Management of Large-Scale Omics Data [I0U19a]:

Visualisation Assignment: HGDP data

Sebastiaan Vanuytven
R0259061
1st Master of Bioinformatics

1. Task description

The Human Genome Diversity Project was a study performed in 2002 with the goal to understand genetic diversity in human populations. Researchers from the Stanford University analysed genomic DNA from 1,043 indigenous individuals from around the world, determining their genotypes at more than 650,000 SNP (single nucleotide polymorphism) loci. The goal of my visualization is to enable users to find how many and which of these SNPs are associated with a disease and where they are located on the human genome. Furthermore, the visualization should give insight in the distribution of risk alleles of these SNPs in the different continents (Oceania, Europe, Asia, Sub-Sharan Africa, America and the Middle East). An allele in this visualisation can be high risk, low risk or non-risk. A high risk allele is defended as an allele that increases the risk of getting a certain disease or condition more than three time, vice versa for low risk alleles. A database with all diseases-associated SNPs was constructed from data found in the Eupedia (http://www.eupedia.com/genetics/medical_dna_test.shtml) and Clinvar (http://www.ncbi.nlm.nih.gov/clinvar?term=human%5Borgn%5D) databases. 159 of these SNPs were found in the HGDP project and the prevalence for each possible allele was calculated for the different continents. To conclude the disease-associated SNPs will be pooled in 8 general categories based on category the disease belongs (Cancer, Autoimmune disease, Neurological conditions and disorders, Cardiovascular disease, Female-specific diseases and disorders, neurodegenerative diseases, Gastrointestinal diseases, Genetically determined addictions and Miscellaneous). For each of these categories an expected distribution of risk alleles was calculated based on the available disease-associated SNPs. Because we didn't find any disease-associated SNPs on the sex chromosomes or mitochondrial DNA in the HGDP dataset, these genomic regions were not visualised

2. Design

See attachments for the drafts of the 3 designed visualisations

Visualisation 1:

The layout of this visualisations will be dominated by a simple chromosome map of the autosomal chromosomes (1 to 22) where all disease-associated SNP's are plotted at the correct location on the correct chromosomes as dots (left-side). The colours of the dots will be based on the disease category to which the disease-associated SNP belongs. The colours were determined by using ColorBrewer2. On the right top side, a legend will be shown where the colours are matched with the corresponding disease category. This visualisation enables us to study the localisation of the disease-associated SNPs, as well as giving insight which disease category contains the most SNPs. In this way we can study potential clustering of SNPs, identify potential hotspots for disease-associated SNPs of find chromosomes that don't harbour any of these SNPs. The second visualisation consist out a stacked bar chart depicting the distribution of risk alleles (high, low and non-risk) at the bottom right side. In the Y-axis, the 6 different world regions are depicted, while the x-axis is used for the percentage. Two different scenarios can be distinguished. When hovering over an SNP dot at the chromosome, the staked bar plot will depict the distribution of risk alleles for the SNP observed in the HDGP dataset. Furthermore, extra information about the SNP like name (RS-number), location, associated disease, high/low/non-risk allele, … will be shown above the bar plot. This visualisation enables us to compare the allele distribution between the different regions studied in the HGDP project.

In the second scenario when hovering over a disease category in the legend, a stacked bar chart will appear that depicts the expected distribution of risk alleles for a SNP belonging to this category.

Visualisation 2:

For the second visualisation I chose to use a bubble map to depict the percentage of the population that carries a high-risk allele for a certain the disease in each region of the world. The colour of the bubble will dependent on the category to which the disease belongs (colours were determined with ColorBrewer2). Above the world map, a selection menu will be implemented, containing the 8 different disease categories in their corresponding colour predeceased by a checkbox. Clicking on one of this categories opens a menu where one disease can be selected. Underneath the world map we can find a stacked bar plot depicting the number of disease-associated SNPs for each disease category on each chromosome (Y-axis, number of SNPs, X-axis, chromosome number). This visualisation enables us to identify which chromosomes that are favoured to harbour disease-associated SNPs (of a certain disease category), as well as chromosomes that contain no SNPs. There is however no information about the location of the SNPs in comparison with each other or about the length of chromosomes, which can potential blur the hypothesis we suggest based on this figure. Furthermore, we can find SNP information sheets that contain all the information know about the SNP (name, gene, chromosome, location, associated disease, category of disease, high/low/risk alleles) and the distribution of risk alleles based on the information obtained from the HGDP project while hovering over the bubble of a disease in the world map. To

