COMBINING GWAS AND BIOLOGICAL NETWORKS : FEATURES SELECTION FOR BREAST CANCER

Pierre-François Saunier

MINES ParisTech - CBIO

Supervisor : Chloé-Agathe Azencott

Mots clés : GWAS, biological networks, mixed models, breast cancer.

Abstract :
This work deals with the identification of genetic factors that increase an individual's susceptibility to breast cancer. Following the appearance of human genotyping projects, GWASs (Genome Wide Association Studies) seem promising. The latter make it possible to study statistical correlations between thousands of single nucleotid polymorphisms (SNPs) and a disease, among a cohort of several thousand individuals. However, they face difficulties, namely in the matter of dimensionality and interpretability. Moreover, GWASs take little account of genetic phenomena leading to disease. Thus, it seems relevant to combine these studies with biological networks based on pre-established genetic interactions, in order to highlight factors whose importance cannot be exhibited by formers statistical studies.

1 Introduction

The goal of my internship was to apply Héctor Climente's work to a new dataset. During his PhD, he built a pipeline to run statistical association studies on datasets and then use the returned results in biological networks to boost the performances. As for the dataset, it is one Asma Nouira is working on at the CBIO.

1.1 GWAS

1.1.1 Principles

The genetic material of an individual, his genome, is composed of essential base pairs called nucleotides. There are four of them : adenine A, thymine T, guanine C and cytosine C. These monomers gather in long polymers of deoxyribonucleic acid (DNA), a double-stranded molecule. The two strands are linked by hydrogen bonds that can only be established between either adenine and thymine (A <-> T) or between cytosine and guanine (G <-> C). Therefore, knowing one strand is equivalent to knowing the two of them. This DNA is split into separate pieces called chromosomes. Humans have 23 pairs of chromosomes. One pair concerns sex chromosomes ; the twenty-two other are autosomal chromosomes. Among each pair, one chromosome comes from the father, the other comes from the mother. The expression of the genome of an individual as observable traits, and also the influence of the environment, is called the phenotype.

Many genetic variations can occur between two genotypes, such as deletions, insertions or inversions. However, the most common variation is called single-nucleotide polymorphism (SNP). It is a substitution of a single nucleotide at a precise locus (see Figure 1) ; this substitution has to be present in more than 1% of the population to be considered a SNP. They are quite frequent in the human genome : there is one SNP every two hundred nucleotides. For most SNPs, only two alleles are present in the population ; they are said bi-allelic, and are characterized by their minor allele frequency (MAF).

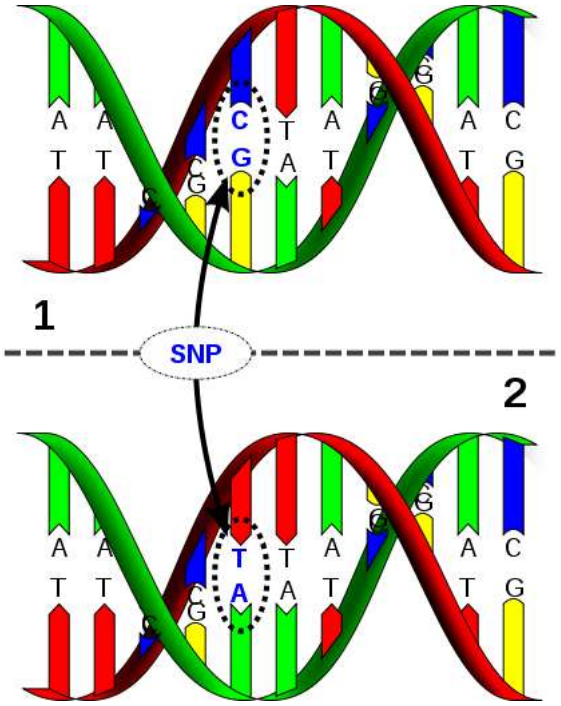


Figure 1: *A schematic SNP*

The invention of DNA chips about twenty years ago made it possible to map the whole genome of an individual, by mapping those regularly disposed SNPs. Even if only a few percentage of the nucleotides are mapped, the notion of Linkage Disequilibrium (LD) makes this mapping relevant. LD is in particular the idea that two close nucleotides have correlated consequences, since they are probably among the same gene and participate in coding the same protein for example. Therefore, mapping only regularly placed SNPs is a good way of having a general overview of which parts of the DNA are significant. When SNPs of interest have been discovered, a fine-mapping of their surroundings will have to be done in order to establish which precise nucleotides are responsible for what we observe.

Genome Wide Association Studies (GWAS) come into play alongside DNA chips. In order to better understand the links between genotype and a certain trait, statistical tests can be lead thanks to the mapping of a large cohort of individuals. For a binary trait, which is typically a disease ($y = 1$ for case or $y = 0$ for control), the first idea is to run a χ^2 test for each SNP, the resulting p-value being an indicator of its statistical signification with regards to the disease. Going further to take into account both the genotype and additional covariates (age, sex ...) is mostly done through a logistic regression. For a given SNP and a binary phenotype :

$$logit(p) = \mu + \beta_i * SNP_i + covariates * weights \tag{1}$$

where p is the probability of having the disease knowing the genotype : $p = P(y = 1|SNP_i, covariates...)$.

