**2019/2020**

# Internship Report

presented by :

**Bassem JABER**

Work conducted starting 17/02/2020 to 14/08/2020 in :

**EMBL-EBI**
**and**
**NOVO NORDISK RESEARCH CENTER**
**OXFORD**

Mission statement :

## Downstream analysis of GWAS and eQTL colocalisation results

Internship Director : **Daniel ZERBINO**

Internship Counselor : **Wojciech PIECZYNSKI**

This work was done for the EMBL-EBI European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD and Novo Nordisk Research Center Oxford, The innovation Building, Roosevelt Dr, Headington, Oxford OX3 7FZ

# Contents

# 1 Abstract

In this internship report, I'll do my best to cover the 6 months of this very special internship in the Zerbino Research Team of the EMBL-EBI.

The Pharmaceutical industry is one of Europe's most dynamic sectors, it shows a steady growth over the last decade. A wealth of companies work in this field but some key actors have a bigger part to play, and I worked in collaboration with one of these actors, Novo Nordisk. These companies develop new medicine through a particularly lengthy, expensive and risky process, called the drug development pipeline.

To understand the process to create a drug, one needs to understand the basics of genetics, molecular biology and genomics, especially to exploit the data produced the algorithms used in statistics study.

I worked in the EMBL-EBI and it plays a key role in the scientific community. It understood the importance it has, in terms of Civil Responsibilty, sustainable development and the security of its employees among other subjects.

In my mission, I worked using the results of an existing algorithm created by my team. This algorithm uses the eCAVIAR method and the Expectation-Maximization algorithmto generate an output with meaningful information. To analyse this information, I developed a code to generate figures using the output created in the form of heatmaps and Manhattan plots, making it easier to find the relevant information we're looking for.

I learned quite a lot in this internship about the range of my coding abilities, my skills in solving crucial problems in stressful times while staying in good terms with the persons involved, it was quite an enjoyable experience and I hope I will continue my journey in the biostatistics field back

in France for my future career.

All my work can be found in my GitHub page at the following URL : https://github.com/BassemJ

## 2 Acknowledgements

I'd like to thank Dr Daniel Zerbino for trusting me in doing this internship with him. He helped me in this internship in multiple ways, cleaning my code, working in a research environment, working in an industrial environment.

Helping me learn by myself, at my own pace, in an encouraging environment and a helping team.

Last, but not least, for helping me in a dire situation during the pandemic when I had to move out in the middle of the crisis.

I'd like to thank Pr Wojciech Piezynski for advising me in my internship.

I'd like to thank Rachel Edgar for helping me write this report. Her comments helped me have an external view and be more concise and more precise on the biology part.

I'd like to thank Yalan Bi for her help with the algorithm and understanding the Pipeline.

And last but not least, I'd like to thank my family and my friends for supporting me in this peculiar internship with lots of ups and downs.

# 3 Economics of the Pharmaceutical Industry

The pharmaceutical industry is the sixth biggest market in the world and in undergoing massive growth [1]. The revenue of the world wide pharmaceutical market has increased from 390.2 billion dollars in 2001 to 1.2 trillion dollars in 2018 [2].

The main objective of the pharmaceutical industry is to alleviate the symptoms, find vaccines or design cures for disease [3]. Advances in the pharmaceutical industry have contributed to increased life expectancy and improved quality of life according to measures like the Human Development Index (HDI) [4].

One example is the reduction in number of deaths worldwide due to HIV, which has dropped from 1,750,000 estimated deaths in 2004 to 770,000 estimated deaths in 2018 [5]. This is the results of a combined effort, people were tested more for HIV and were less likely to contaminate other people. Prevention and raising awareness about HIV made it easier to preserve the health of the people. The medication for HIV, in particular tritherapy, helped raise the life expectancy of seropositive people, and this is a direct effect of the efforts of the pharmaceutical industry [6].

The pharmaceutical industry is a major component of the European economy. More than 396 billion euros worth of medications were exported from Europe in 2017 with a trade surplus of more than 101 billion euros. European pharmaceutical companies employed 760,795 people in 2017 with 114,655 dedicated to Research and Development [7].

Pharmaceutical companies spent more than 35 billion euros in Research and Development in a year. The focus on Research and Development in pharmaceuticals is exceptional. The industry's reinvestment of 15.0% of the net sales is almost twice the size of the next closest industry's rein-

vestment in Research and Development (Technology Hardware & Equipment with 8.7%) [7].
France is the fourth biggest European investor in Research and Development in the Pharmaceutical Industry with 4.451 billion euros invested in 2017 [7]. France is also the fourth biggest drug producer in 2017 with a production of 21.9 billion euros in 2017 [7]. The pharmaceutical industry in France is the second biggest European employer in 2017 with 98,786 people employed [7].
Not only do French people work in pharmaceuticals but they are the leaders in terms of consumption of drugs with a mean of $300 per capita. France is the second biggest European market with a value of 28.419 billion euros in 2017 [7].
The largest pharmaceutical company in France is Sanofi and it's also the the eighth biggest pharmaceutical with 36.126 billion euros of net sales for 2019 [8]. It was the third largest pharmaceutical market with 5.5% of the global market in 2008 [9].

My internship was performed as part of a collaboration between the EMBL-EBI and Novo Nordisk. Novo Nordisk is a pharmaceutical company based in Denmark, ranked 15th worldwide and 7th in Europe with a net revenue of 16.377 billion euros in 2019 and 2.095 billion euros invested in Research and Development in 2019 [9]. The company employs 43,258 people world wide and is specialised in obesity and diabetic care [10].

With such a massive share of French and European economies it is important to consider where the economic resources in pharmaceuticals are allocated. On average, 15.6% of pharmaceutical companies' budgets allocated to the Research and Development for the pre-human and pre-clinical drug development phases.
After drug development, marketing of a new drug was estimated at over 2,558 million dollars in 2013 [7]. Therefore, effort has been put into reducing drug development costs and the time to

market (i.e. time from drug discovery and patent application to commercialization) while ensuring the same quality and safety to patients.

## 3.1 The creation of a new medicine

The process to create a new drug is tightly regulated to ensure patient safety. The complete process is detailed below, as originally published in the European Foundation of Pharmaceutical Industries and Associations' 2019 Report [7].

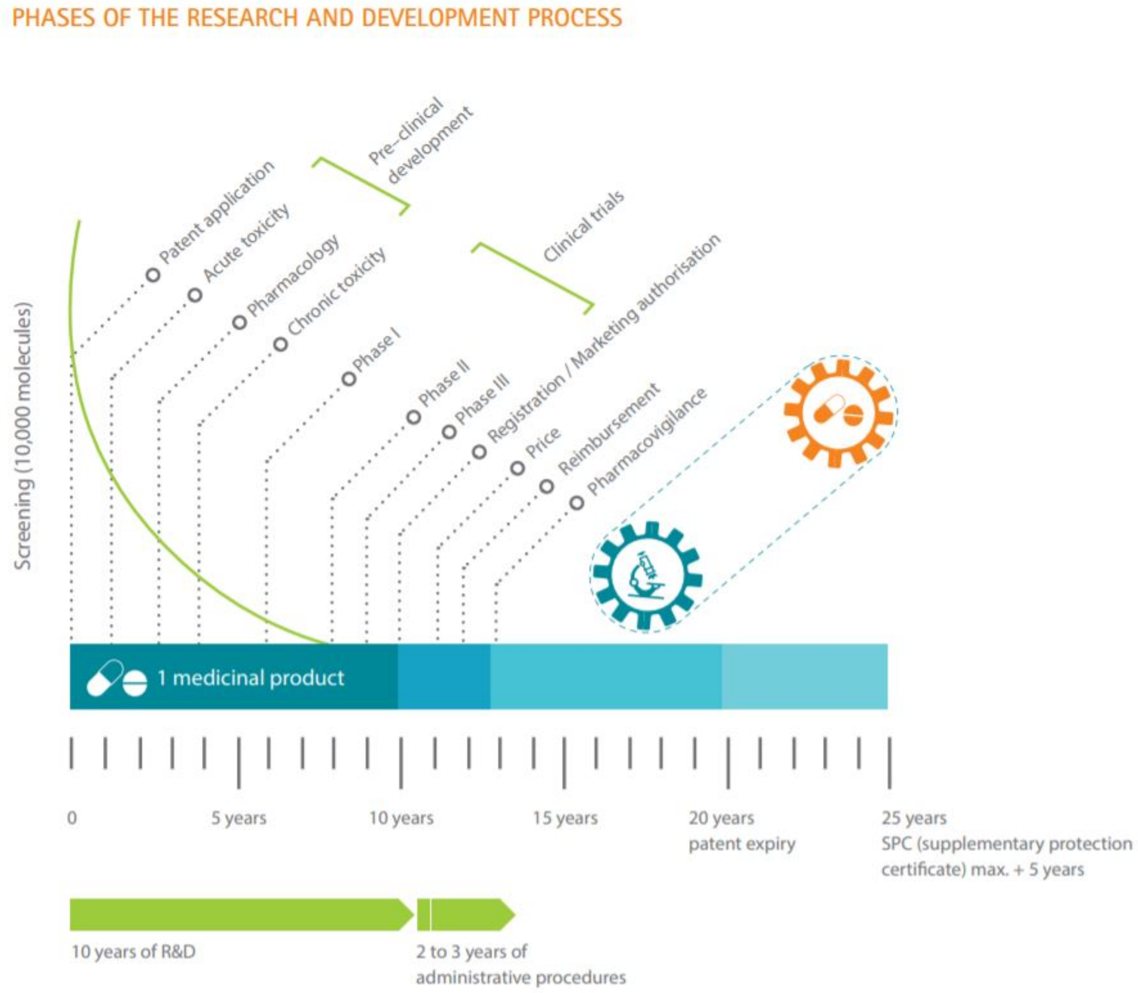


Figure 1: Phases of the process of the development of a new medicine.

The process to create a new medication is systematic and it has several phases that are well established. When a company submits a patent application a countdown begins. Companies have twenty years from patent application to develop a drug that is safe for patients and begin selling in order be profitable (Figure 1).

As it is difficult to reduce the time needed for administrative procedures, an area of focus is the procedure prior to patent application. The goal is to change the entire approach used. The screening of thousands of molecules to find of the right ones for eventual medicinal products is the current method.

This approach has several limitations. Primary among them, is the inclusion of many molecules which ultimately turn out to be useless, upon additional testing. Improving identification and prioritization of molecules having an effect on a condition could prevent a loss of time and money for the company. This would be a huge step for the pharmaceutical industry's Research and Development sector, and get effective treatment to patients in need more quickly [11].

Towards the improvement of drug development, one approach taken by Research and Development teams is the use of genomics. Often teams will analyse genomes of a control population and a population affected by the illness studied to find meaningful differences. For my internship, I was working with a team focused on a similar approach but with use of a new method.

A set of diseases which could greatly benefit from the delivery of drugs to patients is cardiovascular diseases. In Europe, 35.8% of deaths are caused by major diseases of the circulatory system [7], leading to intensive research toward development of new preventative medications.

Toward this goal, expertise will be needed on genomics and the drug development pipeline. Therefore, my internship was a collaboration between the EMBL-EBI and Novo Nordisk Research Center in Oxford. Specifically, my internship's mission was to use the UK BioBank Database [12] to try and find markers in the human genome that could be correlated to the cardiovascular disease.