visualise the distribution of the risk alleles for the different regions in the world, I chose to use pie plots containing 3 potential regions. This visualisation enables us to compare the allele distribution between the different regions, however the stacked bar plot seems more easy to inspect since all of the information is in one figure. Here we have to scan our eyes across 6 different figures. The number of SNP information sheets depends on the number of SNPs that are associated with the same disease (the majority of disease in the obtained dataset only have one associated SNP). A major negative point of this visualisation is that we have to know with which disease a certain SNP is associated to observe the distribution of risk alleles for this individual SNP.

Visualisation 3:

In visualisation I started with a plie plot that consists out of three different circles. The inner circle depicts the different disease categories (in a specific colour scheme, determined with ColorBrewer2), in the middle one we can find the different diseases belonging to a specific category (depicted in the same colour as the corresponding category). The Individual disease-associated SNPs name can be found in the outer circle, again the same colour as the category to which it belongs, is used to depict the name. This figure enables us a quick overview of all disease categories in the dataset as well as which categories contains the most SNPs. Furthermore, we can quickly observe which SNPs are associated with a certain disease and for which diseases we can find associated SNPs with risk alleles are found in the HGDP project. To visualise the distribution of the risk alleles I chose to a bar plot on the world map. Three bars can be observed for for each region in the world (high/low/non-risk allele in the corresponding colour-. Three different scenarios can be observed for the bar plot. When the user hovers over a disease category, bar chart will appear that depicts the expected distribution of risk alleles for a SNP belonging to this category. On the other hand, when the user hovers over a certain disease in the middle circle of the pie plot, a bar chart will appear that depicts the expected distribution of risk alleles for a SNP associated with this disease. Hovering around an individual SNP will reveal an SNP information sheet as described in the previous visualisation and a bar plot on the world map that depicts the distribution of risk alleles for this SNP. In this visualisation we don't obtain any information about the location of the disease-associated SNPs

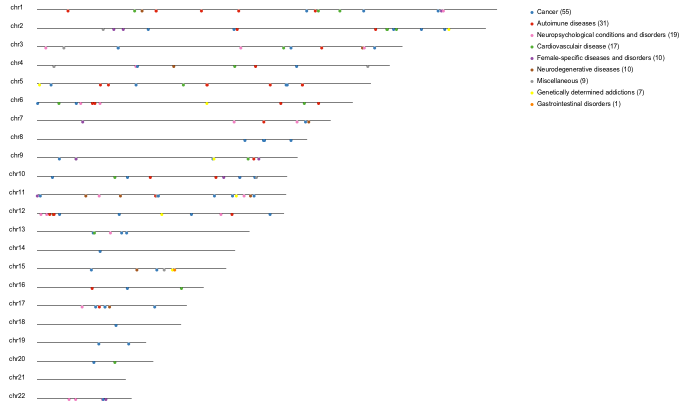
I picked my first design because it enables the user to also visualise and study the location of the SNPs in comparison to each other. Although the location is not new data added by the HGDP project, not a lot of articles and research can be found on clustering/localisation of disease-associated SNPs. It would be interesting to potentially find a chromosome that doesn't contain any disease-associated SNPs or chromosome that harbour a lot of SNPs of particular category (are these genes under control of the same promotors or under the same positive selection?).

For the visualisation of the risk alleles distribution I chose to use the stacked bar plot. The main reason is the fact that all the distribution for the different regions are in one figure and quick visual inspection. If we use the world map with the normal bar plot or pie chart we still would have to remember the individual percentages to compare the distributions. Although in the case of the world map, one could argue it would feel more visual attractive as well as intuitive.