This is not the p-value of the SNP n°i ! The latter is calculated by posing the null hypothesis $(H_0) : \beta_i = 0$ (i.e. the SNP does not influence the phenotype) and the alternative hypothesis $(H_1) : \beta_i \neq 0$ (i.e. the SNP has a non-zero weight and therefore influences the phenotype). Indeed, the goal of GWAS is to highlight that some loci are responsible for a certain trait.
We will see in that some good statistical test appear from this regression.

1.1.2 Difficulties

Hundreds and thousands of GWAS have been conducted since the early 2000. Many regions of interest have been discovered, resulting in a better understanding of biological phenomenons. However, GWAS are not as promiseful as they seem to be, since they still face many difficulties.

Population structure
A major problem in conducting GWAS is the notion of population structure. This is the result of non-random mating between individuals from the cohort, explained by a former geographical separation followed by genetic drift. Different groups emerge from the cohort, with different common ancestors. Since the alleles are not randomly distributed among each group and among the cohort, false discoveries might appear. Population structure is therefore an important confonding factor in GWAS, i.e. a factor that might lead to spurious associations. An important stake of GWAS is to manage to deal with this problem. We will see in (2.1) how this is done.

Dimensionality
This comes with a second difficulty ; these studies rely on huge dataset. The order of magnitude of the cohort size is ten thousands of indidivuals, and up to one million SNPs can be mapped. Thus, this is a complex features selection problem. The number of samples we dispose of is really low compared to the number of SNPs being tested, therefore limiting the statistical power of the test. This particularly induces a risk of overfitting [1].

Statistical power
The main disillusion with GWAS is that they present low statistical power [2]. When working with mendelian disease, we expect to quickly find the only gene or the only nucleotide responsible for the whole disorder. Complex diseases studied through GWAS cannot be explained by a single variation, and each SNP or gene highlighted only account for a few percent of the susceptibility to the disease. In fact, despite the increasing calculation power evereyday available, GWAS alone won't be sufficient to explain such complex diseases [3].

Interpretability
Lastly, what lacks in GWAS is the comprehension of biological and genetic phenomenons that result in the disease. Even if a loci is correctly identified, GWAS alone cannot explain the mechanisms responsible for its role. This can only be brought by a deeper understanding of biological interactions, molecular chemistry and genetics.

1.2 Biological networks

An approach chosen at the CBIO to compensate this is to combine GWAS with biological networks. They are networks whose nodes are either SNPs or genes, and whose edges represent proved biological interactions.
Two genes interact together when they code for proteins that physically interact together through protein-protein interaction for example. For SNPs networks, three types of networks exist. Let's take a SNP X among a gene G. First, one can link X only to its direct neighbors. One could also link X to all SNPs that lie within a certain distance of gene G. Lastly, one could link X to all SNPs that lie within a certain distance of all genes that interact with gene G. To mention it, the network SConES (Azencott et al. 2013) works at the nucleotides level.

Boosting GWAS results with biological networks consists on considering that SNPs or genes linked with the disease are not spread out across the genome but rather connected through biological interactions networks. This complexity is not necessarily detected by statistical tests. The use of networks could help highlighting the role of some loci by trading a bit of statistical significance against biological relevance. They rely, to some extents, on the principle of guilt-by-association. A gene with a high p-value (i.e. weakly associated to the disease) but connected with many significant genes could be displayed

by a network as a new gene of interest. See Figure 2 : gene D is at first not significant enough, but its interactions with other significant genes make it a new candidate. The algorithms mainly consist on searching subnetworks with a high density of some computed score for each gene.