**Biological Background and Key Definitions** After this broad introduction to the pharmaceutical industry, let us now focus on the different biological areas of interest of my internship.

From now on we'll call a patient's condition their **phenotype** and the genome associated the **genotype**.

Let's define two key concepts before continuing further :

**Gene :** A gene was initially defined as a trait that followed Mendelian inheritance rules. It is true, but nowadays scientists of the field of genetics prefer to refer to genes as a DNA sequence which is heritable between generations. Genes are often related to the protein they create. These proteins are the molecules which carry out a specific functions cells [18].

**Nucleotide :** A nucleotide is a small molecular complex made of a sugar molecule named **Deoxyribose**, two phosphate groups and a nitrogenous base. Nucleotide is the basic element that constitute DNA. The difference between the nucleotides that make the DNA lies in the 4 different nitrogenous base that exist: **Adenine (A), Thymine (T), Cytosine (C), Guanine (G).**

As we can see, our gene's definition takes into account Mendel's laws of inheritance, but what are they exactly ? Let's see these laws, stated thanks to the two experiments Mendel did with the pea plants he had.

**1. Law of Dominance :** This law states that a hybrid offpsring ( meaning, two different allele of one gene) will only inherit the dominant trait in the phenotype. The suppressed allele is the recessive trait and the expressed allele is the dominant trait.

**2. Law of Segregation :** This law states that two copies of each hereditary factor are produced during the production of gametes, and are segregated so that the offspring acquires one factor from each parent. In other words, alleles are segregated during the formation of gametes and randomly paired during fertilization.

**3. Law of Independent Assortment :** This law states that the segregation of allele pairs during gamete pairs is done independently, meaning uniformly random, so that all different possible allele pair get an equal chance to occur.

Our focus from a patient's genotype will be **Single Nucleotide Polymorphisms** (SNP). Let's define what a SNP is :

**Single Nucleotide Polymorphism :** A Single Nucleotide Polymorphism (or SNP) is defined by the substitution of a single nucleotide by another one at a specific loci in a genome. This is the most common difference between the genomes of the individuals. When more than 1% of the population has this polymorphism, it is considered as an allele of a SNP [25].

The threshold used for a SNP is usually of 1%, meaning that if a SNP is present in more than one in a hundred people in the population, then this difference will be sufficiently distributed in the population, but some studies include SNPs that are even less distributed among a population. The different versions of a gene, that differ with one or more SNPs, are called the **alleles** of the gene studied [16].

To continue further, we need to introduce two concepts :

**Chromosome :** A chromosome is a microscopic element made of DNA and proteins with parts of the genetic material of an organism.

**Genome :** The genome is all the genetic material of an organism. The size of a genome is measured in base pairs. Human genome has approximately 3.2 billion base pairs and is divided in 23 chromosome pairs.

## 3.2 Essentials of Genetics

The study of genetics is the study of genes, genetic variations and the inheritance of traits in a population through generations. It was first studied by the Austrian monk Gregor Mendel in 1865 by studying the inheritance of observable traits in pea plants.

The observations he made encouraged him to state the laws of inheritance in living beings and his name was given to **Mendelian Genetics** [17]. Although several special cases are now known not to follow these particular rules, an important number of species follow the rules he gave after his study.

Genes seem to be at the heart of genetics, but what are they exactly?

### 3.2.1 Genes

A quick reminder about genes as we already defined it earlier in this report. A gene is a DNA sequence which is heritable between generations. Genes are often related to the protein they create. These proteins are the molecules which carry out a specific functions cells [18].

As a result, genes can be identified by several DNA sequences as amino acids can be coded by different RNA sequences. To illustrate this, let's take for example the Guanine-Cytosine-Thymine triplet. It will generate an Alanine amino acid, but the Guanine-Cytosine-Cytosine triplet also generates an Alanine amino acid, that's why a gene, which is known to generate a specific protein, which is a specific chain of amino acid, can be identified by several DNA sequences.

Genes can be identified by their DNA sequence but also by their loci, which is the physical position of the gene in the chromosome. A DNA sequence generates amino acids with each nucleotide triplet.

The study of genes can be done thanks to the study of the traits of populations in the 1800s by Darwin and Mendel. It is remarkable that their study of inheritance was done without knowledge of the DNA that constitutes it.

### 3.2.2 Complex traits and diseases

**Complex trait :** A Complex trait is a phenotypic observation which can't be explained by Mendelian genetics laws'. These traits result from the expression of multiple genes at the same time all at different levels in combination toward a phenotype [19].

A common research subject is to try to find the most correlated genes or alleles to a specific trait, for example the yield of a wheat crop or the size of a tomato. Various studies have had interests in complex traits in agriculture to enhance the quality of the plants.

But this approach can also be taken to try to better understand the genes which impact someone's phenotype or presence of a disease.

### 3.2.3 Linkage Disequilibrium

When researchers did crosses within a population they noticed exceptions and they tried to find an explanation. It seemed that the traits they studied weren't independent.

Some traits appeared together a lot more than what they were supposed to and other traits appeared together less often than expected. They hypothesized that the genes were linked together. Genes that controlled these non-independent traits were at loci which were very close together. So in the crosses these traits were inherited together on one block of DNA and the traits appeared more often together in subsequent generations.

This is called **Linkage Disequilibrium**. These terms define the fact that alleles in linkage disequilibrium appear together more or less often than they're supposed to. This is the result of a non-random (meaning not uniformly random) association.

When two locis are close to each other, the distance used being the centiMorgan ( 1 centiMorgan = 1 shuffling between the two genes every 100 crosses) [20]. This unit has been historically used, but now the unit geneticians use is the nucleotide.

Linkage Disequilibrium can be measured. Let $p_{AB}$ be the frequency of the occurrence of A and B in the same gamete and $p_A$,$p_B$ be the frequencies for allele A alone and B alone respectively. then if $D_{AB}$ is the coefficent of linkage disequilibrium,

$$D_{AB} = p_{AB} - p_A p_B \tag{1}$$

However, this measure is sometime not convenient because it doesn't give an insight on all the possible values, it doesn't show if it's an extreme value or close to the expected frequencies [21].

A more common derived measure of the linkage disequilibrium is the ratio between the value measured over the most extreme value possible [21]. It is defined by :

$$D' = \frac{D}{D_{max}} \tag{2}$$

Where :

$$D_{max} = \begin{cases} \max[-p_a p_b, -(1-p_a)(1-p_b)] \text{ if D} < 0 \\ \min[p_a(1-p_b), (1-p_a)p_b] \text{ if D} > 0 \end{cases}$$

## 3.3 Essentials of Molecular Biology

Our object of interest is the human genome, but what is it made of ?

The human genome, and in fact, all living organisms' genomes, are made of **Deoxyribonucleic Acid**, also known as **DNA**. DNA is a complex molecule made of nucleotides. A nucleotide is a small molecular complex made of a sugar molecule named **Deoxyribose**, two phosphate groups and a nitrogenous base. The difference between the nucleotides that make the DNA lies in the 4 different nitrogenous base that exist: **Adenine (A), Thymine (T), Cytosine (C), Guanine (G).**

DNA can be seen as the code that runs a living organism, the code of life. It is present in every cell of all living beings. It also serves as the recipient of heredity. When reproduction happens, a single strand of DNA from the father and one from the mother in special cells named **gametes** will fuse to create the first cell of the offspring generated that will carry this shared information and transmit it in the future to his descendant. This reproduction mode using gametes is specific to the species using sexual reproduction.

**DNA Replication** : A cell in a living organism has the DNA of the whole organism. Each cell's DNA is a near exact copy of the DNA of all the other cells of this organism. Initially, it is a single cell and after some time it becomes a complex organism with a variable number of cells, depending on the living being (for example, a human being has approximately one hundred trillion cells in total [22]).

Through cell division a single cell becomes a trillion. Cell division is when a cell divides in two and it creates two new daughter cells. These two new cells are exact copies, they have the same DNA and the same cell organisation. **DNA replication** occurs during cell division and allows a cell to create two identical sets of DNA.

DNA replication happens in three steps: opening of the double helix of the DNA and separation of the two strands, creation of the complementary DNA strand for each single strands, closure and assembly of the two new DNA strands.

Here is a simple and comprehensive illustration of how DNA replication works [23].

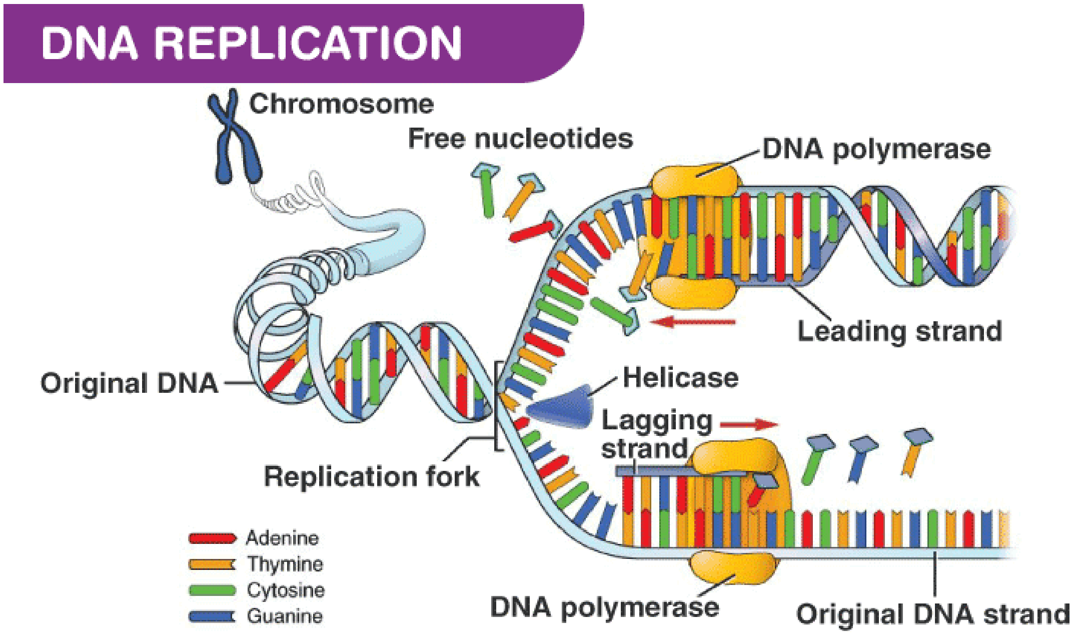


Figure 2: DNA Replication process.

DNA is the support of all the information a cell needs to function and maintain itself. It is organised in genes which are, to put it simply, a sequence of nucleotides that can be read by a complex molecular machinery to create specific proteins coded by the gene.

However, not all genes lead to the synthesis of a protein, some RNA are biologically crucial like ribosomes without being translated into being proteins.

The origin of diversity and variation in a population comes from the **mutations** (duplications, deletions, additions) that occur and errors in the process of replication of the DNA. A single nucleotide differing from a sequence to another can have no apparent effect to completely change the protein created and be the cause of a devastating illness like Progeria [30].

The information contained in the DNA has to be transmitted to the other parts of the cell in order to be used to regulate the cell or create specific proteins. This is the **transcription** phase. An enzyme named **Ribonucleic Acid Polymerase** will create a Ribonucleic Acid (**RNA**) molecule by copying the attaching to a portion of DNA and transcribing those nucleic acids to RNA.

