# DATA VISUALIZATION

## 1. Task description

The data chosen for this assignment was the human genome diversity project data, where genomic DNA of 1043 individuals were collected around the world to determine their genotypes on more than 650,000 SNP loci. The dataset also contains mapping the SNP's to the chromosome region and positions (http://www.hagsc.org/hgdp/files.html) . The aim of my task was to **determine the variation of the SNP in each region** that is: Africa, Europe, Oceania, America, Central-South Asia, Middle East, East Asia. Since the data set were categorical variables, I decided to run a number of SNP at SPSmart (http://spsmart.cesga.es/) website to get a link of the chromosomes and regions together from the CEPH u. Stanford HGDP dataset.

| | SNP | chromosome | position | validation | genes | reference | ancestral | var | population | N | freq_A | freq_C | freq_G | freq_T | MA | MAF | Hobs | Hexp | Fs | Fst | In |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs675333 | chr11 | 99044864 | yes | CNTN5 | G | G | GT | Population Set 1 | 943 | 0.000 | 0.000 | 0.481 | 0.519 | G | 0.481 | 0.463 | 0.499 | - | 0.053 | 0 |
| 2 | rs675333 | chr11 | 99044864 | yes | CNTN5 | G | G | GT | AFRICA | 102 | 0.000 | 0.000 | 0.676 | 0.324 | T | 0.324 | 0.392 | 0.438 | - | 0.047 | 0 |
| 3 | rs675333 | chr11 | 99044864 | yes | CNTN5 | G | G | GT | AMERICA | 64 | 0.000 | 0.000 | 0.219 | 0.781 | G | 0.219 | 0.250 | 0.342 | - | 0.262 | 0 |
| 4 | rs675333 | chr11 | 99044864 | yes | CNTN5 | G | G | GT | EUROPE | 158 | 0.000 | 0.000 | 0.408 | 0.592 | G | 0.408 | 0.475 | 0.483 | - | 0.075 | 0 |
| 5 | rs675333 | chr11 | 99044864 | yes | CNTN5 | G | G | GT | MIDDLE EAST | 163 | 0.000 | 0.000 | 0.433 | 0.567 | G | 0.433 | 0.546 | 0.491 | - | 0.029 | 0 |
| 6 | rs675333 | chr11 | 99044864 | yes | CNTN5 | G | G | GT | CENTRAL-SOUTH ASIA | 200 | 0.000 | 0.000 | 0.495 | 0.505 | G | 0.495 | 0.470 | 0.500 | - | 0.018 | 0 |
| 7 | rs675333 | chr11 | 99044864 | yes | CNTN5 | G | G | GT | OCEANIA | 28 | 0.000 | 0.000 | 0.268 | 0.732 | G | 0.268 | 0.321 | 0.392 | - | 0.254 | 0 |
| 8 | rs675333 | chr11 | 99044864 | yes | CNTN5 | G | G | GT | EAST ASIA | 228 | 0.000 | 0.000 | 0.566 | 0.434 | T | 0.434 | 0.500 | 0.491 | - | 0.048 | 0 |
| 9 | rs6753378 | chr02 | 80096033 | yes | CTNNA2 | G | A | GA | Population Set 1 | 944 | 0.513 | 0.000 | 0.487 | 0.000 | G | 0.487 | 0.407 | 0.500 | - | 0.131 | 0 |
| 10 | rs6753378 | chr02 | 80096033 | yes | CTNNA2 | G | A | GA | AFRICA | 102 | 0.147 | 0.000 | 0.853 | 0.000 | A | 0.147 | 0.255 | 0.251 | - | 0.082 | 0 |
| 11 | rs6753378 | chr02 | 80096033 | yes | CTNNA2 | G | A | GA | AMERICA | 64 | 0.234 | 0.000 | 0.766 | 0.000 | A | 0.234 | 0.313 | 0.359 | - | 0.112 | 0 |
| 12 | rs6753378 | chr02 | 80096033 | yes | CTNNA2 | G | A | GA | EUROPE | 158 | 0.658 | 0.000 | 0.342 | 0.000 | G | 0.342 | 0.430 | 0.450 | - | 0.064 | 0 |
| 13 | rs6753378 | chr02 | 80096033 | yes | CTNNA2 | G | A | GA | MIDDLE EAST | 163 | 0.442 | 0.000 | 0.558 | 0.000 | A | 0.442 | 0.429 | 0.493 | - | 0.024 | 0 |