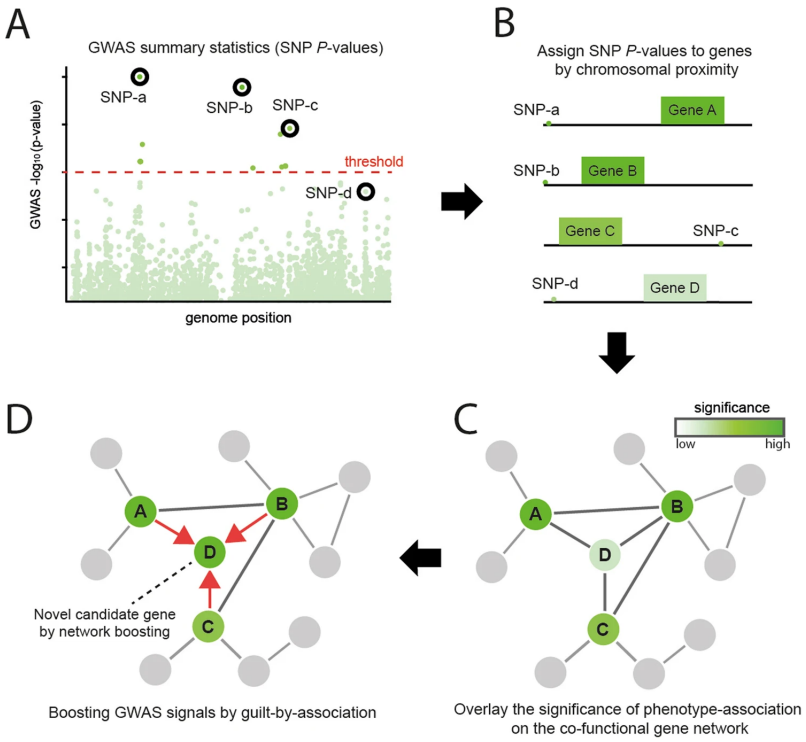


Figure 2: *Boosting GWAS through biological networks*

1.3 Triple-negative breast cancer

The disease I have worked on is triple-negative breast cancer (TNBC). It is one of several forms of breast cancers, whose tumors lack the expression of three proteins : ER (estrogen eceptor), PR (progesteron receptor), and HER2. These cancers represent 10 to 15% of breast cancers, but they are really agressive and tend to be more common among young women [5]. In fact, they grow and spread fast, without the existence of a clear targeted treatment. They are also less detectable by a mammography, have a higher rate of recidivism and a higher rate of familial background.

1.4 The dataset

The dataset I have worked on is a part of the OncoArray initiative ; by developping a custommed SNP chip fine-mapped around interesting regions for cancer, the project aims at better identifying cancer risk loci.

The dataset is composed of the mapping of 314,314 SNPs for 28,281 individuals. 13,846 are cases and 14,435 are controls. These women come from six countries : USA, Australia, Denmark, Cameroon, Uganda and Nigeria. Through a principal components analysis (PCA) of the cohort (Figure 3), we quickly see that the dataset presents a strong population structure.

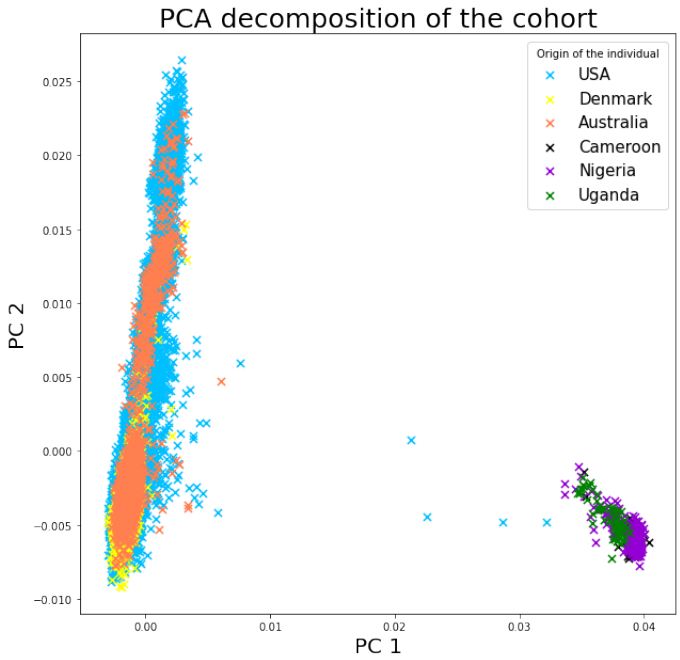


Figure 3: *Two components PCA of the cohort*

2 Methods

2.1 Computing p-values

Notations Quick remark to prevent some potential confusion in the notation. In the following paragraphs, $\mathbf{g}$ represents a vector of genotype for a precise SNP, whose corresponding weight is β, and each line of the vector corresponds to an individual. So when $\mathbf{g}$ is used, we are in the case of a monovariate regression. Whereas $\mathbf{G}$ represents the whole matrix of SNP, with the column weights $\mathbf{b}$: each line still corresponds to an individual, and each column corresponds to a SNP, i.e. $\mathbf{G} = (\mathbf{g}_1;\ ...; \mathbf{g}_S)$. When $\mathbf{G}$ is used, we are running a polyvariants regression.
S is the number of SNPs being tested, N is the number of indidivuals in the cohort.