There are several types of RNA, but the one we're interested in is **Messenger Ribonucleic Acid** (**mRNA**).

The goal of mRNA is to carry information from DNA to **ribosomes**. Ribosomes then **translate** the information contained in mRNA into proteins.

A specific cell division process takes place for eukaryotic organisms that reproduce sexually.

**Meiosis** is a cell division that creates two **gametes** each with unique DNA. The parent cell's DNA is shuffled before gametes are created. This ensures all offsprings are unique because they'll result from the fusion of two gametes each with shuffled DNA, one for the mother, one from the father.

The variations in the human genome from one person to another can come from the initial shuffling of the DNA during meiosis and also from the mutations as stated above that take place on a more or less frequent basis. An estimated rate of mutations in humans is $5 * 10^{-11}$ per division, which means that one mutation will appear every $5 * 10^{11}$ division in general [24]. Mutations can happen due to a number of factors, including external agents' actions, we call these **mutagens**.

Some very key concepts for this work lie in the fact that all cells of a single organism contain the same genetic information (except for the few rare mutations). Despite their identical DNA, a brain cell and a skin cell have two different purposes and functions. This is due in part to **gene regulation**. Complex machinery directs RNA polymerase to the sections of DNA in each cell, which need to be transcribed for the cell to function properly and the transcription then needs to be stopped at the right time.

Some of machinery we describe is another portion of DNA. Some parts of the genetic information encoded in DNA does not code for protein but regulates transcription and translation. It's the **non-coding regions** of the DNA. This is why we have an almost infinite number of possible percentages of expression of a gene, some of these regions can induce, hinder or completely block the expression of a gene.

## 3.4 Essentials of Genomics

The **genome** is composed of an organism's coding and non coding DNA. The genome is divided into chromosomes containing all genetic material [16]. **Genomics** is the study of the genetic material in organisms at the scale of a population and generally the whole genome of each individual, to put it simply, genomics studies one's DNA sequences while genetics focuses on the phenotypes in a population.

### 3.4.1 Single Nucleotide Polymorphism

Let's quickly remind ourselves about the definition of a SNP. A SNP is the substitution of a single nucleotide by another one at a specific loci in a genome. This is the most common difference between the genomes of the individuals. When more than 1% of the population has this polymorphism, it is considered as an allele of a SNP [25].

SNPs occurring in coding regions of the genome can have easily noticeable effects or more subtle ones when occurring in a genes regulation part of the genome for example. These non-coding SNPs could result in a minor difference between individuals. When many SNPs are studied for association with a particular phenotype, this forms the foundation of GWAS.

### 3.4.2 Genome-Wide Association Study

**Genome-Wide Association Study :** A Genome-Wide Association Study is a study of a set of genetic variants (or SNPs) over the whole genome in a designated population to see if any variant is associated with a trait [26].

This study's objective is to analyse the genome of a control population and a case population to examine whether we can find particular SNPs which occur more often in the case population than in the control population.

If this is the case, these SNPs could potentially be causal for a designated disease of interest for the study. Ultimately one could try to understand how to prevent the consequences of these causal variants, to try to alleviate the negative effects of the disease targeted.

### 3.4.3 Functional Fine-Mapping

In the context of Genome-Wide Association Studies, one will try to map SNPs that are most likely to have effects on the condition one is researching about. **Finemapping** consists in finding the causal SNPs. **Functional Finemapping** covers the finemapping methods that take into account information about the functional regions of the genome.

The goal of this procedure is to associate SNPs with the function of the SNP in a map of the chromosome of the genome studied. It can be quite difficult to know if the Functional Finemap is relevant to the study at hand because SNPs can have no effect or one that we don't know of, which can quickly make things harder in order to get relevant information out of it.

### 3.4.4 Expression Quantitative Trait Loci

**Expression Quantitative Trait Loci :** An Expression Quantitative Trait Loci or more commonly called eQTL is a particular SNP which has a significant association to the expression of a gene.

When it comes to the expression of a gene, there is a wealth of possibilities and not only the binary "Present" or "Absent". This binary conception of the expression of a gene was the one used in Mendelian genetics, there was no in-betweens at that time.

Now we know that the expression of a gene can be translated into a real positive number to have a better understanding of the effect the gene will have on an individual. Expression Quantitative Trait Loci have been the most useful piece of information in a lot of research as it can be linked to a phenotype or correlated to a situation or a condition, which is why it is widely used in the research on genetics.

# 4 Presentation of the EMBL-EBI

## 4.1 The EMBL-EBI : European Molecular Biology Laboratory - European Bioinformatics Institute.

### What is the EMBL-EBI ?

The **EMBL-EBI** or **European Molecular Biology Laboratory - European Bioinformatics Institute** is home for biological data services, research and training. It is also a trusted provider of biological data for research all over the world. The European Bioinformatics Institute is part of the European Molecular Biology Laboratory, which is an intergovernmental research organisation with hubs all over Europe ( France, Spain, Italy, UK and Germany).

The structure was the idea of Leó Szilárd, James Watson and John Kendrew. The purpose of this structure was to create an international research centre, like CERN, to have a high quality research in Europe and rival American research which was quite ahead at that time.
John Kendrew was the first EMBL Director general until 1982, and since January 2019, Edith Heard is the fifth director of the EMBL and the first woman to have this role. She was the director of the Institut Curie in Paris just before becoming director of the EMBL.

This structure is quite unique because it is under the supervision of the Foreign Office and not the Ministry of Education and Research as it could be supposed. The structure also has a special status because it is an extraterritorial organization, any potential employee can ask for a special visa no matter his origin, which helps create a meltingpot of cultures in each team.

The goal of the EMBL-EBI is to deliver excellent research in various subjects, provide scientific services like delivering relevant datasets for biology research, train the next generation of scientists for a limited time before letting them go back to the public or private research, engage with the industry in projects and coordinate the bioinformatics in Europe.

The EMBL-EBI has a very diverse environment and promotes multiculturality and diversity. The site houses more than 80 different nationalities and it is a way to provide various point of views on subjects to help think out of the box and think differently.
The EMBL-EBI is primarily funded by EMBL's member states of the European Union. Other major funders are the European Commission, the UK Research and Innovation, the National Institutes of Health, the Wellcome Trust and the Industry Programme.

The research done in the EMBL-EBI is linked to various fields. From research to find a smarter farming for better food security to finding regions involved in diseases in the human genome and mapping biodiversity.

The EMBL-EBI plays a major role in providing Big Data services to researchers all over the world. More than 64 million requests to EMBL-EBI websites are registered everyday from approximately 20 million unique users to access more than 273 petabytes of data. An estimate of 22 500 participants complete EMBL-EBI Training events, online or on site.

The EMBL-EBI is considered essential to the research of 70% of its users and is ranked in the top 1% for the processes of generation and scholarly communication of scientific knowledge. It also enables to reduce the costs of exploiting relevant data to specific research for 79% of users, 93% of users think that it saved them time to find the data, 79% of users think that it enabled them to explore more novel research questions and 81% of users think that it allowed them to reduce the duplication of their efforts in order to complete their research.

The EMBL-EBI also has a strong relationship with the pharmaceutical industry. Businesses can use the servers directly housed on site with the date directly available so that small businesses can also run large-scale analysis on relevant and unchanged datasets.

The EMBL-EBI also has a strong sense of providing services and data for free. The structure is a champion of open data in the life science and provide freely available data services. All teams provide their codes freely on GitHub and it is maintained and frequently updated.

The fact that everything is Open Source also helps stimulate exchange and advances when someone uses it and gives a helpful feedback about how to enhance the user experience for example. Small companies use the services provided by the EMBL-EBI but also large international pharmaceutical companies.

To only cite a few : AstraZeneca, Sanofi and Pfizer use the pharmaceutical and diagnosic services, Unilever and Bayer CropScience use the agri-food and consumer goods services provided. They are all member of the industry program of the EMBL-EBI, making them close collaborators in the scientific processes.

## 4.2 Civil Responsibility, Environment protection and Security in the EMBL-EBI

### 4.2.1 Civil Responsibility

In the EMBL-EBI, there's a clear intention of respecting employee's privacy, personal and professional life balance and protecting the environment.
In terms of civil responsibility, the EMBL-EBI knows it has an important role to play in order to show the young generation the possibilities in the genetics field and science in general.

That's why the EMBL-EBI encourages its employees into doing trainings to be able to properly present in an interesting manner their work to students of all levels when it is possible to go and see a class.
The EMBL-EBI also has a partnership with biology universities and employees can go and show what they do while teaching to a class as a part of their cursus.

This is an excellent way to promote science, biology and research in general to the next generation because a lot of false assumptions can be made by students that may make them take a different path when they could be a perfect fit for the job.

### 4.2.2 Environment protection

The EMBL-EBI promotes environment protection. A free shuttle bus service is available for all employees coming from the city to work. This was offered to all employees in order to limit CO2 emissions for everyday travel to work when you're usually alone in your car.

In all break rooms, there is a variety of different bins to encourage recycling. Each bin is colored and explains what should be thrown in that particular bin, which easily tells you where to throw each things when you don't have this variety at home.

There's also a huge park in the campus where you can have a walk whenever you want. There are all kinds of plants and flowers. It is there to try to protect biodiversity in the campus and every employee is welcome to come and do some gardening if he wants, it is encouraged and a gardener is always there to do trainings.

They also try to limit overprinting. All scientific journals are freely available digitally, you can print it if you want but it is advised to do it only if you know you need to have a paper version. The library also has a limited number of printed books in the campus, if a paper version of a book is needed, it is possible to order it, but a digital resource is available where you can lend the PDF of a book for a limited period of time.

### 4.2.3 Security

The campus is quite a safe place. Everyone, including visitors, has a badge that allows someone to identify the people in the campus if needed.

There is a security checkpoint to enter the campus with a barrier, a reception in each building to help you navigate between the buildings.

The company gives you a company laptop to work on. The work and the code is open source so it's not so much to limit the potential leaks of anything like that, it's more in order to protect

the servers and your saves. If a computer is infected and connected to the network of the campus, it could be a real problem.
All newcomers are invited to an introduction on good behavior online when at work, be careful about your emails, separate private and work, change passwords, etc...

The campus also has a local nursery for employee's children that didn't find a place in public ones, this is a rare service that is worth to be noted, it allows researchers to work at peace without worrying about their children because as you certainly know, places are few and far between and it can stop a parent from working.
I also have a personal anecdote when I felt like I was in need of help and I received more than I could imagine.

After the coronavirus outbreak, my landlord, in the middle of the crisis, without giving me notice, told me I had to find a new place to live because she was worried I would infect her and she didn't want to take this risk. I was in shock and afraid for my security.
I talked about it with my team leader and he immediately tried to find if there was a room available for me using mailing lists, he contacted Human Resources and they helped me try to understand the situation and see if there was a way to do things smoothly. They helped me financially and they reassured me when I was afraid for my security, which is for me an excellent indicator showing that they care about their employees.

All in all, the EMBL-EBI understood well the role it plays in the society and uses its benefits in a smart way to try to be a good place to work with all the services it offers to protect the environment, assure the security of its employees and do its best in order to promote sustainable development.

# 5 My mission

Let's now dive into the heart of my internship, my mission in the Zerbino Research team.

