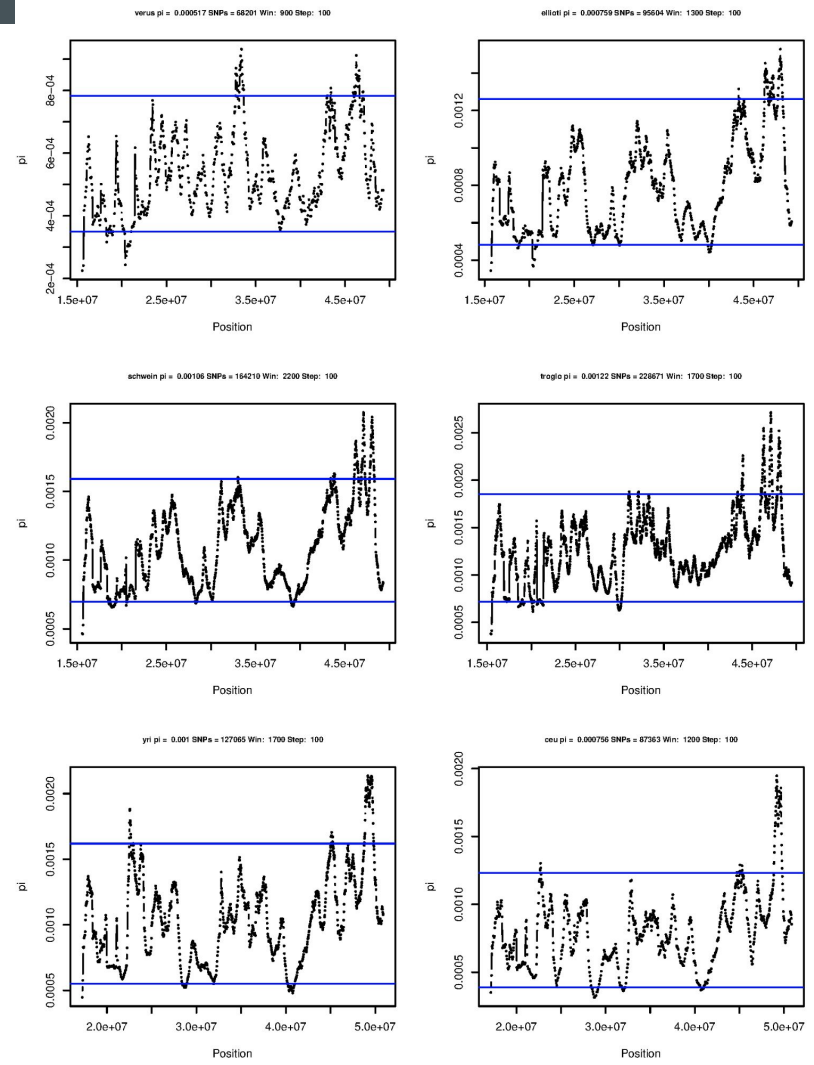
  windowsize <- winsize(nrow(verus))

  mainvv = paste("verus pi = ",format(all_pi["verus"], digits=3), "SNPs =", nrow(verus), "Win: ", windowsize, "Step: ", steps)

  slidingwindowPiplot(mainv=mainvv, xlab="Position", ylab=expression(paste("pi")), window.size=windowsize, step.size=steps, input_x_data=verus$pos,input_y_data=verus$het)