2.1.1 Classical methods

As mentionned in the introduction, a classical logistic regression is run in GWAS. This regression is run for each SNP. Let's therefore fix one particular SNP whose p-value we look for, and write :

$$logit(p) = \underbrace{\mathbf{g}\beta}_{\text{Genotype effect}} + \underbrace{\mathbf{X}\alpha}_{\text{Covariates}} + \underbrace{\varepsilon}_{\text{noise}} \tag{2}$$

where $p = \mathcal{P}(y = 1 \mid \mathbf{g}\ ,\ \mathbf{X})$; $\mathbf{g}$ is the phenotypes vector (i.e. the value of the particular SNP for each indidivual), $\beta \in R$ is the weight associated to this SNP, $\mathbf{X}$ is the covariates matrix, α the associated weights and $\varepsilon \sim \mathcal{N}(0_N, \sigma_e^2 \mathbf{I}_N)$ is noise. This equation is a vectorial equation (one line for each individual). This can also be expressed as

$$y = logit(p) \sim \mathcal{N}(\mathbf{g}\beta + \mathbf{X}\alpha,\ \sigma_e^2 \mathbf{I}N)$$

It follows that the marginal log-likelihood of this model is

$$\log \mathcal{L}(\beta, \alpha, \sigma_e^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}(y - \mathbf{g}\beta - \mathbf{X}\alpha)^\top (y - \mathbf{g}\beta - \mathbf{X}\alpha) \tag{3}$$

And the maximum likelihood estimators are

$$\hat{\beta}, \hat{\alpha}, \hat{\sigma_e^2} \in \underset{\beta,\alpha,\sigma_e^2}{\operatorname{argmax}} \mathcal{L}(\beta, \alpha, \sigma_e^2)$$

The main goal of GWAS being to determine whether our chosen SNP has an effect on the phenotype or not, the natural null hypothesis is $(H_0) : \beta = 0$ which will be tested against $(H_1) : \beta \neq 0$. Therefore :

$$(H_1) : y \sim \mathcal{N}(\mathbf{g}\beta + \mathbf{X}\alpha,\ \sigma_e^2 \mathbf{I}_N)$$

$$(H_0) : y \sim \mathcal{N}(\mathbf{X}\alpha,\ \sigma_e^2 \mathbf{I}_N)$$

The first way of computing the SNP's p-value is based on Wald's test. This is used in the software Plink.

The most precise and therefore the most computationally demanding statistical test is the likelihood ratio of these models :

$$LR = \log \frac{\mathcal{L}(\hat{\beta}, \hat{\alpha}, \hat{\sigma_e^2})}{\mathcal{L}(0, \tilde{\alpha}, \tilde{\sigma}_e^2)} = \log \mathcal{L}(\hat{\beta}, \hat{\alpha}, \hat{\sigma_e^2}) - \log \mathcal{L}(0, \tilde{\alpha}, \tilde{\sigma}_e^2) \tag{4}$$

where $\tilde{\alpha}$ and $\tilde{\sigma}_e^2$ are the maximum likelihood estimators for the null hypothesis.

In fact, Wilks' Theorem guarantees that $2LR$ asymptotically (i.e. for a large cohort of individuals) follows $\chi^2_{(1)}$. We can now assign to this SNP the p-value we were looking for :

$$p_{val} = \int_{2LR}^{+\infty} \chi^2_{(1)}(x)dx = 1 - F_{\chi^2_{(1)}}(2LR) \tag{5}$$

As mentionned in 1.1.2, and noted in 1.4, population structure is a cause of spurious association, when not correctly adressed. The classical way of adressing it is to add the principal components of the data as covariates in the regression (2).

2.1.2 Linear Mixed Models

Another idea to adress population structure is to use mixed models. The main idea behind them is to add in the regression (2) a noise u whose variance is not the identity matrix, and more generally not diagonal. It will express the dependency existing between each SNP. To recall, population structure induces a non-random distribution of the genetic variants, and therefore a correlation between them. Let's write :

$$logit(p) = \mathbf{g}\beta + \mathbf{X}\alpha + u + \varepsilon \tag{6}$$

Where the main addition is the random vector

$$u \sim \mathcal{N}(0, \sigma_g^2 \mathbf{R})$$

As previously, this can be expressed as

$$y \sim \mathcal{N}(\mathbf{g}\beta + \mathbf{X}\alpha,\ \sigma_g^2 \mathbf{R} + \sigma_e^2 \mathbf{I}_N) \tag{7}$$

Several definitions of $\mathbf{R}$, called the genetic relatedness matrix, can be found. The widely-used estimate is

$$\mathbf{R} = \frac{1}{S} \mathbf{G}\mathbf{G}^\top \tag{8}$$

The likelihood of the model can now be computed, and so can the likelihood ratio : the calculations

are detailed in [6].

Note that (7) is a mono-variant regression (we only consider one SNP). Adding this noise is in fact equivalent to considering a multi-SNP regression $\mathbf{y} = \mathbf{Gb} + \mathbf{X}\alpha + \varepsilon$, but with weights $\mathbf{b} \sim \mathcal{N}(0, \frac{\sigma_g^2}{S}\mathbf{I}_S)$, whose exact values are not known but are modeled as random weights driven by a common distribution. By doing this, the definition (8) appears naturally. See these notes for more details.