## 5.1 The objectives of my internship

The Zerbino Research Team of the EMBL-EBI works on the Genome-Wide Association studies to find causal regions of the human genome in diseases of interest to try to find a medicine directly looking into the genomes of the people we study.
Here is my mission : "Study of GWAS and eQTL colocalisation data using statistical analysis with Python for genome analysis. Graphics generation to understand and identify potential genes responsible for health issues to create new drugs."

I had one question I tried to answer which was : " Can we find potential genes of interest and subsequently drugs of interest to alleviate the symptoms and potentially cure cardiovascular diseases by studying Genome-Wide Association Studies and Expression Quantitative Trait-Loci colocalisation results ? "
In order to achieve my objective and answer my question, I had to start by getting more accustomed to the biological studies and the vocabulary used because my background is heavily focused on statistics, even if I had the opportunity to have a class on Biostatistics.

Therefore, first of all, the goal was to have me understand the research going on and get in pace

so the first month was dedicated to getting me on track by having me read about the field of biology and learn enough to be able to understand actual research by myself, without having someone else explain some key elements I didn't get yet.
This is why I have quite a large bibliography for this internship. Most of it is the result of my first period when I was carefully reading and learning about biology.

After this first period of time, I presented the results of my research to my group. The goal wasn't for me to teach them something but more to make sure that I didn't get something wrong that would handicap me in the future when I would have to work all by myself on this kind of subject.
Everything went well and everyone was pleasantly surprised by how much progress I made by learning quite complex subjects in quite a short span of time.

After this first period of time, my goal was divided in two :

- My first objective was to find a way to generate figures from the data we had to test our algorithms to have a working code able to generate everything we need to analyse the data we have at hand with a single line of command.

- My second objective is to analyse the actual code in the GitHub libraries. It is used in the process designed to generate the finemap used to create the results of the Genome-Wide Association Studies. My objective was to try to find a way to enhance the process with my statistical background in terms of finding the right candidate genes and SNPs.

## 5.2 Colocalisation posterior

In order to compute the Colocalization Posterior Probability for a GWAS and eQTL study, we use a method similar to the eCAVIAR method presented in this paper [27].

Before diving right into it, let's first start by understanding the CAVIAR Model used for finemapping that is at the root of the eCAVIAR method used to compute the probabilities.

We have $N$ individuals' genotypes at $M$ specific SNPs or to simply put it, variants. Our data contains values for multiple genes and a phenotype, we assume that there is at least one statistically significant variant for both the phenotype and the gene expression. We assume that there is the same number of people in both the GWAS and eQTL studies and that the pairwise Pearson correlation is the same too.

Let $Y^{(p)}$ be an $N * 1$ vector of the phenotypic values where $y_j^{(p)}$ is the phenotypic value of the $j^{th}$ individual (for example $y_j^{(p)} = 0$ means the individual is a control, he's not affected by the disease studied, and $y_j^{(p)} = 1$ means that it's a case, the individual is affected by the disease).

Let $Y^{(e)}$ be an $N * 1$ vector of gene expression for one gene of interest with one statistically significant variant associated with the expression of that gene. Let $G$ be an $N * M$ matrix of genotype information where $G_i$ is an $N * 1$ vector of minor allele counts for the $N$ individuals at the $i^{th}$ variant. Knowing this,there are 3 possible values for $g_{ji}$, $g_{ji} = 0$ or $g_{ji} = 1$ or $g_{ji} = 2$.

This is the case because humans have a diploid genome, meaning they have 2 homologous sets of chromosomes, so each individual can have no minor allele, a single minor allele in either chromosome or minor alleles in both chromosomes.

We standardize the values of phenotypes and genotypes for a null mean and a unit variance, with $X$ being the standardized matrix of $G$. We assume an "additive" Fisher's polygenic model, which is commonly used for GWAS and in statistics in general to have an estimate of a value. In this model, the phenotypes follow a normal distribution and our assumption that it is an additive model means that each variant contributes to the phenotype linearly.

This leads us to this linear model :

$$Y^{(p)} = \mu^{(p)} 1 + \sum_{i=1}^{M} \beta_i^{(p)} X_i + e^{(p)} \tag{3}$$

$$Y^{(e)} = \mu^{(e)} 1 + \sum_{i=1}^{M} \beta_i^{(e)} X_i + e^{(e)} \tag{4}$$

With $\mu^{(p)}$ the mean of phenotypic values and $\mu^{(e)}$ the mean of gene expression values. $\beta_i^{(p)}$ and $\beta_i^{(e)}$ are the effect size of the $i^{th}$ to the phenotypic values and gene expression values respectively. We add $e^{(p)}$ and $e^{(e)}$ to take into account the environment and measurement errors for phenotypic values and gene expression values respectively. We define the errors as white noises with $\sigma_e^{(p)2}$ and $\sigma_e^{(e)2}$ variances for phenotypic values and gene expression values respectively.

Let $S^{(p)} = [s_1^{(p)}, s_2^{(p)}, ..., s_M^{(p)}]$ be the marginal statistics for the phenotype of interest and $S^{(e)} = [s_1^{(e)}, s_2^{(e)}, ..., s_M^{(e)}]$ be the marginal statistics for the gene expression. The joint distribution of the marginal statistics follows a Multivariate normal distribution, given that we have the true effect sizes.

This gives us :

$$(S^{(p)}|\Lambda^{(p)}) \sim N(\Sigma\Lambda^{(p)}, \Sigma) \tag{5}$$

$$(S^{(e)}|\Lambda^{(e)}) \sim N(\Sigma\Lambda^{(e)}, \Sigma) \tag{6}$$

With $\Sigma$ being the pairwise Pearson's correlation matrix of genotypes and $\Lambda^{(p)} = [\lambda_1^{(p)}, \lambda_2^{(p)}, ..., \lambda_M^{(p)}]$, $\Lambda^{(e)} = [\lambda_1^{(e)}, \lambda_2^{(e)}, ..., \lambda_M^{(e)}]$ being the true standardized effect size for all the variants of the desired phenotype and gene expression respectively. The true effect size is set to zero for non causal variants and not equal to zero for causal variants. $\Sigma\Lambda^{(p)}$ and $\Sigma\Lambda^{(e)}$ are the LD-induced non-centrality parameters for phenotype and genotype expression respectively. We need to have this information in order to have the correct hypothesis to assume the Multivariate normal distribution for our joint distribution.

We define a new variable $C^{(p)}$. This new variable is a binary vector giving the causal status, indicating which variants are causal and which variant aren't causal. If $c_i^{(p)} = 1$ then the $i^{th}$ variant is causal, if $c_i^{(p)} = 0$ then the $i^{th}$ variant isn't causal. In the CAVIAR Method, we define a prior on the vector of effect sizes using the Multivariate normal distribution assumption :

$$(\Lambda^{(p)}|C^{(p)}) \sim N(0, \sigma^{(p)2}\Sigma_c^{(p)}) \tag{7}$$

With $\Sigma_c^{(p)}$ being a diagonal matrix and $\sigma^{(p)2}$ being a constant equal to the variance of our prior using the Genome-Wide Associations Studies non-centrality parameters. The diagonal elements of

$\Sigma_c^{(p)}$ are set to 1 or 0 so that if a variant is selected in $C^{(p)}$, the corresponding element in $\Sigma_c^{(p)}$ is coherent with the values we discussed earlier ( 1 for causal, 0 for non causal).

This prior is used to compute the likelihood of each possible causal status. We find the following joint distribution given the causal status :

$$(S^{(p)}|C^{(p)}) \sim N(0, \Sigma + \sigma^{(p)2}\Sigma\Sigma_c^{(p)}\Sigma) \tag{8}$$

For the gene of interest in which we performed eQTL we have :

$$(\Lambda^{(e)}|C^{(e)}) \sim N(0, \sigma^{(e)2}\Sigma_c^{(e)}) \tag{9}$$

With the same conditions for the values of $\Sigma_c^{(e)}$.

Now that we defined our variables, let's dive further into the heart of our study, the Colocalisation posterior probability. To put it in simpler words, the Colocalisation posterior probability is the probability that the variant is causal in both the GWAS study and the eQTL study. We have the marginal statistic for the GWAS Study in the $S^{(p)}$ variable and the marginal statistic for the eQTL study in the $S^{(e)}$ variable.

To define the probability mathematically, we define the Colocalisation posterior probability by $P(c_i^{(p)} = 1, c_i^{(e)} = 1|S^{(p)}, S^{(e)})$ and we'll call $\phi_i$ the Colocalisation posterior probability of the $i^{th}$ variant. By using the law of total probability, we can compute the summation probability of all causal statuses where the $i^{th}$ variant is causal in both studies and other variants can be causal or non-causal. This gives us :

$$\phi_i = P(c_i^{(p)} = 1, c_i^{(e)} = 1 | S^{(p)}, S^{(e)}) \tag{10}$$

$$\phi_i = \sum_{C^{*}_{/i}{}^{(p)} \in \{0,1\}^{M-1}} \sum_{C^{*}_{/i}{}^{(e)} \in \{0,1\}^{M-1}} P(C_{/i}^{(p)} = C^{*}_{/i}{}^{(p)}, C_{/i}^{(e)} = C^{*}_{/i}{}^{(e)}, c_{/i}^{(p)} = 1, c_{/i}^{(e)} = 1 | S^{(p)}, S^{(e)}) \tag{11}$$

$$\phi_i = \sum_{C^{*(p)} \in \{0,1\}^{M}} \sum_{C^{*(e)} \in \{0,1\}^{M}} P(C^{(p)} = C^{*(p)}, C^{(e)} = C^{*(e)} | S^{(p)}, S^{(e)}) \mathbb{1}(c_i^{*(p)} = 1, c_i^{*(e)} = 1) \tag{12}$$

With $C^{*}_{/i}{}^{(p)}$ and $C^{*}_{/i}{}^{(e)}$ being the vectors of causal status for all variants except the $i^{th}$ variant for the phenotype of interest and gene expression. The $\mathbb{1}$ function is defined as :

$$\mathbb{1}(c_i^{*(p)} = 1, c_i^{*(e)} = 1) = \begin{cases} 1 \text{ if } c_i^{*(p)} \text{ and } c_i^{*(e)} \text{ are causal} \\ 0 \text{ otherwise} \end{cases}$$

We now utilize the Bayes rule to get :

$$\phi_i = \frac{\sum_{C^{*(p)}} \sum_{C^{*(e)}} P(S^{(p)}, S^{(e)} | C^{(p)} = C^{*(p)}, C^{(e)} = C^{*(e)}) P(C^{*(p)}, C^{*(e)}) \mathbb{1}(c_i^{*(p)} = 1, c_i^{*(e)} = 1)}{\sum_{C^{*(p)}} \sum_{C^{*(e)}} P(S^{(p)}, S^{(e)} | C^{(p)} = C^{*(p)}, C^{(e)} = C^{*(e)}) P(C^{*(p)}, C^{*(e)})} \tag{13}$$

With $P(C^{*(p)}, C^{*(e)})$ being the prior probability of the causal status of $C^{*(p)}$ and $C^{*(e)}$ for the GWAS and eQTL Study respectively. We assume that the prior probability over the causal status for the GWAS and eQTL study is independent so we have :

$$P(C^{*(p)}, C^{*(e)}) = P(C^{*(p)}) P(C^{*(e)}) \tag{14}$$

To compute these prior probabilities, we make the widespread assumption that the probability of causal status follows a binomial distribution with the probability of a variant being causal is equal to $\gamma$. Therefore, we have :

$$P(C^{*(p)}) = \prod_{i=1}^{M} \gamma^{c_i^{*(p)}} (1-\gamma)^{1-c_i^{*(p)}} \tag{15}$$

