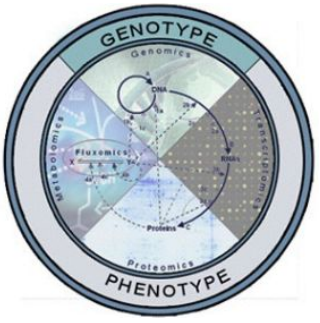
Coralie Visinand Christian Saliba Axel Mermin **2017**

Supervisor : Reyhan Sönmez Flitman

Genome-Wide Association Study (GWAS)

Background:



Every human being, and more generally every living thing, has his own DNA sequence. We call this sequence genotype. Genotype can determine almost every character of a person, and changes between every individual. Every mutation in this sequence can lead to a change in the way to be of a person. DNA is used by the cells to obtain proteins, which in turn synthesize the other substances present in our body. We call the ensemble of this substances the metabotype. The metabotype in turn affect what we call the phenotype, which is the set of observables characteristics, like the colour of the eyes, the colour of the skin and many others, even all the diseases can be grouped under the phenotypic characters. So, we can see that a variation of the genotype can affect the metabotype, which affects the phenotype, taking in some cases to diseases.

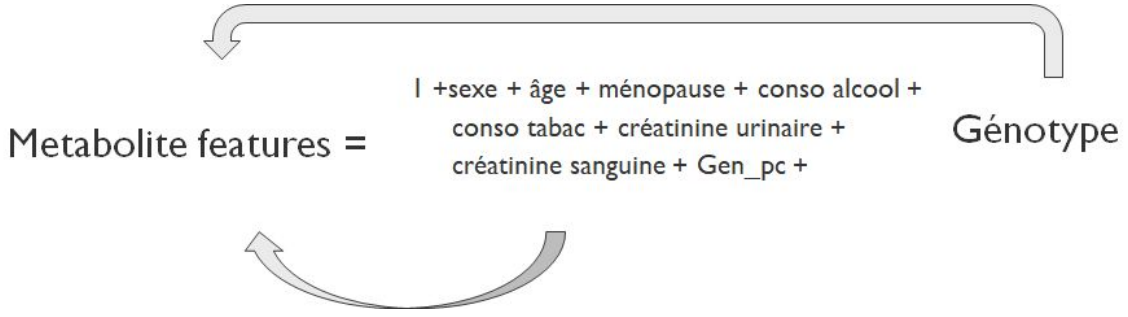
Goal

The goal of GWAS is to identify if there is any significant connection between one or more known mutations in the genotype and the risk factor of a determinate disease.

Dataset: We obtained our data thanks to the COLAUS project. Colaus is a large study based on the population of Lausanne, it involves more than 6'000 persons.

Methods

To reach our goal we did many regressions between each known mutation site (SNP) of the genotype (in our case only of the chromosome 5) and each metabotype feature (taken from the urine). Regression is a statistical technique that allow us to determine the influence of a value in rapport to another one. In our chase we want to determine the influence of the mutations of SNPs and the metabolites expression.
After obtaining all the regressions results, we took the three most significant ones and we checked if this three most significant results are related to a specific gene. The last step is to relate the found gene with a disease or with a phenotypic disequilibrium, thanks to the huge database of researches present on internet.



Mathematical tools:

We used Matlab to create all the plots, to manipulate all the matrices containing the data and mainly to effectuate our about 200 million regressions.

Linkage disequilibrium :

If we do not adjust the regression threshold, it means that all our SNPs are independent. This is not the case biologically. This is called "*Linkage disequilibrium*".

Bonferroni correction :

Generally in a regression we use a threshold of 5% but this threshold is not adjusted for any regression result. Bonferroni correction is just calculated by the threshold divide by the number of test. By convention, the number of 1'000'000 independent SNPs throughout the genome.