What is interesting with this model is that is has proven really efficient. According to [7], whereas we could add a few principal components as covariates in (2), linear mixed models are able to capture the entire eigen-spectrum of the genetic similarity matrix (the eigen-vectors of the genetic relatedness matrix $\mathbf{GG}^\top$ are the principal components of the dataset $\mathbf{G}$). It has been shown that population structure is way more correctly adressed with mixed models.

The implementation I used to run those mixed models is the python package `FaST-LMM`. It stands for Factored Spectrally Transformed Linear Mixed Models.

2.2 Mapping SNPs to genes

The previous way of running GWAS returns per-SNP p-values. Since we aim at working with gene-level networks, we need to transform these p-values into genes p-values, i.e. to affect to each gene a p-value that is coherent with its SNPs. I used the software VEGAS2 to do this.

The principle is described in [8] : given the n SNPs of a gene (and eventually neighbors SNPs within some optionnal distance), a statistical test following $\chi^2_{(n)}$ is computed. Then, a reference population is used to compute a set of simulated data, and the same statistic is calculated for each set. Lastly, if m simulations have been done, and for r of them the simulated statistic exceeds the observed statistic, then the gene is affected the p-value

$$p_{gene} = \frac{r+1}{m+1} \tag{9}$$

The reference population is taken from the 1,000 Genomes project [9] and can be selected from different sets ; European, African, Asian ... or more precise sets ; "British in England and Scotland (GBR)" for example.

One can also specify if all SNPs are used to compute the statistics, or if only a percentage of top most significant SNPs are used. In our case, we only used the top 10% of lowest p-values SNPs, and mapped to each gene every SNP that lies within 50-kb of it.

2.3 Using biological networks

During my internship, I only worked with gene-level networks. The reference database I used to build the networks is the `STRING v11` database presented in [10], which contains both physical and functionnal protein-protein interactions based on previously published litterature or predicted interactions that have been approved by the KEGG (Kyoto Encyclopedia of Genes and Genomes).

As mentionned in 1.2, biological networks methods are essentially searching high-scoring subgraphs. The four methods I used are : Heinz, SigMod, dmGWAS and LEAN. Their functioning is described in [11].

As we will further have a closer look at Heinz, the idea behind this method is to transform each gene's p-value into a score which is positive under strong association and negative otherwise. Then all gene-gene interactions are built according to the STRING database, and the method searches the subnetwork with the highest total score . Since less correlated genes have a negative (or at least low) score, they tend to be excluded from the network, unless they are on the path to highly correlated genes whose scores are high.

3 Results

3.1 GWAS

Running `FaST-LMM` on our dataset returned the following results. The Manhattan Plot of the GWAS is in Figure 4. The threshold computed is a Bonferroni corrected one (ie $\alpha = \frac{5\%}{TESTS} = \frac{0.05}{314\ 000} = 1.6\ 10^{-7}$).

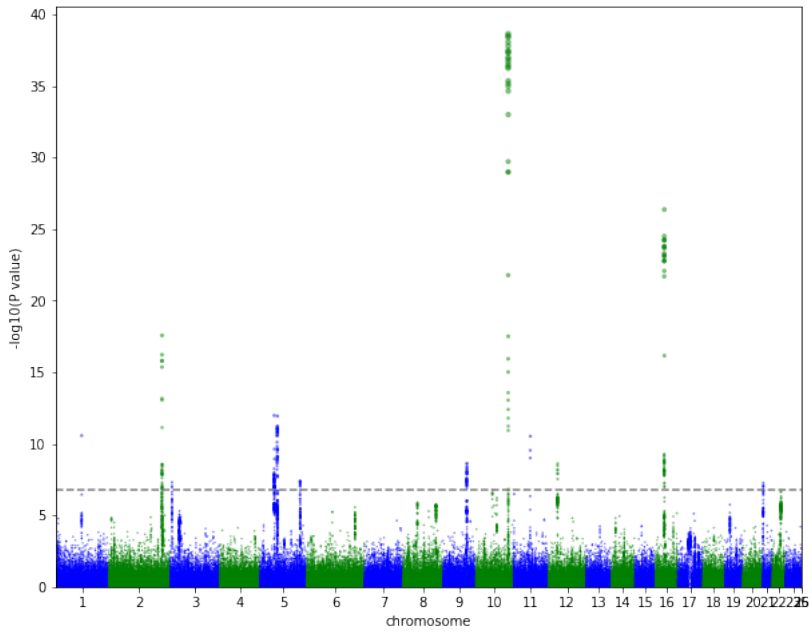


Figure 4: *Manhattan Plot*

The most correlated areas are located among the 10th and 16th chromosomes. To better interpret these p-values, the quantile-quantile plot of the p-values distribution is shown in Figure 5. This plot is used to establish whether two variables are identically distributed. Here, we plotted the observed distribution of p-values against the expected one, a uniform one. Deviation from the line $y = x$ shows that the observed and the expected distributions do not match.