Genome-Wide Associations Studies and eQTL Studies are performed on independent sets of individuals and given the causal status in both studies, the marginal statistics for these two studies are also independent. This leads us to :

$$P(S^{(p)}, S^{(e)}|C^{*(p)}, C^{*(e)}) = P(S^{(p)}|C^{*(p)})P(S^{(e)}|C^{*(e)}) \tag{16}$$

This allows us to simplify the above definition of $\phi_i$ to :

$$\phi_i = \frac{\sum_{C^{*(p)}} P(S^{(p)}|C^{(p)} = C^{*(p)})P(C^{*(p)})\mathbb{1}(c_i^{*(p)} = 1)}{\sum_{C^{*(p)}} P(S^{(p)}|C^{(p)} = C^{*(p)})P(C^{*(p)})} * \frac{\sum_{C^{*(e)}} P(S^{(e)}|C^{(e)} = C^{*(e)})P(C^{*(e)})\mathbb{1}(c_i^{*(e)} = 1)}{\sum_{C^{*(e)}} P(S^{(e)}|C^{(e)} = C^{*(e)})P(C^{*(e)})} \tag{17}$$

According to the simplified version of our last definition for $\phi_i$, the probability that the same variant is causal in both the GWAS and eQTL Study is independent. It is equal to the multiplication of two probabilities, the probability of the variant being causal in the GWAS and the probability that the same variant is causal in the eQTL study. This means that the Colocalisation posterior probability is computed as :

$$P(c_i^{(p)} = 1, c_i^{(e)} = 1|S^{(p)}, S^{(e)}) = P(c_i^{(p)} = 1|S^{(p)}) * P(c_i^{(e)} = 1|S^{(e)}) \tag{18}$$

We compute the two parts of this probability using the definition of $\phi_i$, $P(c_i^{(p)} = 1|S^{(p)})$ is equal to the first part of the equation and $P(c_i^{(e)} = 1|S^{(e)})$ is equal to the second part of the equation and with this information we can compute the Colocalisation posterior probability for our sets of studies.

## 5.3 Expectation-Maximization and Maximum Likelihood Algorithms

Our goal with our work is to try to find causal variants linked to specific diseases.
We have access to some information, the genotype, the phenotype, the localisation of genes, the SNPs. But a lot of data is also unknown, it is a latent variable.
Knowing that, a fitting algorithm designed to do an estimate of a latent variable is finding the Maximum Likelihood estimator using the Expectation-Maximization Algorithm.

In our study, it is possible to have the observed data using the complete data available ( observed data and latent variable) by applying a certain non injective function to the complete data. We don't know the function, but we know it exists.

The EM Method giving birth to the EM algorithm is as follows :

The goal is to replace the log-likelihood $l_x(\theta) = ln(p(x;\theta))$ of the observed data by the log-likelihood of the complete data $ln(p(y;\theta))$ by its approximation $E[ln(p(y;\theta))|x]$, the expected value knowing the observed data $x$. By doing that, the log-likelihood becomes a function of the variable $x$ and the variable $\theta$. But $\theta$ is hidden, we don't know the value or else we wouldn't need to do all this procedure.

That's why we use this iterative method :

Initialization step : Initialization of $\theta$ by choosing a $\theta_0$ ( it can be random or advised by some information we can use)

Expectation step : Compute

$$Q(\theta,\theta_i) = E_{\theta_i}[ln(p(y;\theta)|x] = \int_{\gamma} ln(p(y;\theta)p(y;\theta_i|x)dy \tag{19}$$

Maximization step : Approximation of the maximum

$$\theta_{i+1} = arg\max_{\theta} Q(\theta,\theta_i) \tag{20}$$

This is the general form of the algorithm, but we work in a bayesian context, meaning that we can simplify the algorithm :

Because we have

$$\theta_{MAP} = arg\max_{\theta} p(\theta|x) \tag{21}$$

The Expectation step becomes :

$$E_{\theta_i}[ln(p(\theta|y)|x] = E_{\theta_i}[ln(p(y|\theta))|x] + ln(p(\theta)) - E_{\theta_i}[ln(p(y)|x] \tag{22}$$

$$E_{\theta_i}[ln(p(\theta|y)|x] = E_{\theta_i}[ln(p(y|\theta))|x] + ln(p(\theta)) + C, \tag{23}$$

With $C \in \Re$.

We find this relation by using Bayes' Theorem :

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{24}$$

And by noting that in the first relation, $\theta$ is a fixed parameter and that $E_{\theta_i}[ln(p(y)|x]$ is independent of $\theta$ because in bayesian context, $p(y)$ is the marginal law of y

Finally, knowing that, the Maximization step becomes :

$$\theta_{i+1} = arg\max_{\theta} \left(Q(\theta, \theta_i) + ln(p(\theta))\right) \tag{25}$$

## 5.4 Generating the figures to analyse the results.

Figures are needed in order to find relevant information in the output of our algorithm, but to start things off, one has to know about the data.

In order to analyse the data, we'll use four different python libraries :

**1. Pandas :** Pandas is a python library used to generate DataFrames. A DataFrame is a way of organizing data in a table. The power of Pandas resides in the wealth of methods applicable to the DataFrames that eases the analysis of data, and the multiple techniques it offers to easily modify your Dataframe as you wish. It is widely used in Datascience and for the reasons stated above.

**2. NumPy :** NumPy is a python library used to do mathematical operations on your data with ease. It takes into account all kinds of data structures so you don't have to change your original a lot to use NumPy's methods. It is optimised for operations in large dimension and is widely used in Datascience too.

**3. Matplotlib :** Matplotlib is a python library used to plot the data you have at your disposition and display it in various ways. It is widely used in academical research and computer science classes but new libraries based on Matplotlib's original way of displaying data emerged to have new and more graphical ways to display your data.

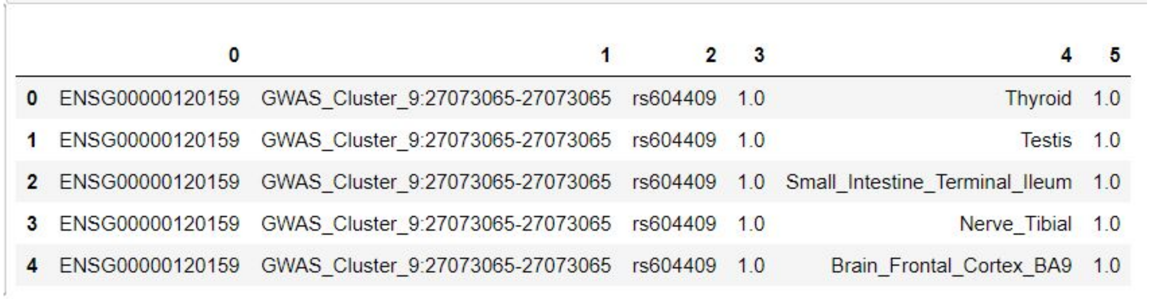
**4. Seaborn :** Seaborn is a python library used to do data visualization. Seaborn is better than Matplotlib in some aspects. For example, it already has an implemented method to generate graphs and figures that are more graphical and with a lot of possible tweaks to have your ideal visual. That's why growing number of people tend to use these new libraries over Matplotlib for example.

### 5.4.1 The data

The data I had at my disposition was a tab-delimited text file. I had already used this type of document in my classes so I had no problem using it, but I had to understand what kind of information I had within my dataset and how to use it in order to obtain what I wanted.

My raw dataset, without any modification was a DataFrame with 78478 rows and 6 columns. Each row is an entry and each column is a piece of information about the data.

Here are my raw dataset's first rows.



| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **0** | ENSG00000120159 | GWAS_Cluster_9:27073065-27073065 | rs604409 | 1.0 | Thyroid | 1.0 |
| **1** | ENSG00000120159 | GWAS_Cluster_9:27073065-27073065 | rs604409 | 1.0 | Testis | 1.0 |
| **2** | ENSG00000120159 | GWAS_Cluster_9:27073065-27073065 | rs604409 | 1.0 | Small_Intestine_Terminal_Ileum | 1.0 |
| **3** | ENSG00000120159 | GWAS_Cluster_9:27073065-27073065 | rs604409 | 1.0 | Nerve_Tibial | 1.0 |
| **4** | ENSG00000120159 | GWAS_Cluster_9:27073065-27073065 | rs604409 | 1.0 | Brain_Frontal_Cortex_BA9 | 1.0 |

Figure 3: Raw dataset with no modifications.

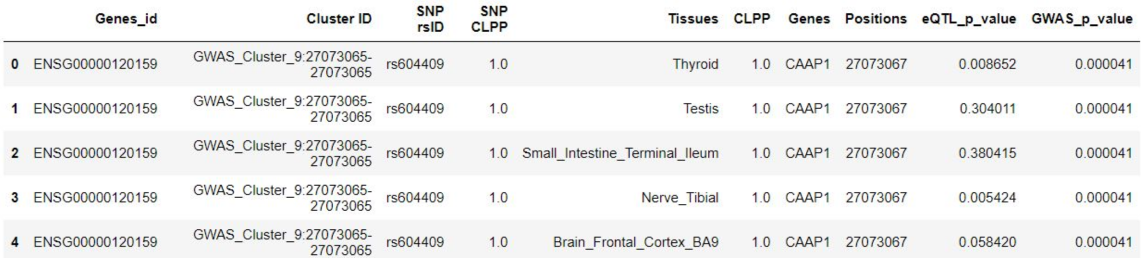
The first column is the identification number of the gene in the ENSEMBL ( EMBL-EBI genome database) database, the second column tells you the GWAS cluster identification number in the database, the third one gives you the SNP rsID, which is a generic way of identifying the single nucleotide polymorphims all over the human genome, the fourth column gives you the colocalisation posterior probability for this particular SNP, the fifth column gives you the tissue used to do the analysis and the sixth and last column gives you the colocalisation posterior probability for the whole cluster.

My second tool to generate the figures in order to analyse the results is another a file we'll call the results file. This file is used in the form of a DataFrame with 6283 rows and 144 columns. It is used as a dictionary.

Before diving further into the tools, let's try to understand why we use it. Our goal is to generate two types of figures, Manhattan plots and Heatmaps. These figures will allow us to check for statistical significance in our dataset using CLPPs and p-values. Our goal will be to find the correct gene and tissue couple under a designated threshold (for example $1 * 10^{-5}$ for the p-value). Then we use this information to validate existing publication or find new potential candidates for causal variants.

Our objective with this document is to find the correct gene identification number and SNP for a given entry in our first output, the raw dataset. For this given entry, there's one and only one row in the results file that correspond.
This row has 144 columns, each column has a single information that has potential interest for a user of this database. For our study, we're interested in the conventional name of the gene in the international standards, the position in the genome of the SNP ( think of the genome as an extremely long double strand, it gives a unique location given by international standards), the expression quantitative loci p-value for this specific gene and SNP for each tissue sample we have and the genome-wide association study p-value for each SNP.