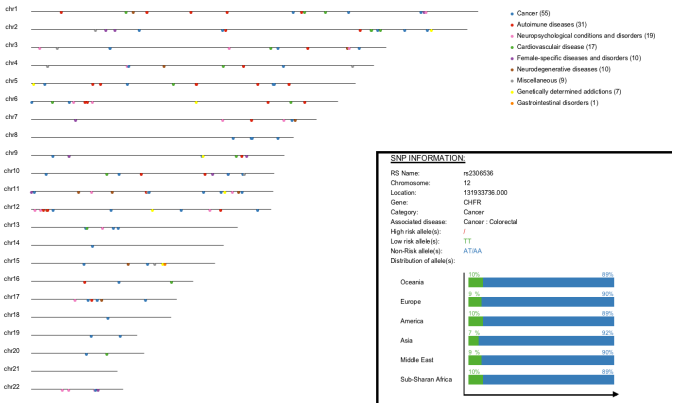
3. Implementation

The first visual design (see question 2.) was implemented with the Processing software.

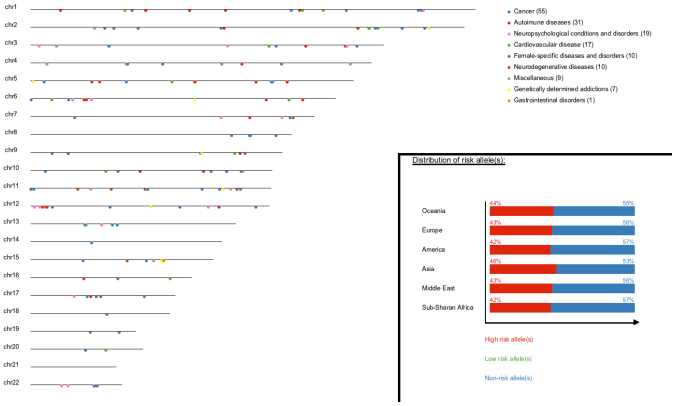
General overview of the visualization:



Interaction 1: Hover over SNP on the chromosome will reval all the information about the SNP as well as the distribution of the risk alleles for the different continents.



Interaction 2: Hover over disease category will reveal the suggested distribution of risk alleles for disease-associated SNPs that belong to this category.



4. Insights

A) Localisation of disease-associated SNPs in the human genome:

Chromosome 21 doesn't contain any disease-associated SNPs and is therefore the only chromosome that doesn't harbour a cancer-associated SNP. It is no surprise that cancer has the most associated SNPs since the distortion of a lot of cellular pathways can lead to cancer. The spectacular finding here is that chromosome 21 doesn't harbour any disease-associated SNPs. Is there a potential reason why no disease-associated genes are found on chromosome 21, when we know trisomy 21 causes the syndrome of down? Another hypothesis is that the end of chromosome 8 could potentially play an important role in the regulation of the human excretion since it system contains 4 SNPs in 4 different genes in close proximity of each other that are associated with cancers that manifest in the excretion systems of humans. Lastly we observed that SNPs associate with autoimmune disease are only found on the 12 first chromosomes with the exception of chromosome 2. What could be the reason that we can only find SNPs at these chromosomes and not at the other?

B) Expected distribution of risk alleles for an SNP belonging to a certain disease category

Generally speaking, the distribution doesn't differ much between the different regions in the world when looking to the expected distribution of risk alleles for an SNP belonging to a certain disease category. For cardiovascular disease this a strange finding since it is commonly viewed as a typical western conditions. This can give us the incentive to investigate if environmental and lifestyle factors are actually the reason why cardiovascular disease like stroke and myocard infarct are more observed. When looking to the individual SNPs belonging to the disease category we observe that the distribution of risk alleles strongly differs between the different cardiovascular disease-associated SNPs. This finding suggests that it's better to look to the individual distribution than making an expected distribution.

