KU Leuven

Master of Bioinformatics

Management of Large-Scale Omics Data - I0U19a

Data Visualization Report

Author:
María del Carmen Bravo González-Blas

May 25, 2016

1 Task description

The presented visualization will make use of the data provided by the Human Genome Diversity Project and Eupedia in order to compare how likely is to develop a disease or trait based on the average genotype in each of the countries present in the study. We will make use of:

- The genotypes of the HGDP participants: The data contains the allele variant present in each of the 1043 participants for 660918 SNPs.
- The information about the country of origin of each participant [1].
- The risk associated to each allele variant of the SNPs related to a certain disease or trait: There are 675 main SNPs related to disease listed in Eupedia. Other databases, as SNPedia, contain more SNPs related to disease than Eupedia, however, Eupedia provides free access to the risk information while SNPedia would require using their private software Promethease or more extensive data mining to retrieve all the information. We will also assume that the effect of the variants is additive.

The data has been processed as shown in *Figure S1* in order to obtain a matrix with the required information for the selected visual, in which the columns represent countries, the rows the diseases and traits and the cells the corresponding risk score. Scripts and data can be found clicking here. The screencast of the visualization can be watched here.

2 Design

In this section, we will present three different designs, describing the advantages and drawbacks of all their components. The third design will be the one selected for its implementation.

2.1 "Group" design

The first of the designs is shown in *Figure 1.* In this visual, we propose to group the individuals by countries, being each individual shown as a dot. When the combination of allele variants results in a higher risk of developing the disease or trait, the dot will be red; while if it has a beneficial effect it will be green. The radius of the circle will indicate how high is the risk or benefit (the bigger the radius, the bigger is the effect). The disease or trait can be selected in a drop-down menu.

However, we find more difficulties than advantages when interpreting this visual. Its main advantage is that we can see the effect in each of the patients, so if there is one outlier in one of the groups it could be easily spotted. However, visualising by individual would not provide a clear first insight. Even grouping by country, dots may overlap, it

[1] For one of the participants, this information was found to be missing.

may be difficult to extract a conclusion in those countries were many participants are present and the comparison between the radius of the circles may be hard. Furthermore, we would also loose the distances between countries (in our visualization, we would also expect to see more similarities between countries in the same continent than the further ones) or its size, since the size of the cluster is determined by the number of participants of that country. If we tried to represent sizes of countries also (e.g as size of the group), it would be impossible to draw the dots corresponding to the 145 patients in Israel! This way the size of the group should be large enough to fit all dots, not exactly the size of the country.

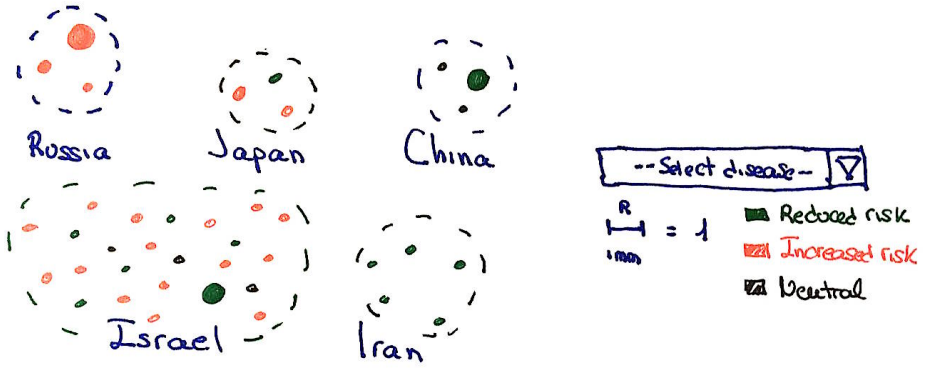

Figure 1: "Group" Design. Countries are represented as groups of dots that represent their native patients. The size of the cluster will depend on the amount of patients, not the size of the country. Red dots indicate individuals that have a higher risk of developing a disease/trait; while green ones represent reduced risk. The radius of the circle represents is used to quantify how high or low the risk is. The disease/trait to be displayed is chosen with a dropdown menu.

When it comes to the way of selecting the disease or trait, if the list is long it may be hard to find a concrete disease or trait. In this visualization we ended up with data for 86 diseases/traits, so it may be interesting to find a more friendly way to interact with it.

2.2 Barplot design

The second of the designs is shown in *Figure 2*. In this visual, we would display a barplot upon selection of disease/trait per country. We previously saw that grouping by individual does not seem to be a good approach, so in this case each barplot would represent the average risk per country. The bars are grouped by continent. To improve the disease/trait selection, the diseases/traits will be grouped by type. Once a type is selected, there would be another drop-down menu in which only the diseases related to that type would be displayed, reducing the list of diseases to select.