Table: *A sample of the dataset*

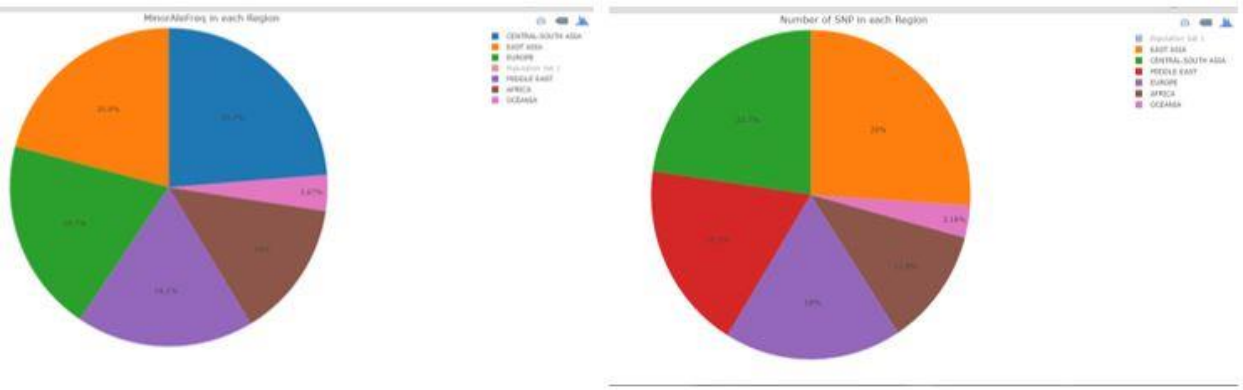
## 2. Design:

The software tools used for the designs were JMP, Tableau and RStudio, an R package known as plotly which is an interactive browser-based charting library built on the open source JavaScript library, plotly.js. ( https://plot.ly/r/)

My first design was to get a general view of the data with respect to each region and try to find out which region or population differ the most amongst the other regions. So I create a pie chart of the number of SNP's in each region and also the minor allele frequency of each region for comparison. In both case we can see that the region

## DATA VISUALIZATION ASSIGNMENT

Oceania has the least SNP's in its region. But this does not permit the contrast between the SNP's and the population. This can be considered a good visualization if we just want to know the number of individuals in each continent denoted by different colors and portion sizes according to the number of people from that region.




*Pie chart of the SNP's and minor allele frequencies of each population created with R package plotly.*

- My second design was to make a histogram plot of the SNP's and how they vary in each region(population). The idea is to get a clear visual of the data so you can easily identify which SNP's differ in which region/population. The different colors of each bar represent the country and from the plot we can see which SNP's differ. Each bar represents a particular SNP's and the different colors represents the population or continent. This visualization permits use to see that variability of the minor allele frequencies in each region and the stacked color contrast makes it easy to interpret. I would consider this a good visualization with respect to my task.