Risk alleles for female specific disorders seems to be less observed in Asia, even when we look to the distribution of risk alleles for the individual SNP. What could be the reason that Asian people carry les of this risk-associated variant?

C) Distribution of risk alleles for individual disease-associated SNPs:

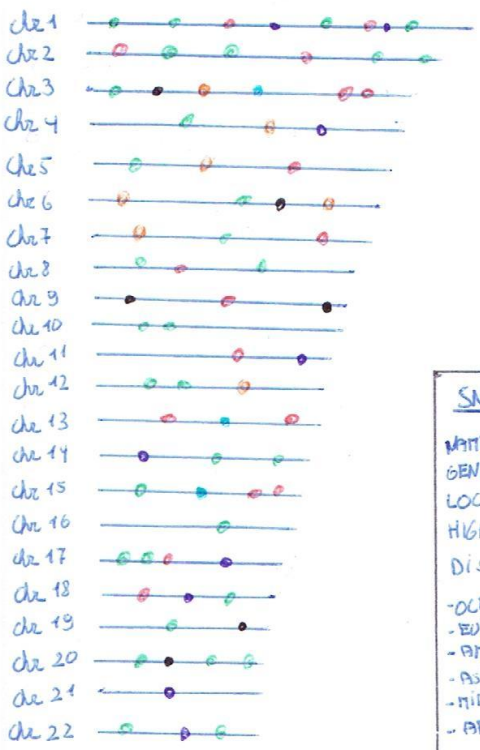
The distribution of the risk alleles for the SNP (gene is NTF3) associated with ADHD indicates that people living in the Middle–East (29%) and Europe (31%) have a higher chance of developing ADHD, while Asian people have the least chance. It would be interesting to study the prevalence of people indicated with this disorder for different regions and see if both findings correspond with each other. This could possibly reveal that we in the western world focus more on this kind of "learning" disorders. High risk alleles for other "learning disabilities" like autism seem to be less prevalent in the Asian people, this could possibly be a reason why Asian students are considered better and smarter students. A higher percentage (30%) of Asian people carry the high develop alcoholism in comparison with the other studied regions. If the prevalence of alcoholism in indeed higher in Asia, this would seem contradictory since around 40% of the Asian people also witness alcohol flush reaction, where there is a heavier showing of the acetaldehyde-based side effects ("Hangover feeling"). This would actually enable the people to quickly stop drinking and giving them an aversion for drinking alcohol. Lastly we observe that a high percentage of the Asian people (almost 90% of the people) carry an allele variant that increases the risk of developing heroin addiction. In the rest of the world, around 80% carry this gene. Since heroine and opium are products mainly stemming from Asiatic countries like Vietnam and Thailand, it would have seemed more logical to have a lower percentage of this allele in Asia from an evolutionary standpoint. What could be the reason that the high percentage in the Asian population is maintained?

5. Screencast

You should also make a screencast (max 5 minutes) in which you demonstrate what the tool does. See here for one possible way to do this: http://www.labnol.org/software/create-youtube-screencast/27936/

Screencast can be found at
https://www.youtube.com/watch?v=Pk3KsCWx9ac

LAYOUT



- CANCER (#)
- AUTOIMUNE (#)
- CARDIOVASCULAIR (#)
- NEURODEGENERATIVE (#)
- GASTROINTESTINAL (#)
- MISCELLANOUS (#)

. . . .

