Project

Albet Vidal, Roger Bosch and Òscar Casals

2022-10-04

Part A

Observing the data

For this project we will take a look at the information collected from the 1000 genome project by Fernando Racimo, Davide Marnetto and Emilia Huerta-Sánchez who wanted to find cases of Adaptive Introgression in humans using this data.

As in any other project of this kind the first step is to take a look at the data we are studding. We will do that by loading into R a table with the information of interest extracted from the ***Signatures of Archaic Adaptive Introgression in Present-Day Human Populations*** article written by the previously mentioned researchers.

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
DATA <- read_excel("TableS3_excelfile.xlsx")
```

If we take a look at the information each column of our table holds we will find that the third column of the data set is hard to read, this is because it holds the information of three different variables instead of just one. By dividing this variables in three different columns the contents of the table become more pleasant to the eye and easier to work with.

```
data <- DATA %>% separate(`Chr:Start-End`, c("Chr", "Start", "End"))
data
```

```
## # A tibble: 2,041 x 16
##    Mode       Outgro~1 Ingro~2 Moder~3 Archa~4 Chr   Start End   Der_Q~5 Uniq_~6
##    <chr>         <dbl>   <dbl> <chr>   <chr>   <chr> <chr> <chr>   <dbl>   <dbl>
##  1 continents     0.01     0.5 EUR     Nea     chr9  1664~ 1668~   0.588       2
##  2 continents     0.01     0.5 EUR     Nea     chr9  1672~ 1676~   0.698       6
##  3 continents     0.01     0.5 EUR     Nea     chr9  1676~ 1680~   0.698       4
##  4 continents     0.01     0.5 EUR     Nea     chr9  1680~ 1684~   0.689       2
##  5 continents     0.01     0.5 EUR     Nea     chr19 3360~ 3364~   0.65        1
```

```
##  6 continents     0.01     0.5 EUR     Den     chr17 1888~ 1892~   0.65        1
##  7 continents     0.01     0.5 EUR     Both    chr9  1672~ 1676~   0.68        1
##  8 continents     0.01     0.5 EUR     Both    chr9  1676~ 1680~   0.7         1
##  9 continents     0.01     0.5 EUR     Both    chr9  1680~ 1684~   0.712       3
## 10 continents     0.01     0.5 EAS     Nea     chr1  2326~ 2326~   0.56        8
## # ... with 2,031 more rows, 6 more variables: `D(Nea)` <dbl>, `D(Den)` <dbl>,
## #   `fD(Nea)` <dbl>, `fD(Den)` <dbl>, Genes <chr>, Lit_overlap <chr>, and
## #   abbreviated variable names 1: Outgroup_Max_Freq, 2: Ingroup_Min_Freq,
## #   3: Modern_pop, 4: Archaic_pop, 5: Der_Quantile, 6: Uniq_Shared
```

Meaning behind the columns

Now that the data is in a more presentable state we can start looking at the meaning of each column:

- **Mode:** It indicates from where the outgroup used for each sample comes from.
- **Outgroup_Max_Freq:** It indicates the maximum frequency each segment has in the outgroup used.
- **Ingroup_Min_Freq:** It indictaes the minimum frequency of each segment in the ingroup used.
- **Modern_pop:** Population where the sample comes from.
- **Archaic_pop:** It indicates if the sample is Neanderthal-specific, Denisova-specific or could belong to both genomes.
- **Chr:** It indicates to which chromosome the segment belongs.
- **Start:** It indicates in which position of the chromosome the selected segment starts.
- **End:** It indicates in which position of the chromosome the selected segment ends.
- **Der_Quantile:** 95% empirical quantile under neutrality.
- **Uniq_Shared:** ?
- **D(Nea):** ?
- **fD(Nea):** ?
- **Genes:** Gene encoded by the segment selected.
- **Lit_overlap:** ?

Transforming character vectors into factors

Some of the non-numerical values in the data set are translated as characters instead of factor in r, to prevent future mistakes we will transform set values into factors.

