Emma Houben r0303329 Management of large-scale omics data 1st master of bioinformatics

Data visualization exam assignment

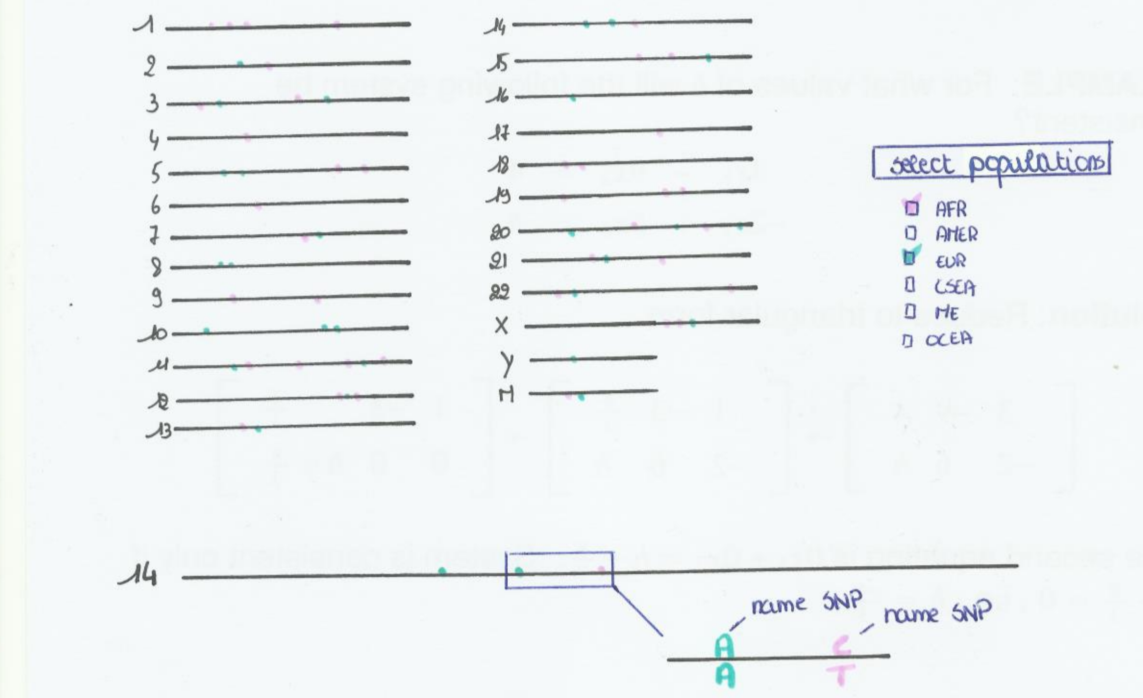
1. Task description

Task: Find unique SNPs for each of the 7 populations and determine on which chromosome they are located.

After cleaning the data, there appeared to be no unique SNPs for the East-Asian population. All other populations (Africa, Europe, South-Central-East-Asia, Middle East, Oceania and America) had a lot of unique SNPs (from over 30 000 for Africa to 560 for Oceania). Because of the amount of the data, it would be useful to distinguish them based on chromosome and population. The main point of the visualization is to find unique SNPs for a specific population, on which chromosome they are mainly located and to compare it to the other populations.