VISUALISATION OF DISEASE ASSOCIATED SNP'S AND THE DISTRIBUTION OF THEIR RISK ALLELES

AUTHOR: SEBASTIAAN VANOUTVEN

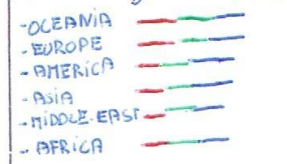
DATE: 23-5-16

TASK: DESIGN VISUALISATION

SHEET NUMMER: 1

SNP INFORMATION

NAME ...
GENE ...
LOCATION
HIGH/LOW/NON-RISK ALLELES ...
DISTRIBUTION



OPERATIONS

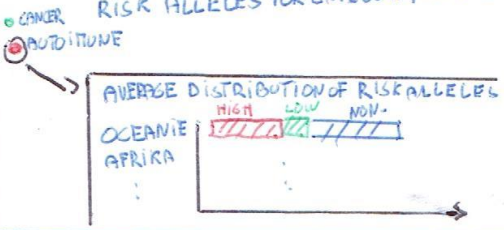
① HOVER OVER SNP ON CHROMOSOME MAP

=> INFORMATION OVER THE SNP
+ distribution OF RISK ALLELES

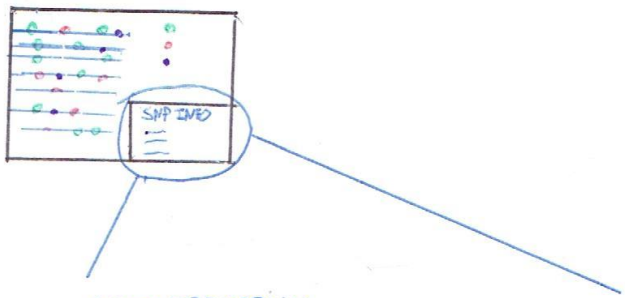
SNP INFORMATION
NAME ....
GENE ....

② HOVER OVER DISEASE CATEGOR

=> DISPLAY AVERAGE DISTRIBUTION OF RISK ALLELES FOR CATEGORY OF DISEASE

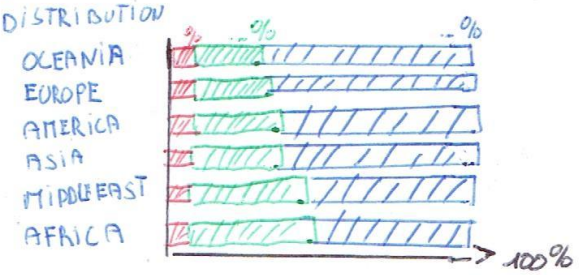


FOCUS/ZOOM



SNP INFORMATION

NAME:
CHROMOSOME:
LOCATION:
CATEGORY:
ASSOCIATED DISEASE
HIGH RISK ALLELE
LOW-RISK ALLELE
NON-RISK ALLELE
DISTRIBUTION



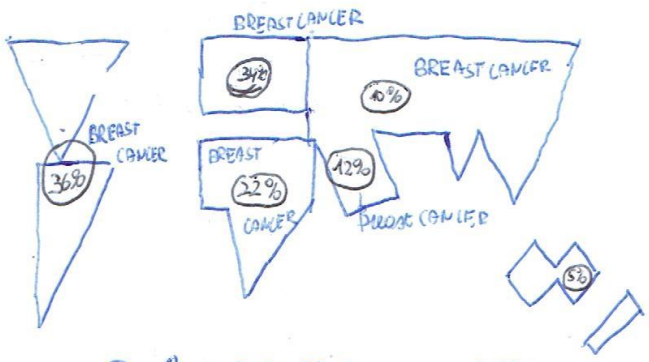
DISCUSSION

⊕ · LOCALISATION OF EACH SNP ON CHROMOSOME
- CLASSIFICATION ACCORDING TO TYPE OF DISEASE
- DISTRIBUTION OF ALLELES FOR EACH SNP
- POTENTIAL PATTERNS IN LOCALISATION FOR CATTEGORY OF DISEASE CAN BE OBSERVED
- STACKED PLOT OFFERS MORE THAN WORLD MAP IN DISPLAYING % FOR each TYPE ALLELE
- MOST DISEASES ONLY GIVE SNP associaties · CLUSTERING PER DISEASE NOT USEFUL

⊖ · NO OVERVIEW PER DISEASE
- NOT POSSIBLE TO FIND MOST SIGNIFICANT DIFFERENT DISTRIBUTION OF ALLELES FOR INDIVIDUAL
- NOT USABLE ON DATASETS CONTAINING > 1000 SNPS
- SNP'S IN NEIGBOORHOOD WILL OVERLAP ON CHROMOSOME MAP