```
data$Mode = factor(data$Mode)
data$Modern_pop = factor(data$Modern_pop)
data$Archaic_pop = factor(data$Archaic_pop)
data$Chr = factor(data$Chr)
data$Genes = factor(data$Genes)
data$Lit_overlap = factor(data$Lit_overlap)
```

Wide format or Long format?

Finally we must determine if the table presented is in a long or wide format. It seems the data is in wode format due to the fact that ...

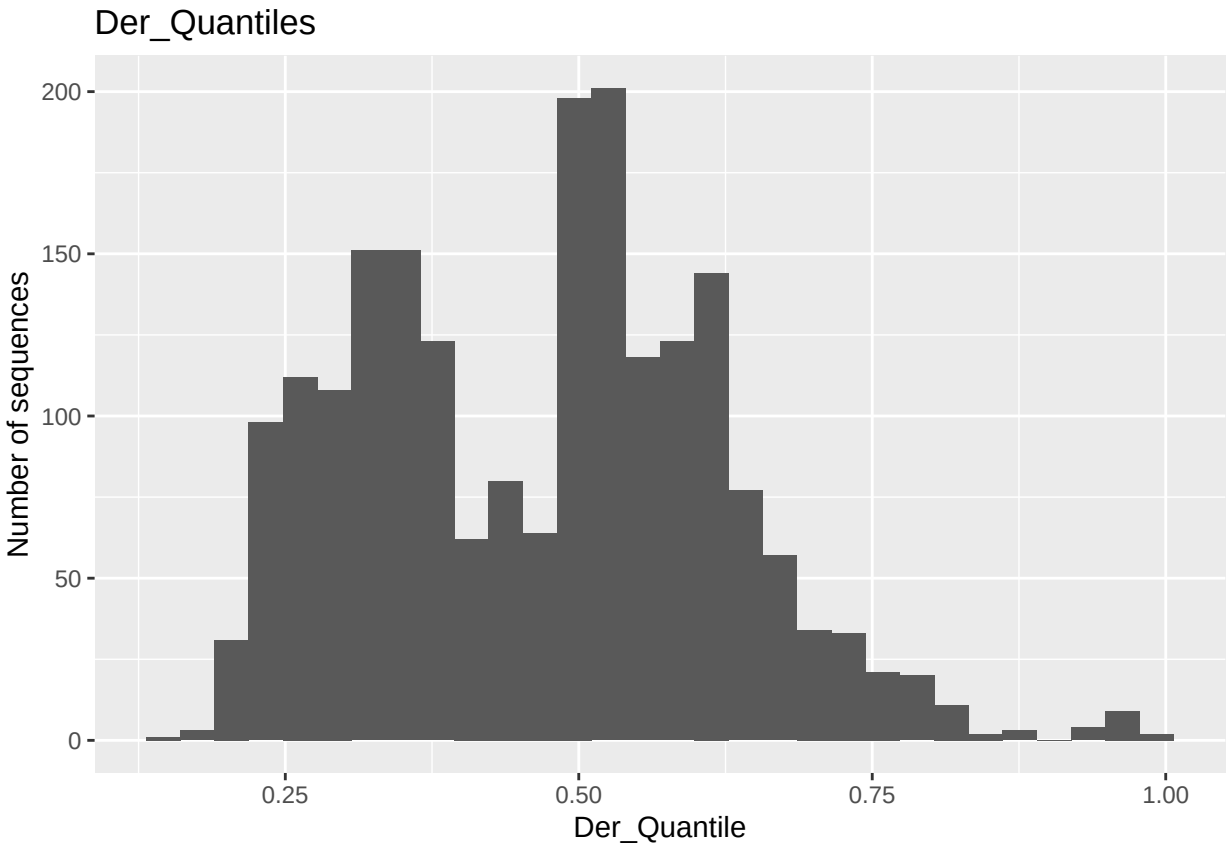
Part B

Distribution of the variables

Now that we know in what our data consists it is time to find out which distribution are they following. In order to accomplish this task we have to select a continuous and a discrete variable and represent them in a bar plot. The selected continuous variable was Der_Quantile and the discrete variable was the number of segments each quantile contains.