After doing all the modifications, adding the information we want to the original dataset, here is our final dataset we will use to generate the figures.



|   | Genes_id | Cluster ID | SNP rsID | SNP CLPP | Tissues | CLPP | Genes | Positions | eQTL_p_value | GWAS_p_value |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ENSG00000120159 | GWAS_Cluster_9:27073065-27073065 | rs604409 | 1.0 | Thyroid | 1.0 | CAAP1 | 27073067 | 0.008652 | 0.000041 |
| 1 | ENSG00000120159 | GWAS_Cluster_9:27073065-27073065 | rs604409 | 1.0 | Testis | 1.0 | CAAP1 | 27073067 | 0.304011 | 0.000041 |
| 2 | ENSG00000120159 | GWAS_Cluster_9:27073065-27073065 | rs604409 | 1.0 | Small_Intestine_Terminal_Ileum | 1.0 | CAAP1 | 27073067 | 0.380415 | 0.000041 |
| 3 | ENSG00000120159 | GWAS_Cluster_9:27073065-27073065 | rs604409 | 1.0 | Nerve_Tibial | 1.0 | CAAP1 | 27073067 | 0.005424 | 0.000041 |
| 4 | ENSG00000120159 | GWAS_Cluster_9:27073065-27073065 | rs604409 | 1.0 | Brain_Frontal_Cortex_BA9 | 1.0 | CAAP1 | 27073067 | 0.058420 | 0.000041 |

Figure 4: Modified dataset with all the information needed.

We can see some differences with the first dataset I had (3). First, the columns are now named to easily access what we need in the algorithm I develop. I also added 4 new columns, which correspond to the data we needed from the results file in order to correctly generate the figures.

### 5.4.2 The procedure

In order to find the right information at the right place and use it at the right moment, a lot of work went into how to assemble the data so that the users only have to give the input and everything is done with a single line of command using the code available on the GitHub page.

The goal was to have 3 different stages when it came to coding the algorithm.

We want it to run, then run correctly, by giving the correct result and finally run correctly and quickly.

The dataset I was using was a toy dataset, it didn't have much data and it was basically here to allow me to work on data without having to wait too much for the code to run.

Let's now talk about the two figures I used in order to analyse the results.

## 5.5 The Heatmap

The first figure I had to do in order to understand the results was a heatmap.

A heatmap is a data visualisation tool used in order to better understand complex statistical data. It's origin comes from the classic weather maps and it uses the same methodology. You have a lot of data in a matrix or a 2D map of something, and you want to know which areas are the most interesting regarding your value of interest.

In order to produce it, we first had to know what we wanted to see. We wanted to know what gene in what tissue had the highest CLPP for each genome-wide association study cluster.
Looking at my dataset, each row has the gene, the tissue, and the genome-wide association study cluster, we only had to group by one of this values and we would obtain a list of sub-datasets.

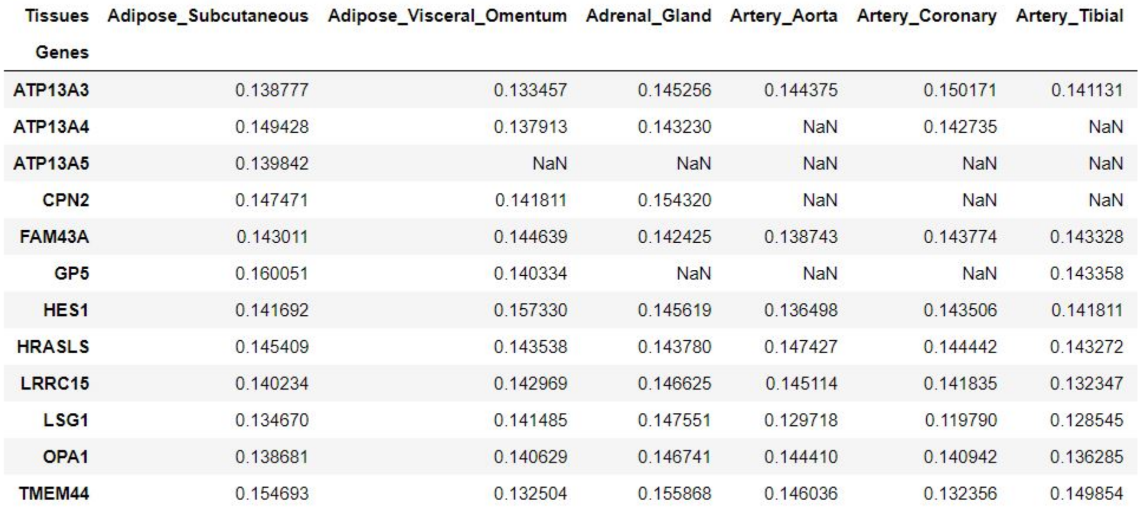
It became clear that the most logical approach was to group by genome-wide association study cluster, because the location of the SNP in the gene is what interests us the most in order to understand what differs between the phenotypes.
I grouped my dataset by clusters and I had approximately 120 genome-wide association study clusters to work with.

Then, for each sub-dataset, with only one genome-wide association study represented, you had a matrix with the same columns as the original dataset, only with less rows.

In order to have a heatmap, we want to transform our matrix into a 2D map with the genes and the tissues in the axis.

To do that, I extract the genes, the tissues and the CLPPs from the matrix and I use the CLPPs as data to put in the heatmap.

Doing that, I find myself with a 2D map, with genes on one axis, tissues on the other axis and the correct data for each cell. Here's an example of 2D map without the colors.



| Tissues | Adipose_Subcutaneous | Adipose_Visceral_Omentum | Adrenal_Gland | Artery_Aorta | Artery_Coronary | Artery_Tibial |
|---|---|---|---|---|---|---|
| Genes | | | | | | |
| ATP13A3 | 0.138777 | 0.133457 | 0.145256 | 0.144375 | 0.150171 | 0.141131 |
| ATP13A4 | 0.149428 | 0.137913 | 0.143230 | NaN | 0.142735 | NaN |
| ATP13A5 | 0.139842 | NaN | NaN | NaN | NaN | NaN |
| CPN2 | 0.147471 | 0.141811 | 0.154320 | NaN | NaN | NaN |
| FAM43A | 0.143011 | 0.144639 | 0.142425 | 0.138743 | 0.143774 | 0.143328 |
| GP5 | 0.160051 | 0.140334 | NaN | NaN | NaN | 0.143358 |
| HES1 | 0.141692 | 0.157330 | 0.145619 | 0.136498 | 0.143506 | 0.141811 |
| HRASLS | 0.145409 | 0.143538 | 0.143780 | 0.147427 | 0.144442 | 0.143272 |
| LRRC15 | 0.140234 | 0.142969 | 0.146625 | 0.145114 | 0.141835 | 0.132347 |
| LSG1 | 0.134670 | 0.141485 | 0.147551 | 0.129718 | 0.119790 | 0.128545 |
| OPA1 | 0.138681 | 0.140629 | 0.146741 | 0.144410 | 0.140942 | 0.136285 |
| TMEM44 | 0.154693 | 0.132504 | 0.155868 | 0.146036 | 0.132356 | 0.149854 |

Figure 5: 2D Map of the Genome-Wide Association Studies CLPPs for each Gene and Tissue couple.

This is exactly what we need in order to assess the statistical significance of each gene for each tissue. By only using this, the problem is that it is difficult to know which cell is the most interesting one, which one is the least interesting one, etc...

That's why we want to use a heatmap.

Instead of using numbers and having to read every number one by one, we use a graphic legend and it allows us to easily know which gene is the most statistically significant and which tissue sample has the highest value. Let's turn this 2D Map into a heatmap and read the most statistically significant gene and which tissue had the highest CLPP.

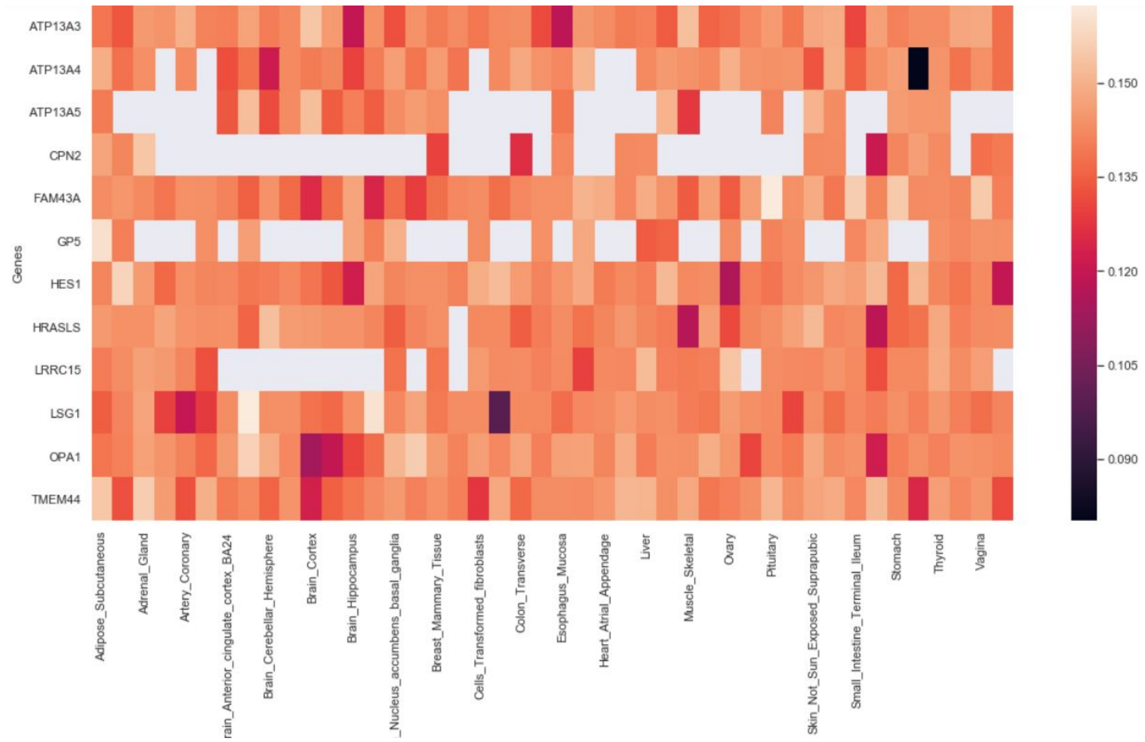


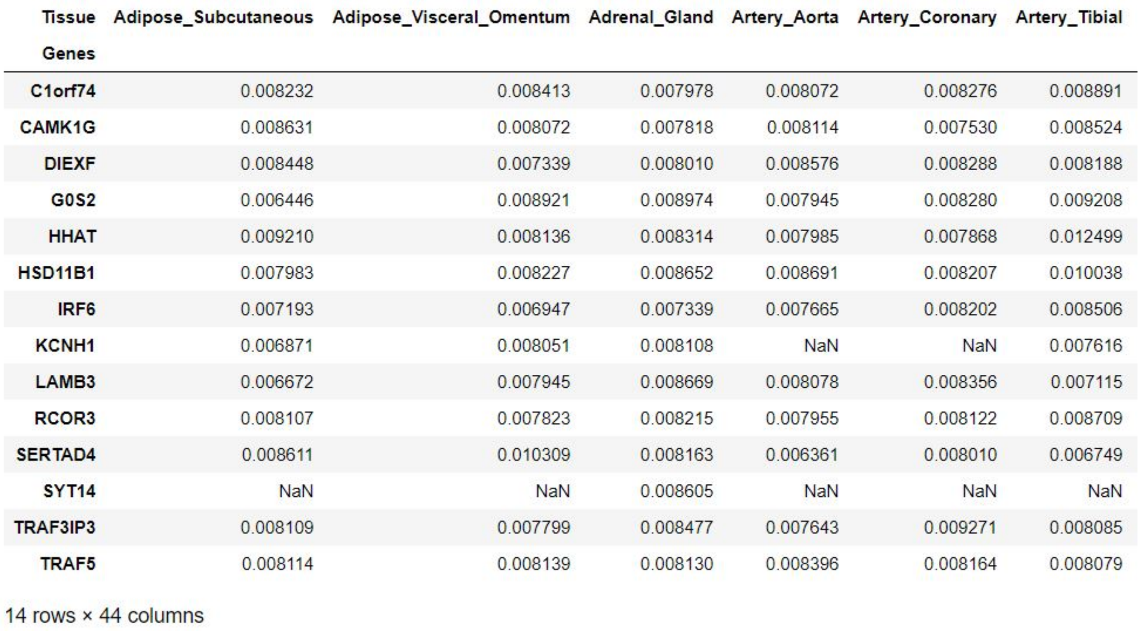
Figure 6: Heatmap of the Genome-Wide Association Studies CLPPs for each Gene and Tissue couple.