ESTIMATING THE NUCLEOTIDE DIVERSITY ALONG THE CHROMOSOME

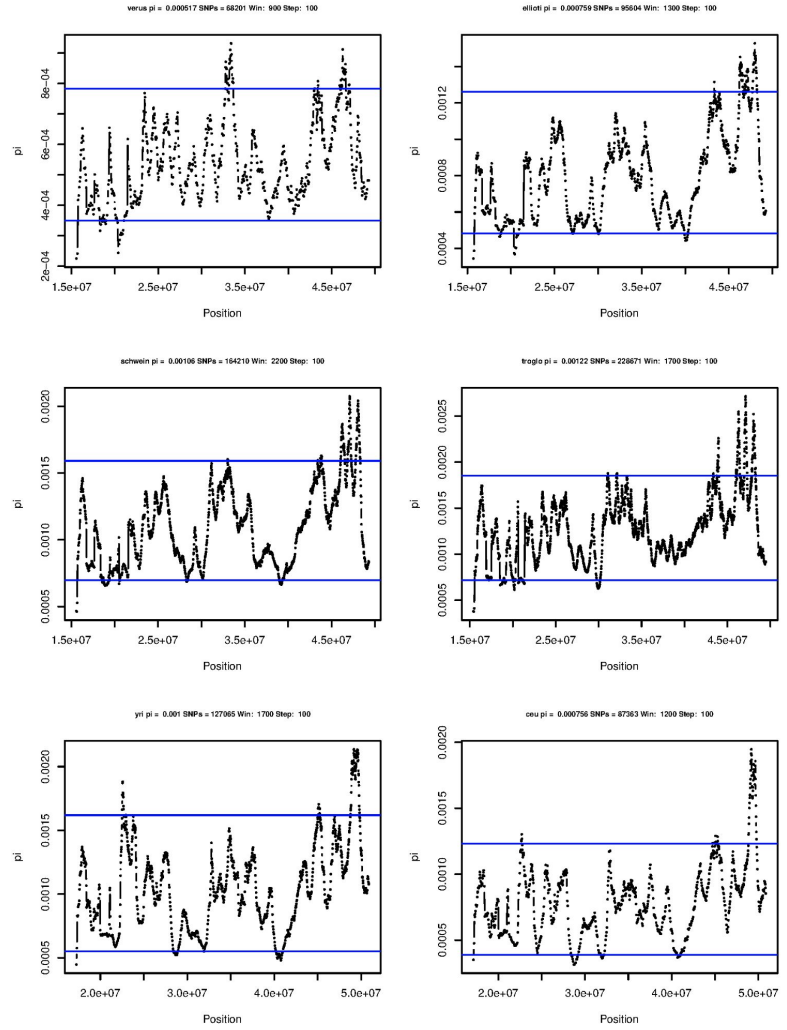
- **Q16:** Why is there a difference in pi along the chromosome?
- **Q17:** Why is the pattern different among the populations?



ESTIMATING THE NUCLEOTIDE DIVERSITY ALONG THE CHROMOSOME

- **Q16:** 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. There are also differences in recombination rate and mutation rate across the chromosome, that will also influence the amount of diversity.

- **Q17:** Why is the pattern different among the populations?
  Populations may be adapted to different environments, so there will be differences in which regions of the chromosome are more constrained by selection. They also 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 --bfile verus --het --out 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

- **Q18**: Is there a sign of inbreeding in some of the humans?
- **Q19**: Do some of the chimpanzees show signs of inbreeding?
- **Q20**: If so, how related do they seem to be?
- **Q21**: What is going on here? Why are the inbreeding coefficients so high?
- **Q22**: This effect is more pronounced in some populations (CEU in humans, verus, ellioti and scweinfurhii in chimpanzees), can you guess why is that?

ESTIMATING INBREEDING COEFFICIENT PR. INDIVIDUAL

- **Q18**: Is there a sign of inbreeding in some of the humans?
  No, all have an F value close to zero or negative.
- **Q19**: 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.
- **Q20**: 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.
- **Q21**: 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.
- **Q22**: This effect is more pronounced in some populaiton (CEU in humans, verus, ellioti and scweinfurhii in chimpanzees), can you guess why is that?

  These popualtions have less diversity, so the increase in the actual number of observed homozygous sites with respect to the expected when pooled together with the more diverse populations is higher than in the more diverse populations (YRI and troglodytes populations in humans and chimpanzees, respectively).

SUMMARY

- PLINK format
- Summary statistics in R
- Practical application of plink formatted data
- Graphical representation of the data and biological interpretation of results