2. Design

Design 1

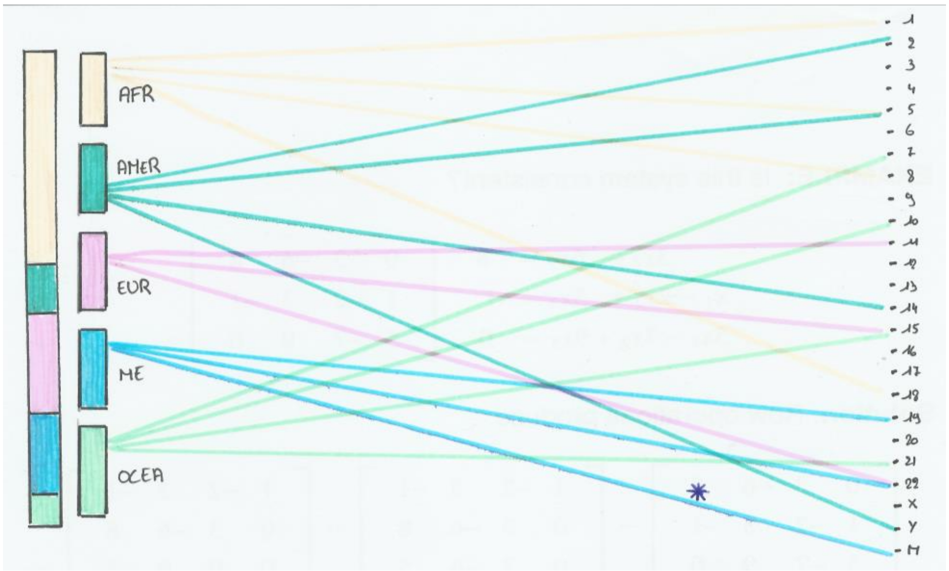


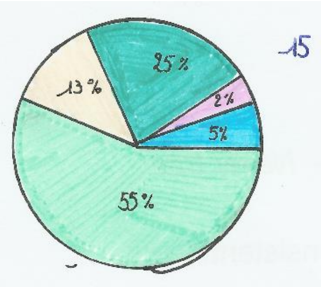
This design will show the chromosomes and the user will be able to select the populations from which he wants to see the unique SNPs. This SNPs will be shown as a coloured dot on the chromosomes. For example, if you click on chromosome 14, you will see an enlarged version of the chromosome. From here you can still zoom in on the unique SNPs to see what their exact position is, what there allele is and what the scientific name of the SNP is.

The advantages of this design: (1) you get an idea of where the SNPs are situated and (2) you can see the allele of the unique SNPs together. The disadvantages of this design: (1) there are a lot of SNPs found (more than expected) so the overview will be too small to show all coloured SNPs. (2) It will be difficult to estimate the difference between the populations.

Design 2

This design shows at the left side the populations and at the right side the chromosomes. The most left bar represents the average of SNPs per person. This is there to give an idea how many SNPs are found per population. The second row of bars at the left side represent the different populations. From here links go to the chromosomes where unique SNPs are found. When you click on the populations, you get the number of SNPs per person. When you click on a link (*), it will give you the percentage of SNPs for that population going to the chromosome. It will also give you the possibility to get a list with those SNPs. At the right side you see the chromosomes, when you click on those, you will see a pie chart (example 15), containing the proportion of populations with unique SNPs on that chromosome.