LAYOUT

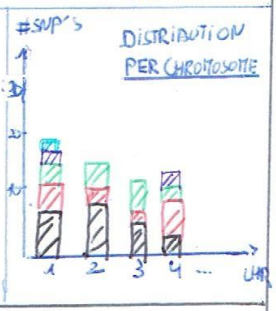
☑ CANCER ☐ AUTOIMMUNE DISEASE ☐ CARDIOVASCULAR
☐ GASTROINTESTINAL ☐ MISCELLANOUS



◯ = % OF POPULATION WITH RISK ALLELE FOR DISEASE

INFORMATION SNP 1
NAME ...
GENE ...
LOCATION ...
HIGH / LOW / NON-RISK ALLELE
DISTRIBUTION

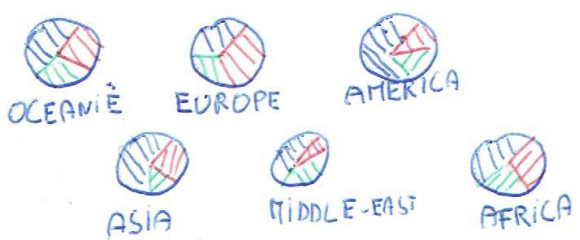
(INFORMATION SNP 2)
NAME ...
GENE ...
LOCATION ...
HIGH / LOW / NON-RISK ALLELES
DISTRIBUTION
(CAN ALSO BE EMPTY)



FOCUS/ZOOM

INFORMATION CONTRIBUTING SNP 1

NAME: ...
CHROMOSOME: ...
LOCATION: ...
ASSOCIATED DISEASE: ...
HIGH RISK ALLELE:
LOW RISK ALLELE:
NON-RISK ALLELE:
DISTRIBUTION OF ALLELES



VISUALISATION OF DISEASE ASSOCIATED SNPS AND THE DISTRIBUTION OF THEIR RISK ALLELES

AUTHOR: SEBASTIAAN VANUYTVEN
DATE: 23-5-16
TASK: DESIGN VISUALISATION
SHEET NUMBER: 2

OPERATIONS

(1) CHECKBOXES TO CHOSE DISEASE FROM DISEASE CATEGORY. DEPICTS BOBBLE (SIZE ~ %) ON WORLD MAP WITH % OF POPULATION CARRYING A RISK ALLELE FOR DISEASE. ALSO THE INDIVIDUAL INFO FOR EACH CONTRIBUTING SNP IS SHOWN + ALSO DISTRIBUTION RISK ALLELE

☐ CANCER
  ☐ BREAST
  ☐ PROSTATE
  ☐ LIVER
  ...

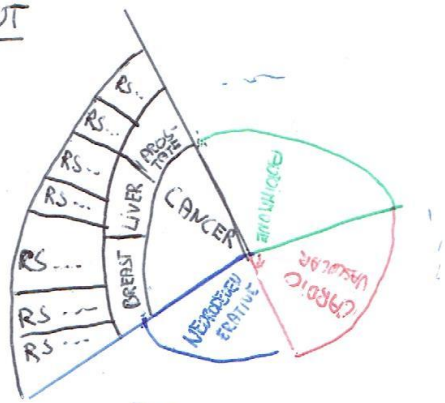
MULTIPLE DISEASE CANNOT BE DEPICTED DUE TO POTENTIAL OVERCROWDING WORLDMAP

DISCUSSION

⊕ QUICK IDENTIFICATION OF CONTINENT WITH HIGHEST %
- MAP WILL NOT GET OVERCROWDED DUE TO ONE DISEASE AT THE TIME
- INFO ON LOCALISATION OF SNPS (WHICH CHRI
- WORLD MAP IS MORE INTUITIVE THEN THE WORDS FOR THE CONTINENT