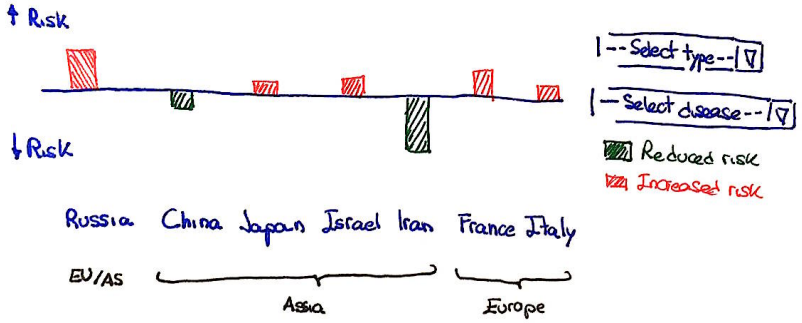

Figure 2: Barplot Design. Each bar represent the average risk for the selected disease/trait in the corresponding country. Country bars are organised by continent. Two drop-down menus are available, one for selecting the type of disease and another one, whose content will depend on the selected type, containing the diseases.

This representation supposes an increase in clarity compared to the first one. By looking the barplots we can get a quick idea if there are differences between countries, depending on how plane the barplot profile results. However, even if we are grouping the bars by continent, we still do not have good references for distances between countries.

2.3 Map design

The third of the designs is shown in *Figure 3*. In this visual, we display a map that will be coloured by countries upon selection of disease/trait. If the country is not represented in the study, then it will remain in white. In this case, we include a diverging color scale to represent the risk . The method for selecting the disease/trait remains the same as in the second proposal.

Figure 3: Map Design. Given a world map, the countries for which data is available will be coloured in function of the risk of disease/trait. The colour scale is given at the bottom left and ranges from the minimum to the maximum value for the selected disease/trait. Two drop-down menus are included, similar to those in the second design.

This representation supposes an increase in clarity compared to the others. Now we not only have information about the average risk per country, but we can also capture distances between countries. Of the three proposed designs, this is the one that provides a more clear first insight of the data to answer our question.

3 Implementation

We have decided to implement the map design, using p5.js (v 0.5.10). The visualization can be used clicking here. A screenshoot is shown in *Figure 4*.

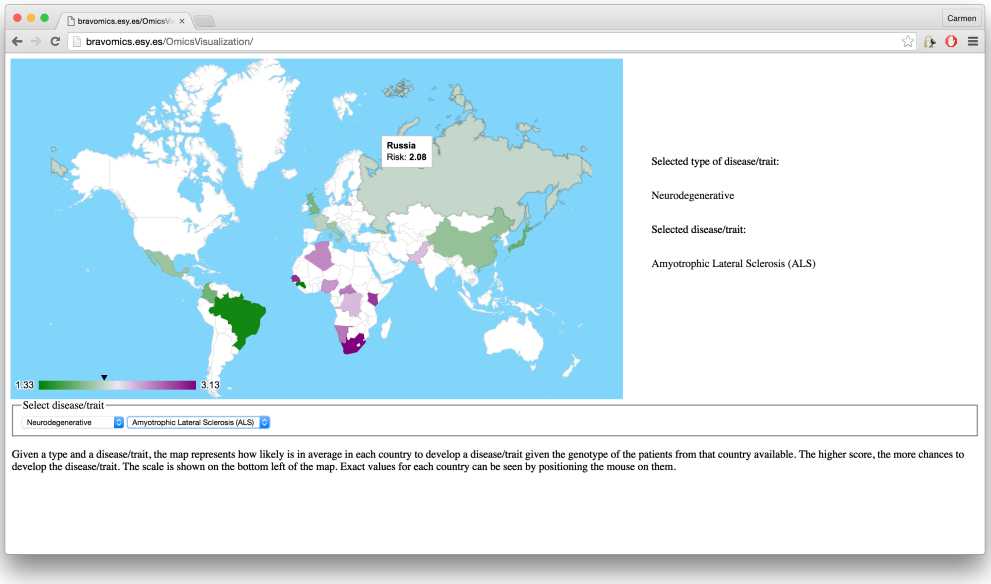

Figure 4: Visualization. Screenshoot of the visualization tool developed based on the map design.

We can see all the elements mentioned in the map design. We have chosen a divergent "green-purple" color scale (instead of e.g "green-red") because it is colorblind friendly, according to Harrower and Brewer[2]. With this scale, countries with lower risk of developing the disease or trait will be coloured bright green; the ones with higher risk, bright purple; and the ones with intermediate risk will be in the change of hue zone and coloured lighter. Furthermore, we have added instructions and a panel at the right showing the disease/trait that is being displayed. In order to improve clarity, we have added another interaction: if the user positions his mouse in a country, the name and numerical value of the risk are also displayed. The "geoChart" library from Google Developers has been used for the development of the tool, together with p5.js libraries. All the files used for this visualization can be found here.

[2]Harrower, M., & Brewer, C. A. (2003). ColorBrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal, 40(1),* 27-37.

4 Insights

One of the main patterns that we can see in our visualization is that countries that are nearer seem to have similar risk of developing a disease or trait. For example, in *Figure 4*, in which we display the risk for Amyotrophic Lateral Sclerosis (ALS), we see that there is more risk of developing it in African countries that in the rest of the world. Similar trends are seen in most of the other diseases/traits, e.g. lung cancer is much predominant in Europe; leukemia is much less common in African countries. We can expect that someone from Spain or Belgium, for which we have no data, will present a risk to develop a disease or trait similar to the ones found for UK, France or Italy. These results suggest that self-reported ancestry can facilitate assessments of epidemiological risks. Rosenberg *et al.* (2002)[3] had already reported that there is more genetic variation between distant populations, also using the HGDP data.