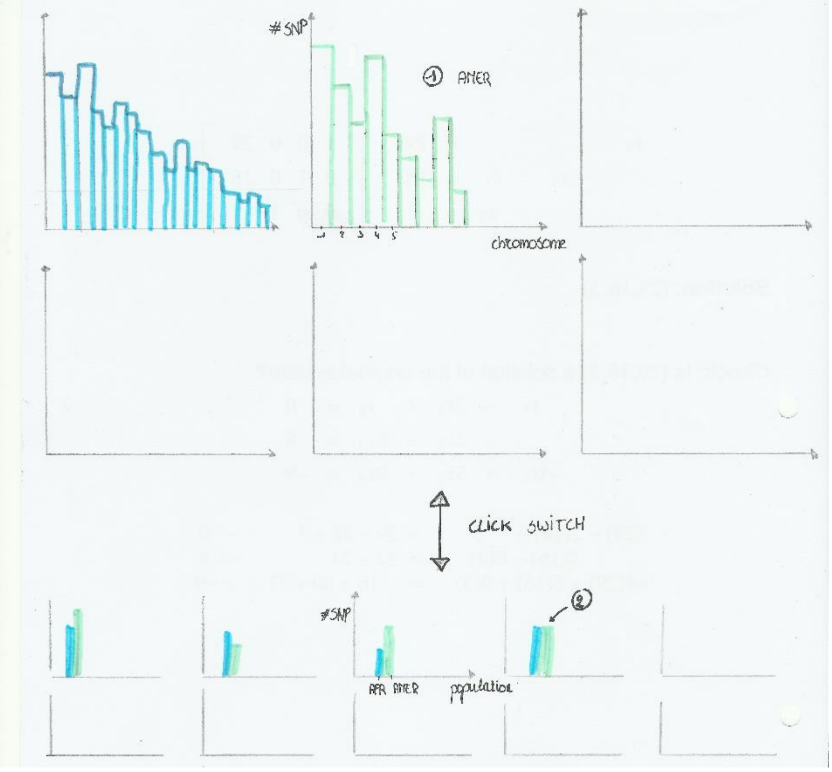




The advantages of this design is that you get an idea on which chromosomes the SNPs are, without visualizing all SNPs individually. The disadvantage: there can be a lot of links.

Design 3

This last design shows a lot of bar plots. In the first screen you see a bar plot for each population with the number of SNPs found at each chromosome. Here you have the possibility to select a bar to see where the SNPs are located on that chromosome or to retrieve a list of these chromosomes (1). You also have the possibility to change the view to get bar plots from every chromosome with the number of SNPs per population displayed. Here you also have the possibility to click on a bar and retrieve the chromosome with the SNPs on it, or a list (2). An advantage of this design is that you get a really good idea of the proportions, but it still is just a bar plot.

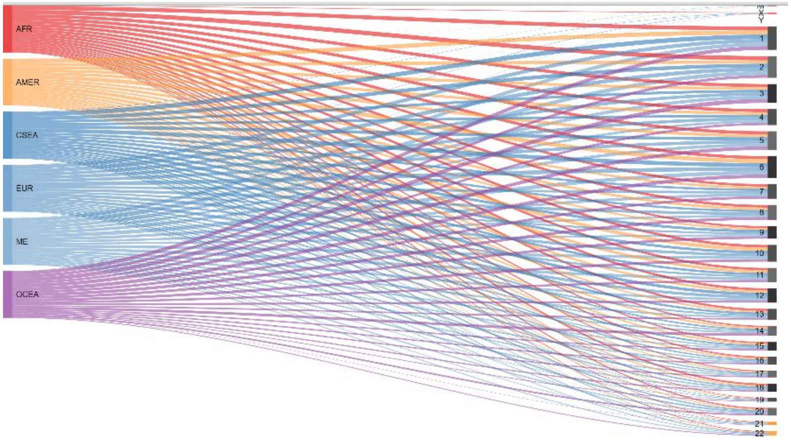


Conclusion

I choose to work with the second design because it has more advantages over the first design, and based on the data, the third design would not have been possible because for the African population there were over 30.000 SNPs, which is 10 times more than every other country. This would have been impossible to plot with the third design.

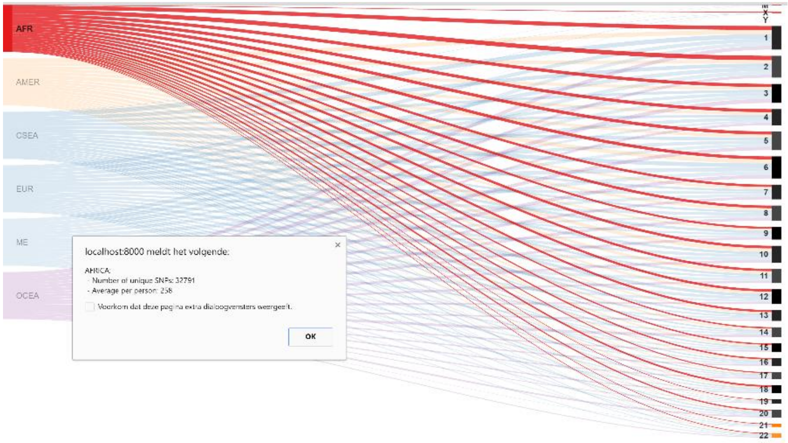
3. Implementation

The implementation was a lot harder than expected. I used Google visualization charts to make the Sankey diagram, but apparently this is still under construction, so a lot of basic things were not possible. Here is an example:



Here you can see the overall look of the visualization. The links from the population to the chromosome are proportional to the amount found for one population.