The strong deviation at the end of the plot is what we look for ; we expect that some SNPs are not distributed uniformly among the cohort. They will be the SNPs of interests for TNBC. But our observed distribution strongly diverges from the expected one for the most of it, with a genomic inflation factor $\lambda = 1.124$ (defined by $\lambda = \frac{\text{Median (observed } \chi^2)}{\text{Median (expected } \chi^2)}$, i.e. we run a per-SNP χ^2 association test and we compare it to the expected distribution of this statistic, as explained in [12]). This is totally unexpected since linear mixed models are supposed to correctly adress population structure, therefore reduce the λ closer to 1.0 and fitting the observed distribution more correctly to the theorical one.

In fact, it appears that `FaST-LMM` obtained the exact same results as classical methods that I ran in the same time (with Plink and PCA correction, adding the five first principal components as covariates). It indicated us that there is a stronger population structure hidden among our cohort that `FaST-LMM` could not get rid of. Asma is working on finding what is happening there. It seems that running LMM on each separated groups of Figure 3 correctly adresses population structure, but that it is too strong when the whole cohort is used.

Therefore, here comes a first strong result : linear mixed models were, in our case, not more powerful than classical methods. `FaST-LMM` gave the same results but at the price of a much longer wait time (a few days versus a few hours).

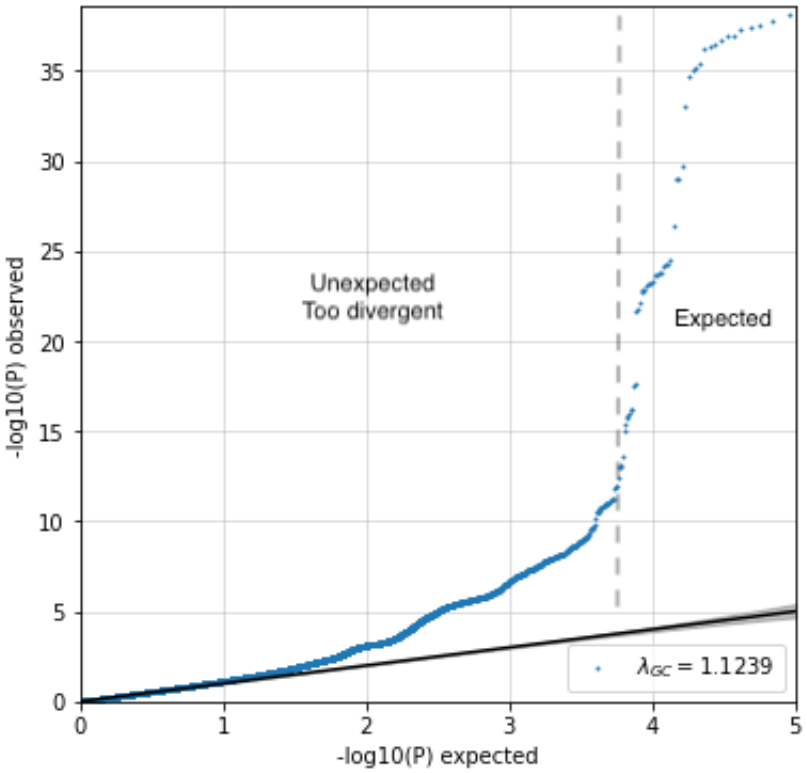


Figure 5: *QQ-plot*

3.2 Biological networks

The file I used is their `protein.links.v11.0.txt`. The interactions are given through protein-protein couples, and the accessory file `protein.aliases.v11.0.txt` allow to convert these to gene-gene couples. I have run five networks with these values ; Heinz (with two differents false discovery rate : 0.5 and 0.05), SigMod, Lean and dmGWAS. The size of the resulting network is greatly variable :

	Heinz 0.05	SigMod	dmGWAS	Heinz 0.5	Lean
Number of nodes	13	232	291	1183	17611

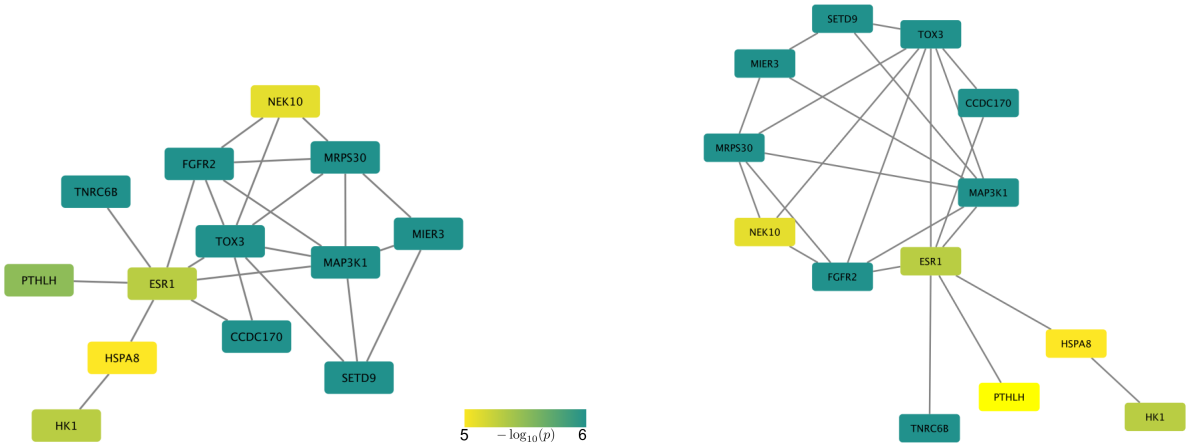
Table 1: Number of genes given by each network