⊖ - NO DISTRIBUTION OF ALLELE DISTRIBUTION PER DISEASE CATEGORY
- INFO ON INDIVIDUAL SNP → NEED TO KNOW WITH WICH DISEASE ASSOCIATED
- NO CHROMOSOMAL OVERVIEW, GIVES MORE INFORMATION
- PIE DIAGRAMS

LAYOUT



VISUALISATION OF DISEASE ASSOCIATED SNP'S AND THE DISTRIBUTION OF THEIR RISK ALLELES

AUTHOR: SEBASTIAAN VANUYTVEN

DATE: 23-5-16

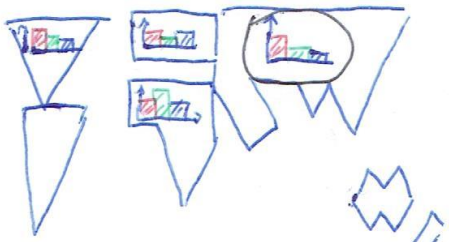
TASK: DESIGN VISUALISATION

SHEET NUMBER: 3

SNP INFORMATION

NAME:
CHROMOSOME:
LOCATION:
CATEGORY:
ASSOCIATED DISEASE:..
HIGH RISK ALLELE
LOW RISK ALLELE ..
NON-RISK ALLELE

DISTRIBUTION OF RISK ALLELES



O = 200M

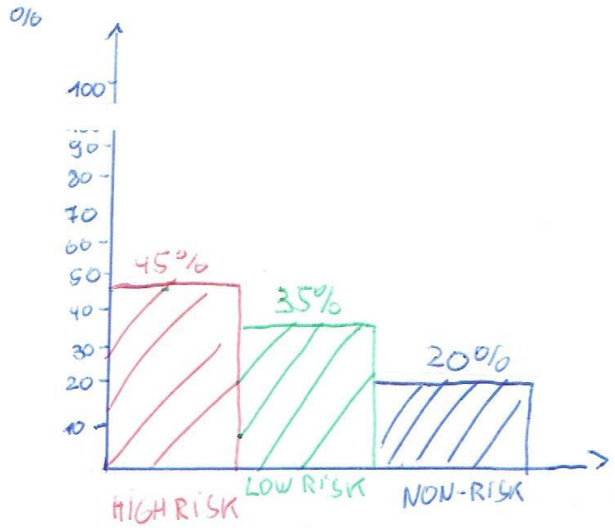
OPERATIONS

(1) HOVER OVER DISEASE CATEGORY (outer circle)
=> EXPECTED DISTRIBUTION OF RISK ALLELES FOR THIS CATEGORY + SNP INFO = BLANK

(2) HOVER OVER DISEASE (MIDDLE CIRCLE)
=> EXPECTED DISTRIBUTION OF RISK ALLELES FOR THIS DISEASE
+ SNP INFO = BLANK

(3) HOVER OVER INDIVIDUAL SNP (OUTER CIRCLE)
=> DISTRIBUTION OF RISK ALLELES
+ DISPLAY ALL SNP INFORMATION

FOCUS/200M



DISCUSSION

(+) DISTRIBUTION OF ALLELES PER DISEASE (CATEGORY) AND INDIVIDUAL SNP
- OVERVIEW OF ALL SNP'S ASSOCIATED WITH PARTICULAR DISEASE (CATEGORY)
- WORLDMAP IS MORE LOGICAL TO USE THEN TERMS "OCEANIA", EUROPE
- QUICK IDENTIFICATION, WHICH DISEASE (category) HAS THE MOST SNP'S ASSOCIATED

(-) NO INFO OVER LOCALISATION SNP ON THE CHROMOSOME
- PIE-FIGURE CAN BECOME OVERCROWDED QUICKLY
- BARPLOT ON MAP IS NOT THE FAST WAY TO COMPARE DISTRIBUTIONS
- MOST DISEASES IN IN DATA HAVE 1 SNP ASSOCIATED => CLUSTERING PER DISEASE USEFUL?