```
library(ggplot2)
graph_quant <- ggplot(data = data, mapping = aes(x = Der_Quantile)) + geom_histogram() + labs(y = 'Numbe
graph_quant
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

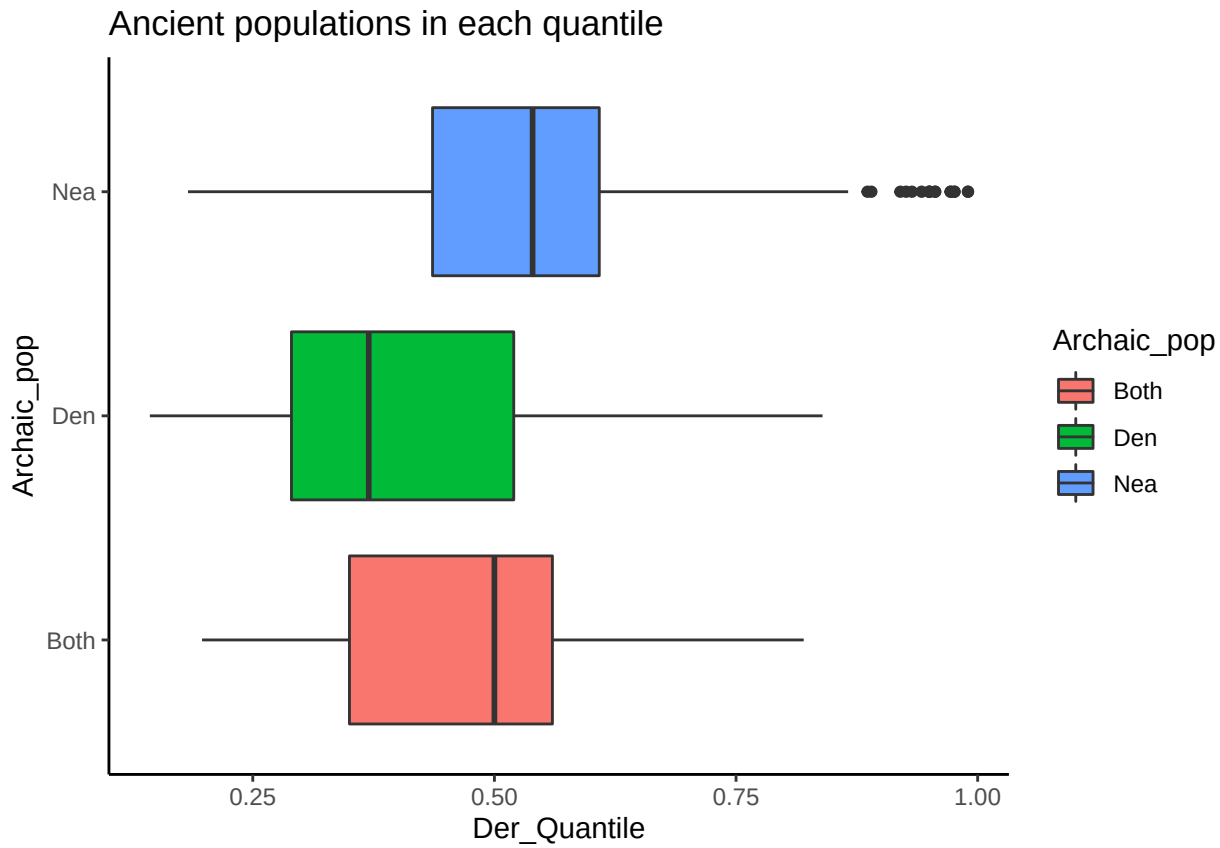


By looking at the plot above becomes apparent that our data follows a multimodal edge peak distribution.

We can also notice how most of the sequences are located in the third quantile(0.50-0.75) and how shallow is the quantity of sequences in the last quantile in comparassion to the rest of the graph.

##Sumerise the data A part from finding out the distribution of the quantiles we would also like to know the ancient populations that can be found in each of the quantiles. The best fitted graph for this task is the boxplot.

```
ggplot(data = data, mapping = aes(x = Der_Quantile, y = Archaic_pop, fill = Archaic_pop)) + geom_boxplo
```



Thanks to this boxplot we can observe that: neanderthal specific segments tend to be in the third quartile(0.50-0.75), Denisova specific fragments are usually in the second quantile(0.25-0.50) and the fragments that are not specific to any of the populations studied are mostly found in the second quantile(0.25 - 0.50).