The objective of our visualization is to give insight on how the risk for developing certain disease/trait is by country, however, it also has its limitations. First, we have only used the most important related SNPs retrieved from Eupedia, not all of them. More SNPs and their risk are available in SNPedia, however, the software for its retrieval could not be installed. The visualization could be more realistic if we added these other SNPs.

Furthermore, in Eupedia the risk per allelic variation is represented as more (or less) than 3x chances of (not) developing the disease/trait or neutral. We have given 3 points when the chances increased/decreased more than 3 times, 1 point when they were lower (negative if it reduced the risk, positive if it increased it) and 0 when neutral. In addition, we have also considered that there is an additive effect (e.g, if you have an allelic variation with 3 points and another with 1 points, then your risk of developing the disease/trait is 4 times higher).

The number of patients per country was not homogeneous neither. The country with more patients in the study was Pakistan (197), while Namibia only had 6 participants. More sampling should be made in order to obtain more trustable results. *Figure 5* shows the number of participants per country.

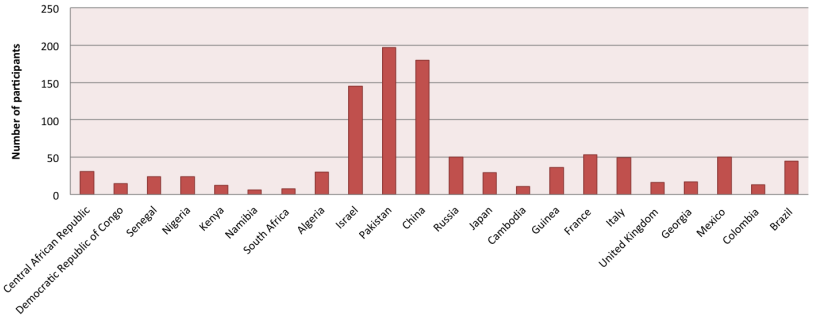

Figure 5: Number of participants per country.

[3]Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science, 298*(5602), 2381-2385.

Supplementary Material: Data Visualization Report

Before developing our visualization, the data had to be preprocessed. *Figure S1* shows the pipeline used. All the scripts were written in Python (v 2.7.10). We start with the data of the HGDP genotypes and Eupedia SNPs. The first step is to filter both files in order to keep only the information about the SNPs that are present in both files (*Filtering.py*). Next, we merge the outcomes from the first step in a matrix, in which columns will be the patients and rows the SNPs (*MatrixGen.py*). In the cells, the symbols '++', '+', '.', '-' and '- -' represent the effect of the corresponding allelic variation in each patient (positive means that the risk increases; if negative, it decreases). Next we give a numerical score to the symbols ('++': 3, '+': 1, '.': 0, '-': -1, '- -': -3) and group the SNPs by disease/trait (*Punctuation.py*). Finally, we group the patients by country and divide the overall risk score by the number of participants, obtaining a matrix in which columns represent countries, rows diseases and the cells the corresponding average risk score (*byCountry.py*). The final text file is converted to csv to be used in the visualization. All scripts can be found clicking here.

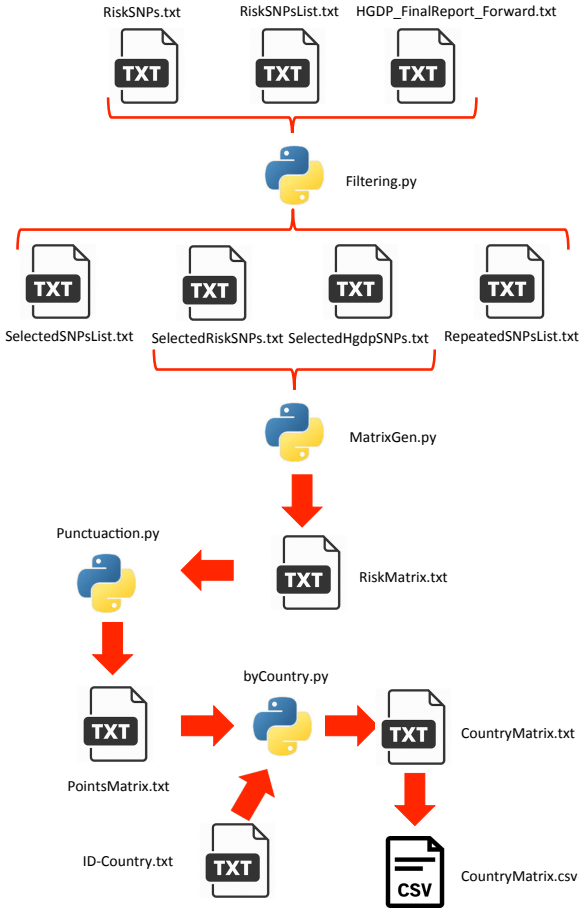

Figure S1: Data Processing pipeline.