I could not manage to interpret Lean due to how different the returned file is from the other methods. An important result is how much the other four networks overlap :

	Heinz 0.05	SigMod	dmGWAS	Heinz 0.5
Heinz 0.05	13	13	9	13
SigMod		232	50	230
dmGWAS			291	88
Heinz 0.5				1183

The diagonal obviously contains the number of nodes on each network.
There is a core of 9 genes that are returned by every network. And with the exception of dmGWAS, the thirteen genes given by the smallest network (Heinz 0.05) are all present in the others.

Let's have a look at the Heinz 0.05 FDR network (visualization done through Cytoscape, Figure 6):



(a) Organic layout **(b)** Circular layout

Figure 6: *Two visualizations of Heinz (0.05)*

The default configuration of Cytoscape renders the network as shown in **(a)**. The app `yFiles Layout Algorithms` proposes other visualizations styles ; the circular one **(b)** shows the graph from another angle, which can be interpreted differently. A graph structure seems to emerge from **(b)**, with an interconnected core and four *outsiders* genes.

The dark green genes are the ones with the lowest p-values. We understand here the effect of the network. HSPA8 has a p-value ten times higher than TOX3, i.e. it is less significantly linked to the cancer, but it appears in the method thanks to the interactions it has with other genes. This gene has already been pointed out in [13] as a gene of interest for triple-negative breast cancer, despite being ranked 249th (by ascending p-values) in the results of VEGAS2. This reinforces the idea that this approach is relevant.

MAP3K1 5	SETD9 7	TOX3 8	TNRC6B 11	FGFR2 17	MRPS30 18	MIER3 20
CCDC170 37	PTHLH 49	HK1 56	ESR1 57	NEK10 62	HSPA8 249	

Table 2: *Rank of each gene*
(VEGAS2 results sorted by ascending p-values)

The nine common genes to all four networks are all the eight dark green genes, and HK1, i.e.

TNRC6B ; FGFR2 ; TOX3 ; CCDC170 ; SETD9 ; MAP3K1 ; MIER3 ; MRPS30 ; HK1.

The four remaining genes are PTHLH, ESR1, NEK10 and HSPA8. Despite not being well-ranked according to VEGAS2, all of them have already been pointed out as genes of interest for cancers respectively in [14], [15] (this article focuses on ER-positive breast cancer), [16] and [13]. Lastly, [17] also confirms ESR1 as a gene of interest with regards to TNBC. This supports the relevance of the Heinz method, and therefore supports that working with biological networks truly boosts GWAS results. According to our VEGAS2 results alone, we would have missed PTHLH, HK1, ESR1, NEK10 and HSPA8, but they are catched up by the network.

4 Conclusion

This internship was a really exciting time, since the topics were both genetically and mathematically interesting in depth.

As Héctor showed through his thesis, working with biological networks really helps boosting GWAS results, and enhances the selection of loci correlated to triple-negative breast cancer. During the short period of my internship, I managed to run several of the methods selected by Héctor, almost completely fulfilling the expectations of the term. Additional time would be needed to properly analyze the results, although the real work that would need to be done would be a geneticist or biologist's work to understand the mechanisms that involve the genes sorted by the networks.

The limitations of this approach remain the poor interpretability of the results; these algorithms do not explain how these genes are involved in the disease, which does not allow a clear causality to be established. The fact that we found genes that had already been studied elsewhere shows, however, that the GWAS + biological networks approach does not highlight genes at random. What I lacked, since I was only able to run the networks late, is time to interpret the final networks, but it seems to me that this work goes beyond the role of the engineer and is much more biological. Nevertheless, this approach allows to sort out the loci of interest, and it allows to clear the ground for more precise research on the selected loci.

Acknowledgments

I am grateful to my supervisor Chloé-Agathe Azencott for the implementation of this trimester ; for the extra time this supervision has demanded of her ; and obviously for the captivating courses she taught at Mines ParisTech during my first and second years.

I am also thankful to Héctor Climente and Asma Nouira for the help they both provided me, without them being expected to do so. This helped me to better work on the subject, despite the short duration of the internship.

References

[1] Rita M. Cantor, Kenneth Lange, and Janet S. Sinsheimer. "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application". en. In: *The American Journal of Human Genetics* 86.1 (Jan. 2010), pp. 6–22. ISSN: 00029297. DOI: 10.1016/j.ajhg.2009.11.017. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002929709005321.