## DATA VISUALIZATION ASSIGNMENT

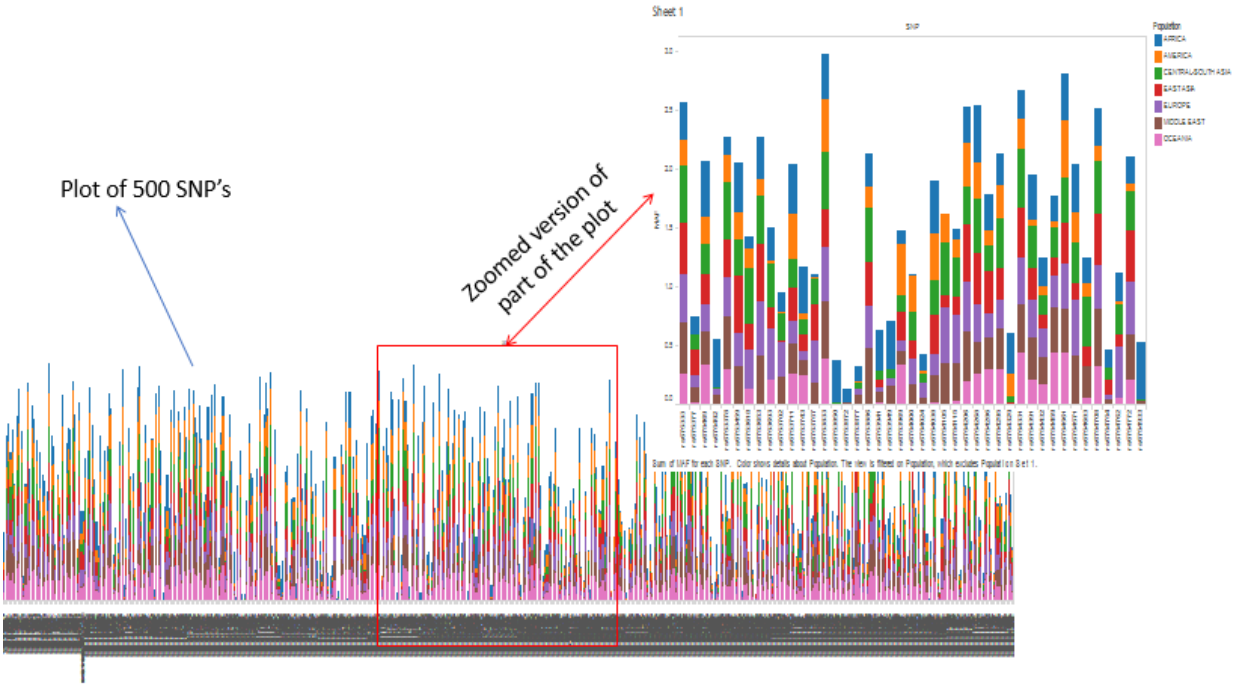



*Figure 1. plot showing the variation of SNP in each region using Tableau*

**3. Implementation of final design.**

- My final decision was to create a bubbles plot or scatter plot so I can clearly identify which part of the plots has missing points/bubbles and in which regions they belong. According to the Fds design methodology we should create a visualization in such a way that the user can understand the task while looking at the data. So with this in mind, I thought a bubbles plot would a good way to represent the variability of the SNP' s in each region. Each population is represented by a single color and the missing bubbles in each population would indicate the SNP that is different compare to other population. Additional when you place the mouse on a bubble you get other information like the chromosome number, chromosome position, number of individuals in each population, minor allele, the gene name if available since some of the SNP's when scanned with SPSmart did not have a gene name assigned to it.