My first concern were the colours. I tried to give all the population nodes a unique colour (based on colorbrewer and colour blind safe) and the chromosomes just black, but this was not possible with the Sankey diagram.

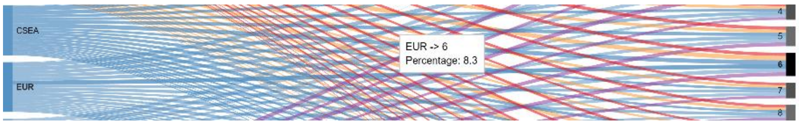


localhost:8000 meldt het volgende:

AFRICA:
- Number of unique SNPs: 32791
- Average per person: 258

□ Voorkom dat deze pagina extra dialoogvensters weergeeft.

When you click on a population, its links get highlighted and a message is shown with the number of unique SNPs and an average per person. I wanted a bar next to the graph (as in the design part) that gave an overview of the number of SNPs per person from each population, but this did not work.

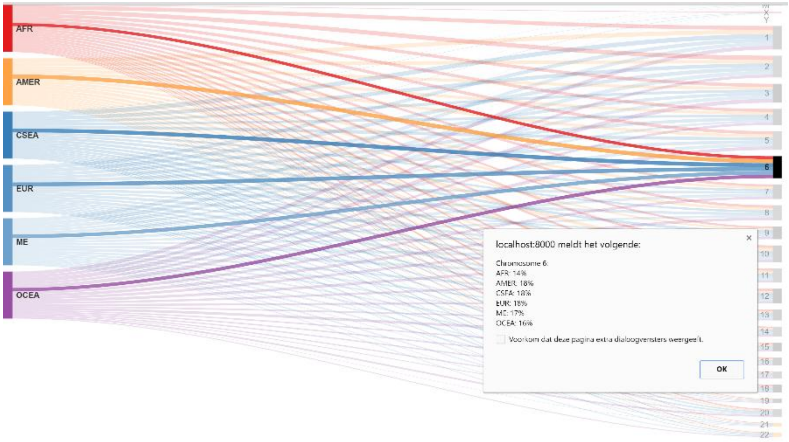


EUR -> 6
Percentage: 8.3

When you hoover over a link, it mentions the population, the chromosome and the percentage relative to the total SNPs for a population.

I wanted to give the user the chance to get the list of SNPs for that specific population on that specific chromosome when you click on the link, but this is only possible for other charts. The Sankey diagram does not yet support this feature.

The last feature appears when you click on a chromosome. Here you can see which population a unique SNPs had on this chromosome and it gives a messages with the percentage of the population on this chromosome. Here I wanted to make a pie chart pop up with the populations.



localhost:8000 meldt het volgende:

Chromosome 6:
AFR: 14%
AMER: 18%
CSEA: 18%
EUR: 18%
ME: 17%
OCEA: 16%

□ Voorkom dat deze pagina extra dialoogvensters weergeeft.

So in general, the Google Sankey diagram has been a valuable tool to visualize majority of the data analytics I had in mind, although not all features were currently available to implement everything.

4. Insights

There are a few points that can be improved after creating this visualization: because of the big amount of unique SNPs (what I didn’t expect), it would be more useful to split the population in smaller groups (based on land or clan). Especially the African population should be divided more because of its enormous amount of unique SNPs.

A second thing that can be done after this is try to link the unique SNPs to diseases. With this information you can determine the disease causing SNPs unique for one population. This can have an be used in the upcoming domain of “personal medicine”.

A last application with this visualization can be used in criminology. When DNA is found, people can look for specific SNPs and determine the suspects origin. For this, the visualization should support the selection of a specific SNP and mention in which population it is found. For this application, there can be looked into combination of different unique SNPs for 1 population to make the result more accurate.

Apart from new ideas for the further investigation, I also would try to improve my visualization when new features are available for this Google chart.

5. Screencast

Note: the screencast does not support the pop up messages in my visualization. Please find attached the code in my email in case you want to validate this.

https://youtu.be/wjd9GDNa9dk