The first thing that we see is that there are some cells that are white. These cells are empty

because in the samples we used to generate the dataset there was no eQTL for that gene in that tissue, possibly due to low expression values, which hinder measurement.
Using the legend on the right hand side, we see that the higher the CLPP is, the lighter it gets, the darker it is, the lower it gets. We're not looking at the lowest CLPPs, but the highest ones.

Reading the heatmap, one can see that the gene GP5 seems to be the one with the highest values, and the highest value is for the Adipose Subcutaneous' type of tissue.
These were the results for the toy dataset of approximately 78 000 entries. Let's see if there's a noticeable difference for a bigger dataset, used in new medical research, the UK BioBank dataset, which has more than thirty-two million entries.

Here are the results for a Genome-Wide Association Studies cluster in the UK BioBank dataset.



| Tissue | Adipose_Subcutaneous | Adipose_Visceral_Omentum | Adrenal_Gland | Artery_Aorta | Artery_Coronary | Artery_Tibial |
|---|---|---|---|---|---|---|
| Genes | | | | | | |
| C1orf74 | 0.008232 | 0.008413 | 0.007978 | 0.008072 | 0.008276 | 0.008891 |
| CAMK1G | 0.008631 | 0.008072 | 0.007818 | 0.008114 | 0.007530 | 0.008524 |
| DIEXF | 0.008448 | 0.007339 | 0.008010 | 0.008576 | 0.008288 | 0.008188 |
| G0S2 | 0.006446 | 0.008921 | 0.008974 | 0.007945 | 0.008280 | 0.009208 |
| HHAT | 0.009210 | 0.008136 | 0.008314 | 0.007985 | 0.007868 | 0.012499 |
| HSD11B1 | 0.007983 | 0.008227 | 0.008652 | 0.008691 | 0.008207 | 0.010038 |
| IRF6 | 0.007193 | 0.006947 | 0.007339 | 0.007665 | 0.008202 | 0.008506 |
| KCNH1 | 0.006871 | 0.008051 | 0.008108 | NaN | NaN | 0.007616 |
| LAMB3 | 0.006672 | 0.007945 | 0.008669 | 0.008078 | 0.008356 | 0.007115 |
| RCOR3 | 0.008107 | 0.007823 | 0.008215 | 0.007955 | 0.008122 | 0.008709 |
| SERTAD4 | 0.008611 | 0.010309 | 0.008163 | 0.006361 | 0.008010 | 0.006749 |
| SYT14 | NaN | NaN | 0.008605 | NaN | NaN | NaN |
| TRAF3IP3 | 0.008109 | 0.007799 | 0.008477 | 0.007643 | 0.009271 | 0.008085 |
| TRAF5 | 0.008114 | 0.008139 | 0.008130 | 0.008396 | 0.008164 | 0.008079 |

14 rows × 44 columns

Figure 7: 2D Map of the Genome-Wide Association Studies CLPPs for each Gene and Tissue couple using the UK BioBank Dataset.

We can see that even if I had a toy dataset, the tissue samples used were the same, so even if the dataset is bigger, all the methods coded for the toy dataset work for the larger one.

Let's now take a look at the Heatmap.

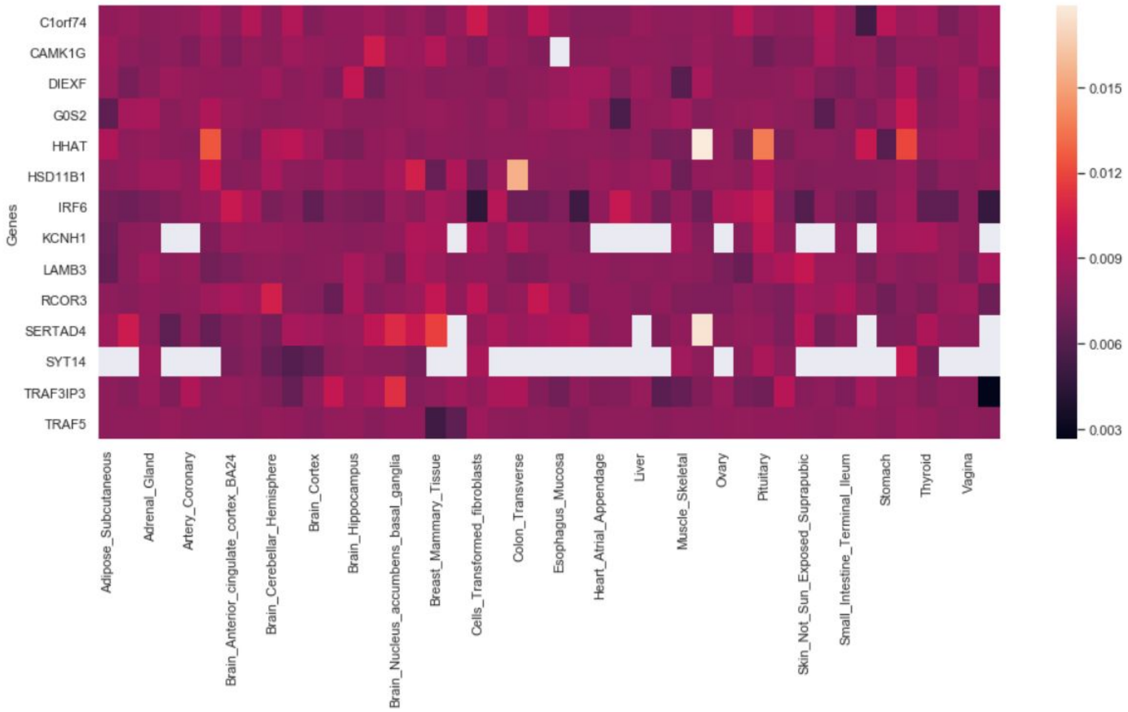


Figure 8: Heatmap of the Genome-Wide Association Studies CLPPs for each Gene and Tissue couple using the UK BioBank Dataset.

Here one can see that there are less missing values, meaning that for each tissue sample used, all or most of the genes were successfully tested. However, there are still some missing bricks, especially for the gene SYT14 for this heatmap, and this is something that someone has to know when working in the research, especially in human related research, data is imperfect.

It lacks completion, it can be biased, it is rare and a lot of work is put into trying to overcome these flaws by generating new and better datasets like the UK BioBank.

Let's get back to reading the heatmap, two rows seem to be quite brighter than the others for this genome-wide association studies cluster. The genes HHAT and SERTAD4 seem to be the most statistically significant, with several tissues samples showing a high CLPP, especially for the pituitary tissue sample for the HHAT gene and the Breast Mammary Tissue for the SERTAD4 gene.

This last heatmap wasn't chosen randomly. The goal of our study is to find evidence linking a gene to a particular disease. A way to check if everything is correct and no error go unseen is to find a publication in the same field that found evidence linking genes to diseases and then trying to validate this study with our results analysis in our dataset.

That's why we used a publication identifying 64 novel causal variants related to Coronary artery disease [28].

In this publication, a genome-wide association study over 34 541 cases and 261 984 controls of the UK BioBank database allowed to identify new locis that were not reported in previous research with various thresholds regarding the p-value to have a limited set of potential candidates for causal variants.

My goal was to see if I could find similar results by generating figures to analyse the datasets we use to validate our method.

A statistically significant loci was found with several candidate genes and the SERTAD4 and HHAT genes were possible candidates. The observation of the heatmap can help validate this result and reassure us regarding the method used to generate our figures, all seems to be well and it can be used to investigate without validating existing research again.

Now that we went over the heatmap in detail, let's take a look at our other figure in order to analyse the results of our algorithm, the Manhattan plot.

## 5.6 The Manhattan plot

**Manhattan plot :** A Manhattan plot is used to display a large number of data points in a plot, it is commonly used in genome-wide association studies to identify the statistically significant SNPs. In general, the position of the SNPs in the genome is used as the X axis, and $-log10$(p-value) is used as the Y axis [29].

Here is an example of Manhattan Plot for the whole genome :

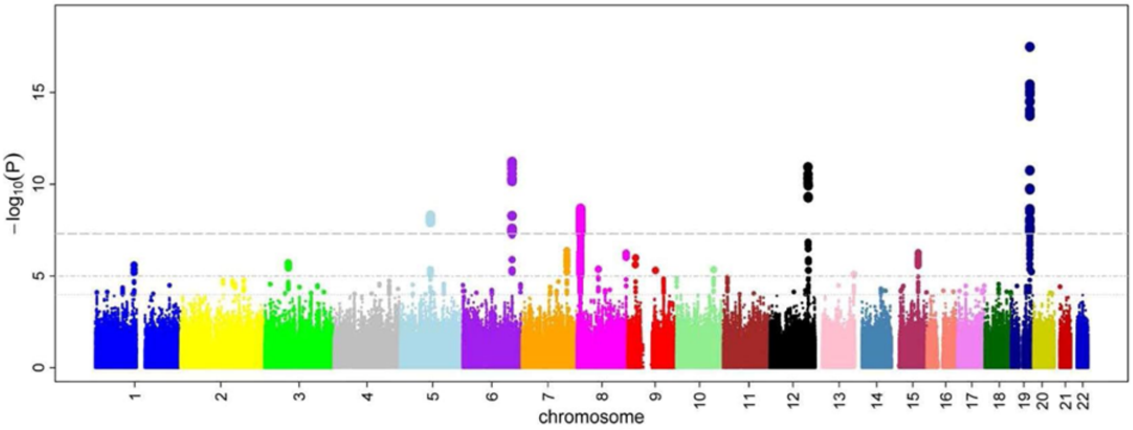


Figure 9: Manhattan Plot showing statistically significant loci.

Each Manhattan plot is generated for the expression of a gene in a specific tissue. The goal is to find the most statistically significant SNP's position in the genome to study it.

For our toy dataset, the main problem with the Manhattan plots was that there were a lot of possible Manhattan plots and each one was available but most of the time it consisted of only one or two points in the plot.

This isn't really interesting to have one or two points in a Manhattan plot, the goal is to have a good number of points to have a mean and select the statistically significant p-values that stand out.

To try to have the most significant possible Manhattan plot for the toy dataset, we checked the length of each Manhattan plot that we can generate and selected the maximum length.

We found out that seven points was the maximum and selected a random Manhattan plot with seven points.