## DATA VISUALIZATION ASSIGNMENT

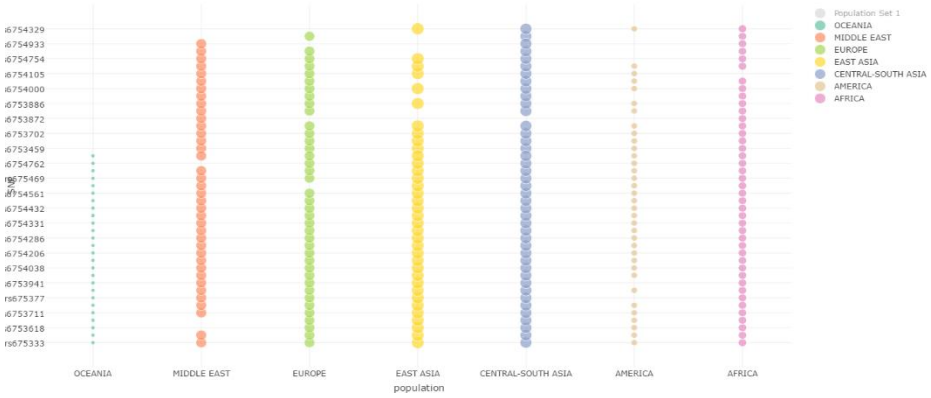



*Figure showing the variability of SNP in each region*

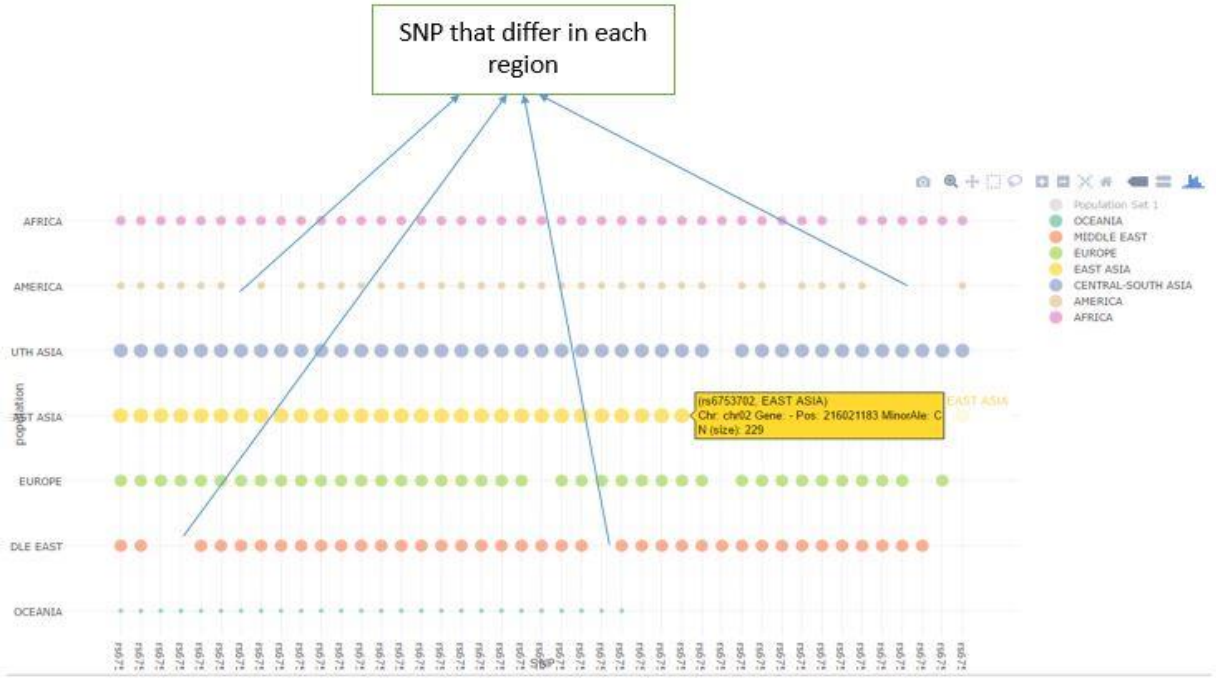


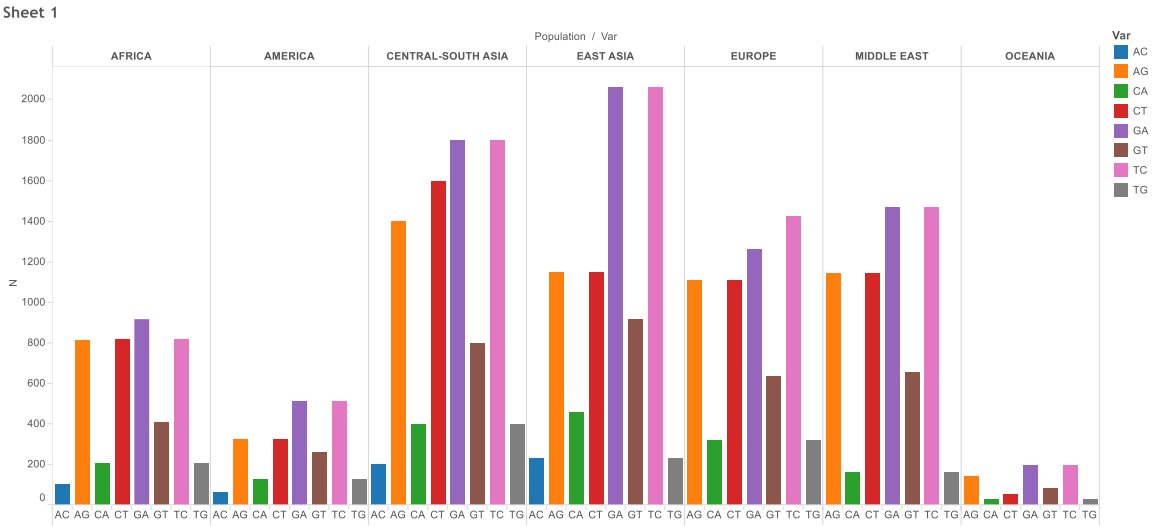

*Figure showing the various regions with missing SNP in each population.*

The screencast of the visualization can be found at :https://www.youtube.com/watch?v=ne5K2YLqwrM .

DATA VISUALIZATION ASSIGNMENT

## 4. Insights

Let says I wish to determine the variation of the alleles in each region that is searching for a genotypic insight of their DNA structure and to know which population is more susceptible to a disease. A combination of certain alleles can predispose an individual(population) to a certain disease so looking at the plot we can see what combination of alleles are in each population. The different colors represent the different alleles and from the plot we can see that East Asia, Europe, MiddleEast and Oceania has the blue bar absent which is the AC combination




Sum of N for each Var broken down by Population. Color shows details about Var. The view is filtered on Exclusions (Population,Var) and Population. The Exclusions (Population,Var) filter keeps 60 members. The Population filter excludes Population Set 1.

The code for the pie chart and bubbles plot were derived with reference from : https://plot.ly/r/reference/#pie and https://plot.ly/r/