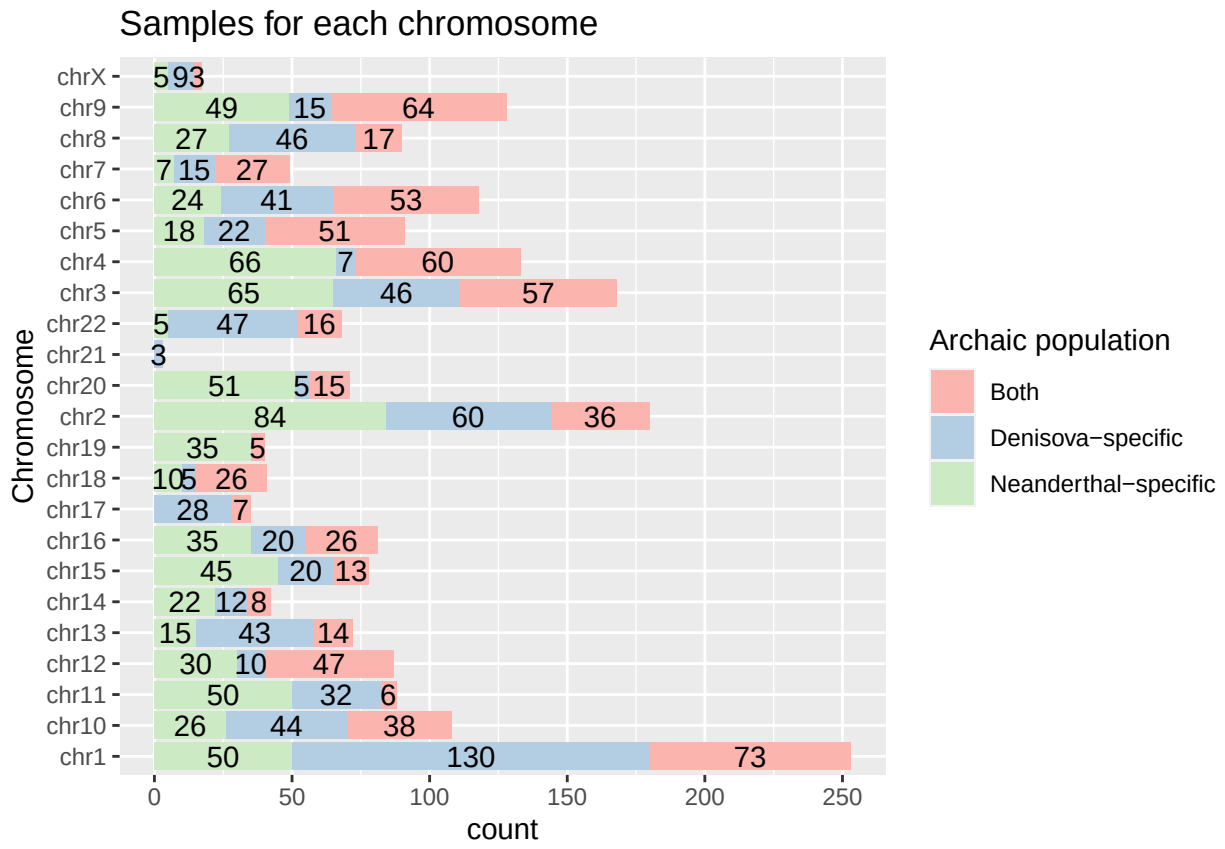
We can also see how that the few sequences present in the fourth quantile(0.75 - 1) are neanderthal specific.

Describing our data

Plotting the chromosomes

Now it is time to start looking at some other variables a part from the quantiles, this time we will see how many segments we have of each chromosome and if those segments are specific to an archaic population or not.

```
ggplot(data = data, mapping = aes(x = Chr, fill = Archaic_pop)) + geom_bar() + geom_text(mapping = aes(
```



The graph displayed above shows that most of the fragments in our graph come from chromosome 1 and most of them are Denisova-specific. The reason why we have so many samples of this chromosome could be because it is the largest in humans, therefore there is a lot of segments to extract from there.

We can also notice how we barely have samples of chromosome 21 and that all of the segments from it are Denisova-specific. We suppose this happended because chromosome 21 is the smallest one in humans.

Finally we can also notice that chromosome 17's samples are or Denisova-specific or Neanderthal-specific.A possible reason could be that this chromosome contains the the Homeobox B gene cluster, a DNA sequence involved in the regulation of patterns of anatomical developement.

Plot 2

????

Plot 3

????

Sources

Signatures of Archaic Adaptive Introgression in Present-Day Human Populations by Fernando Racimo, Davide Marnetto and Emilia Huerta-Sánchez

The information of each chromosome used to try to explain the data in *Samples for each chromosome* plot was collected from the wikipedia entries of each of the highlighted chromosomes.