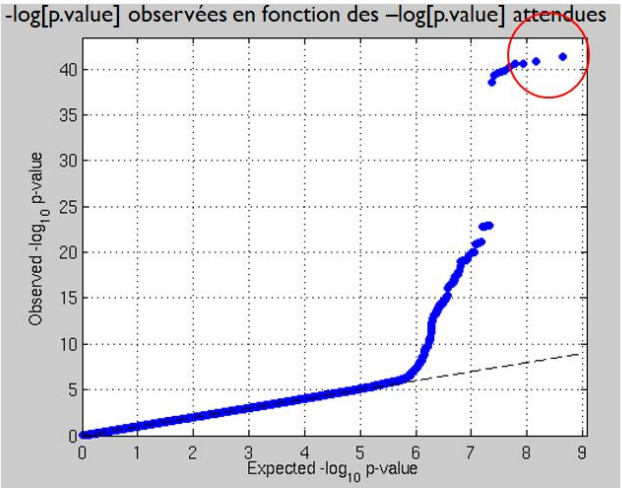
We used the correction of bonferroni to modify our threshold precisely because of this "linkage disequilibrium" and this allowed us to adjust our threshold correctly.

We obtain : $$P_{Bonferroni} = \frac{0.05}{1'000'000} = 5 * 10^{-8}$$

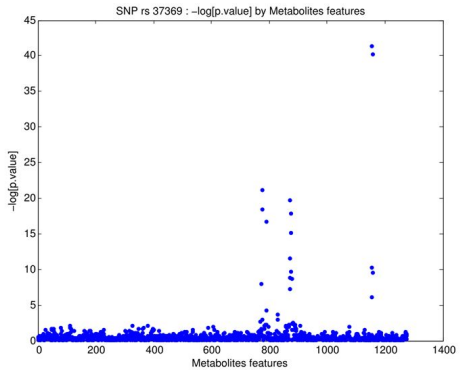
This is our new threshold.

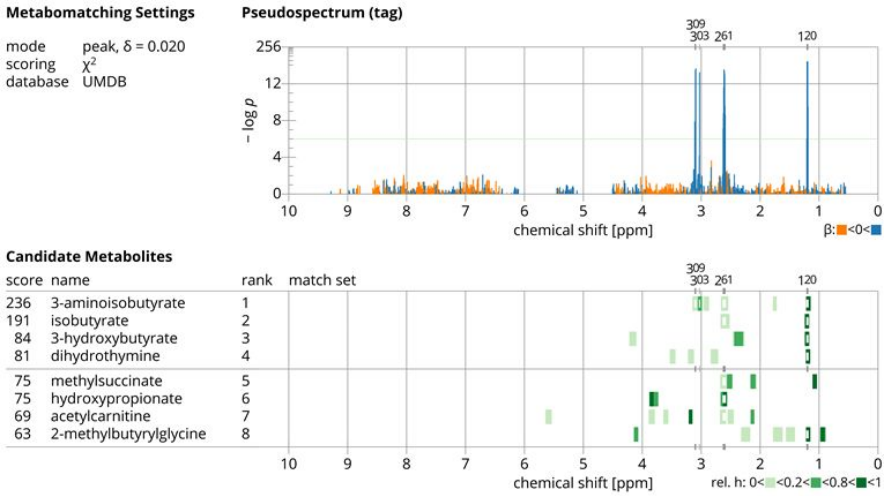
Results

This is the QQplot transformed in *-log* of our regression results.



After taking the three most significant results we found the SNPs to which they are related: rs37369, rs37370, rs40200. We took this three SNPs and we created pseudospectras like the figure aside and then we used the metabomatching method.





The metabomatching method is compared with NMR spectra of different metabolites and give us the ones which fit best.

We noticed that they was all related to SNPs presents on the same gene, AGXT2. This gene is related to the synthesis of 3-aminoisobutyrate, which is an enzyme

Conclusion for medical research:

We checked on internet if the gene AGXT2 has already been studied. Effectively we found many studies which relate this gene to diseases or to an increased risk to develop a determinate disease. For example on the platform NCBI we found some studies that relate that gene to cardiovascular diseases, like coronary heart disease or carotid atherosclerosis.
So we can conclude that GWAS can effectively help many people in the medical field, simplifying the diagnosis of a determinate disease. Maybe in a not so far future we will be able to diagnose a specific disease without invasive exams, just taking a sample of urine from the patient.