[2] Jonas Patron et al. "Assessing the performance of genome-wide association studies for predicting disease risk". en. In: *PLOS ONE* 14.12 (Dec. 2019). Ed. by Joseph Devaney, e0220215. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0220215. URL: https://dx.plos.org/10.1371/journal.pone.0220215.

[3] N. Risch and K. Merikangas. "The Future of Genetic Studies of Complex Human Diseases". en. In: *Science* 273.5281 (Sept. 1996), pp. 1516–1517. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.273.5281.1516. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.273.5281.1516.

[4] Tak Lee and Insuk Lee. "araGWAB: Network-based boosting of genome-wide association studies in Arabidopsis thaliana". en. In: *Scientific Reports* 8.1 (Dec. 2018), p. 2925. ISSN: 2045-2322. DOI: 10.1038/s41598-018-21301-4. URL: http://www.nature.com/articles/s41598-018-21301-4.

[5] Kartik Aysola Akshata Desai. "Triple Negative Breast Cancer – An Overview". In: *Hereditary Genetics* (2012). ISSN: 21611041. DOI: 10.4172/2161-1041.S2-001. URL: https://www.omicsonline.org/triple-negative-breast-cancer-an-overview-2161-1041.S2-001.php?aid=13213.

[6] Francesco Paolo Casale. "Multivariate linear mixed models for statistical genetics". en. In: (Nov. 2016). Publisher: Apollo - University of Cambridge Repository. DOI: 10.17863/CAM.13422. URL: https://www.repository.cam.ac.uk/handle/1810/267465.

[7] Gabriel E. Hoffman. "Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions". en. In: *PLoS ONE* 8.10 (Oct. 2013). Ed. by Marie-Pierre Dubé, e75707. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0075707. URL: https://dx.plos.org/10.1371/journal.pone.0075707.

[8] Aniket Mishra and Stuart Macgregor. "VEGAS2: Software for More Flexible Gene-Based Testing". en. In: *Twin Research and Human Genetics* 18.1 (Feb. 2015), pp. 86–91. ISSN: 1832-4274, 1839-2628. DOI: 10.1017/thg.2014.79. URL: https://www.cambridge.org/core/product/identifier/S1832427414000796/type/journal_article.

[9] Ewan Birney and Nicole Soranzo. "The end of the start for population sequencing". en. In: *Nature* 526.7571 (Oct. 2015), pp. 52–53. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/526052a. URL: http://www.nature.com/articles/526052a.

[10] Damian Szklarczyk et al. "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets". en. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D607–D613. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gky1131. URL: https://academic.oup.com/nar/article/47/D1/D607/5198476.

[11] Héctor Climente-González. "Network-guided genome-wide association studies". en. PhD thesis. Feb. 2020. URL: https://www.theses.fr/en/2020PSLEM001.

[12] Silviu-Alin Bacanu, B. Devlin, and Kathryn Roeder. "The Power of Genomic Control". en. In: *The American Journal of Human Genetics* 66.6 (June 2000), pp. 1933–1944. ISSN: 00029297. DOI: 10.1086/302929. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002929707635459.

[13] Jiarui Chen et al. "KEGG-expressed genes and pathways in triple negative breast cancer: Protocol for a systematic review and data mining". en. In: *Medicine* 99.18 (May 2020), e19986. ISSN: 0025-7974, 1536-5964. DOI: 10.1097/MD.0000000000019986. URL: https://journals.lww.com/10.1097/MD.0000000000019986.

[14] Gloria Assaker et al. "PTHrP, A Biomarker for CNS Metastasis in Triple-Negative Breast Cancer and Selection for Adjuvant Chemotherapy in Node-Negative Disease". en. In: *JNCI Cancer Spectrum* 4.1 (Feb. 2020), pkz063. ISSN: 2515-5091. DOI: 10.1093/jncics/pkz063. URL: https://academic.oup.com/jncics/article/doi/10.1093/jncics/pkz063/5556295.

[15] Derek Dustin, Guowei Gu, and Suzanne A. W. Fuqua. "ESR1 mutations in breast cancer". en. In: *Cancer* 125.21 (Nov. 2019), pp. 3714–3728. ISSN: 0008-543X, 1097-0142. DOI: 10.1002/cncr.32345. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cncr.32345.

[16] Shahana Ahmed et al. "Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2". In: *Nature Genetics* 41.5 (May 2009), pp. 585–590. ISSN: 1546-1718. DOI: 10.1038/ng.354. URL: https://doi.org/10.1038/ng.354.

[17] Kristen N. Stevens, Celine M. Vachon, and Fergus J. Couch. "Genetic Susceptibility to Triple-Negative Breast Cancer". en. In: *Cancer Research* 73.7 (Apr. 2013), pp. 2025–2030. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-12-1699. URL: http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-12-1699.