There are three kinds of Manhattan plots we want to generate, depending on the Y axis we want to use.

We'll only discuss one kind of Manhattan plot because the other ones follow the exact same procedure, and the plot doesn't change that much. We'll use the Manhattan Plot using the p-value for a particular SNP here as an example of Manhattan plot.

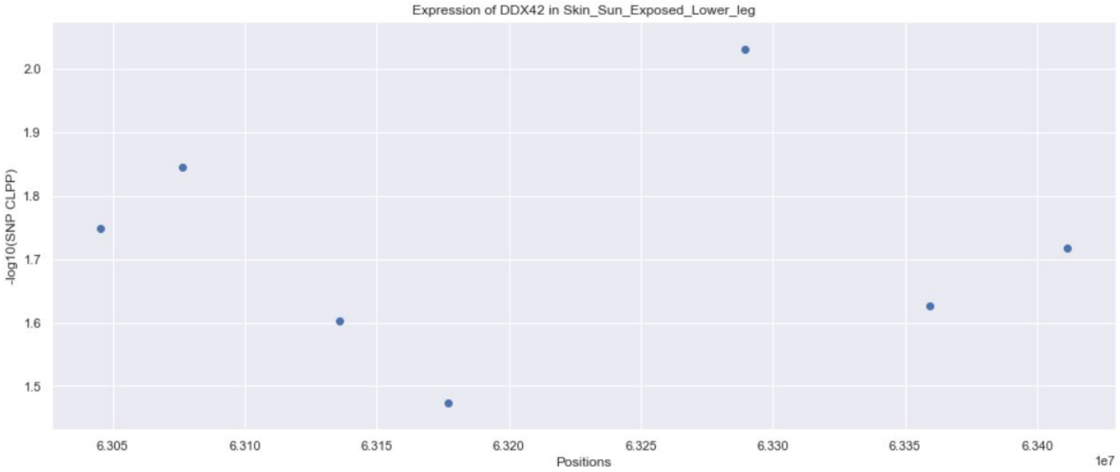


Figure 10: Manhattan Plot of the SNP Colocalisation Posterior Probability Between the expression of Gene DDX42 in the Skin Sun Exposed Lower Leg Tissue (from GTEx) and CAD (from the toy dataset).

We can see that there are a few points, but most probably the value over two near the position $6.330 * 10^7$ is the most statistically significant for this sample, even if the threshold is quite higher, for values over 5 or 8, as we use the $-log10$ of the quantity we're interested in.

Now let's have a look at a Manhattan Plot for the real UkBioBank dataset, using the gene and tissue couple we found to be interesting in the Heatmap.

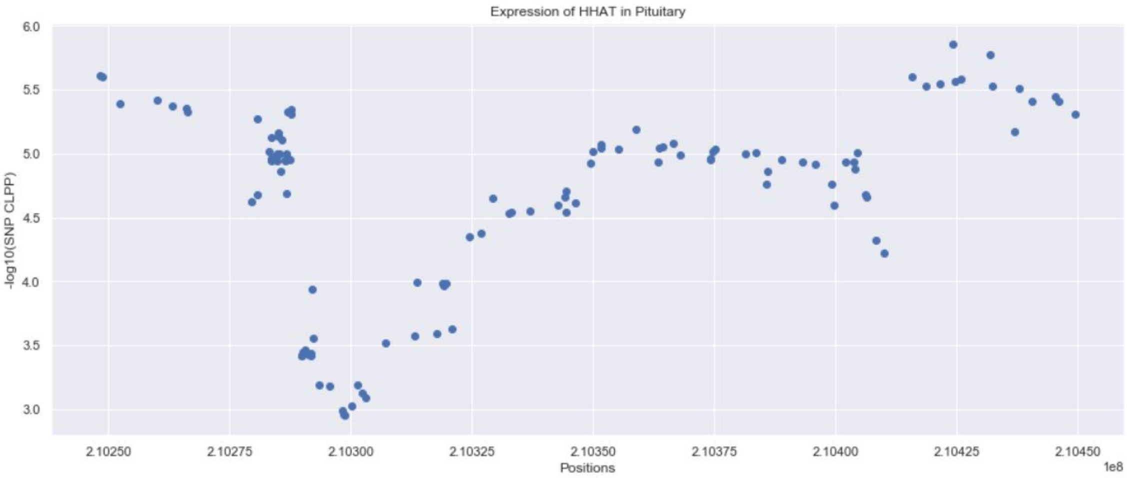


Figure 11: Manhattan Plot of the SNP Colocalisation Posterior Probability Between the expression of Gene HHAT in the Pituitary Tissue (from GTEx) and CAD (from UKBioBank).

We see now that with the UK BioBank Dataset, we have a "real" Manhattan plot with quite a lot of data points, 117 to be precise. We also note that when we had a little over 2 as a maximum for the last Manhattan plot with the toy dataset, we now have a lot of data points over 5, which can be used a solid threshold to look at potential candidate for causal variants linked to the disease studied.

Once again, this result confirms the results found in the publication, which allows us to have some confidence in the algorithm we developed.

The same procedure is used to generate the Manhattan plots using Genome-Wide Association Studies p-values and expression quantitative trait loci p-values. The only difference lies in the position of the points in the plot, we'll leave the figures to the annex to try not to be too long.

Now that we detailed the generation of the figures in order to analyse the results of the algorithm developed, let's try to understand what's behind this algorithm, what are the statistics and mathematics principles used to have these statistical results.

## 5.7 The problems encountered during the internship

During my internship, I had to face a variety of problems, on all kinds of subjects.
First of all, the first problem rose after the beginning of COVID-19 crisis. My landlord wanted me out of my apartment as soon as possible just after lockdown had been declared effective in the UK. This took me by surprise, I wasn't ready because I didn't expect it.

After some discussion she made it clear it wasn't negotiable so I resigned myself and tried to find something else. Fortunately, my team leader and my team in general have been very supportive, I was quite in a shock at that moment and I felt down but everything unraveled rather quickly, I found a new accommodation the next week and moved out the week after that.
It was easier than I expected and I truly want to thank my team and my team leader, Pr Daniel Zerbino in particular, for his help in this difficult period of time.

Secondly, regarding the code, I wasn't used to work on someone's code and try to enhance the performances or understand it from scratch without a really well detailed course on how it works. I lost quite some time trying to adapt and do it by myself but I found it quite hard and I hope it serves me as a learning experience in order to be more productive in my next missions.
Thirdly, regarding the code, there were a lot of tweaks I was partly aware of about the complexity that I wasn't used to use in my algorithms because we were using toy datasets but it became clear that I needed to use coding techniques in order to drastically reduce memory consumption and

CPU usage.

This allowed me to increase my knowledge about coding methods and good manners to try to be optimal while writing the algorithm as soon as possible and not wait future iterations with the team to try and optimize the code.

All in all, my internship allowed me to gain experience in various fields, human relations, algorithmic, code comprehension, and I hope it will serve me well for my working career.

# 6 Conclusions and perspectives

To conclude this report, I'd like to say that this internship has been a great experience for me. I had the opportunity to work in a wonderful environment, in a great team and have encouraging results even in this difficult situation where everyone had to work from home and asking for help was a little bit tougher.

To sum up the internship I can say that I learned and re-learned a lot about Biology. The goal was for me to gain competence and skills and without any pressure, which was enjoyable.

I consolidated my skills in figure generation, data cleaning, statistics.

I learned about working in a team and more importantly on an ongoing project that started years ago. It is not something that we learn in school as we usually code everything from scratch.

I think I will continue my journey into datascience applied to biology and biostatistics because I enjoyed working in this field. It is still in an exploratory phase, a lot of work has to be done in order to enhance the performances of the algorithms used and when it will be done, it could improve the life of thousands of people, which makes me feel like continuing in this path.

# 7 Bibliography

[1] E. Moyou, Marché pharmaceutique : chiffre d'affaires mondial 2001-2017, 2019

[2] Matej Mikulic, Global Pharmaceutical Industry Statistics & Facts, Statista, 2019.

[3] Bozenhardt, Erich H. & Bozenhardt, Herman F., "Are you asking too much from your filler ?", Pharmaceutical Online, VertMarkets, 2018.

[4] Max Roser, Human Development Index (HDI), Our World in Data, 2019.

[5] Number of Deaths due to HIV, World Health Organization Data, 2020.

[6] Rambaut, A., Posada, D., Crandall, K. et al. The causes and consequences of HIV evolution. Nat Rev Genet 5, 52–61 (2004).

[7] EFPIA, The Pharmaceutical Industry in Figures, Key Data, 2019.

[8] Fourth Quarter and Full Year Results 2019, Sanofi, 2020.

[9] Pharmaceutical Industry in France, Finesco Inc., 2012.

[10] Novo Nordisk Annual Report 2019, 2020.

[11] G. Koscielny et al., Open Targets: a platform for therapeutic target identification and validation, Nucleic Acids Research, Vol. 45, 2017.

[12] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. (2015) UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of

Middle and Old Age. PLoS Med 12(3)

[13] Vamathevan, J., Clark, D., Czodrowski, P. et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18, 463–477 (2019).

[14] Eggert, U. The why and how of phenotypic small-molecule screens. Nat Chem Biol 9, 206–209 (2013).

[15] How Different Antidepressants Work, WebMd

[16] T.A. Brown, Genomes Third Edition, Garland Science Publishing, 2007

[17] Griffiths, Anthony J.F.; Miller, Jeffrey H.; Suzuki, David T.; Lewontin, Richard C.; Gelbart, eds., An Introduction to Genetic Analysis, Genetics and the Organism: Introdution 2000.

[18] Miko, I. & LeJeune, L., eds. Essentials of Genetics. Cambridge, MA: NPG Education, 2009.

[19] D.J. Balding; M. Bishop; C. Cannings, Handbook of Statistical Genetics Third Edition, Wiley, 2007

[20] Slatkin; Montgomery, Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics. 9 (6): 477–485. 2008.

[21] Lewontin, R. C., The interaction of selection and linkage. I. General considerations; heterotic models, Genetics. 49 (1): 49–67, 1964.

[22] Iryna Nikolayeva, MAT5013 Statistical learning in high dimension for biological data, Télécom SudParis, 2019.

[23] DNA Replication - Machinery And Enzymes, Byju's, 2019

[24] John W. Drake,* Brian Charlesworth,† Deborah Charlesworth† and James F. Crow, « Rates of Spontaneous Mutation », the Genetics Society of America, 1998

[25] Single Nucleotide Polymorphism, Scitable, Nature Education, 2015

[26] Gibson, G. Hints of hidden heritability in GWAS. Nat Genet 42, 558–560 (2010).

[27] Hormozdiari, Farhad et al. "Colocalization of GWAS and eQTL Signals Detects Target Genes." American journal of human genetics vol. 99,6 (2016): 1245-1260. doi:10.1016/j.ajhg.2016.10.003

[28] Van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ Res. 2018;122(3):433-443. doi:10.1161/CIRCRESAHA.117.312086

[29] Laura Sanders, "40 more 'intelligence' genes found", Science News, 2017

[30] Lamin A Truncation in Hutchinson-Gilford Progeria, Annachiara de Sandre-Giovannoli, Rafaëlle Bernard, Pierre Cau, Claire Navarro, Jeanne Amiel, Irène Boccaccio, Stanislas Lyonnet, Colin L. Stewart, Arnold Munnich, Martine Le Merrer, Nicolas Lévy, Science, 2003