Facultad de Ciencias

Universidad Zaragoza

Application of Mendelian Randomization in the inference of gene regulatory networks

Aplicación de la Aleatorización Mendeliana en inferencia de redes de regulación genética

Trabajo Fin de Grado

Author:

Claudia Llop Moreno

Tutor:

Dr. Joaquín Sanz Remón

Grado de Física.
2021-2022

Índice

1. Summary

The regulation of gene expression in living systems is a process that depends on the participation of thousands of genes, proteins and metabolites that interact with each other generating network patterns that constitute the paradigmatic example of a complex system. These gene regulatory networks (GRNs), which in their most fundamental formalization are composed of genes whose expression levels are regulated by each other, constitute a convergent object of study of the physics of complex systems, genomics and systems biology.

However, inferring the directionality of the causal relationships that form these networks is not a simple task, unless gene expression experiments (e.g. RNA-seq) are combined with additional experimental evidence addressing, for example, the presence of interactions between regulatory proteins and DNA elements (e.g. ChIP-seq). When such multi-modal data sets are not available, and the only molecular phenotype available is gene expression, it is much easier to obtain co-expression networks (where links denote correlation) than to obtain regulatory networks (where links are directional and indicate causality). The literature (1)(2) on inference of regulatory networks is extensive, but it is often complicated to infer (directed) causal relationships from (undirected) co-expression patterns, as interactions are sometimes not clearly mediated by a single molecular mechanism, among other problems.

Building on previous work by the research team, in this dissertation we aim to explore the potential of Mendelian randomization (MR) to infer the directionality underlying the correlation patterns that form the basis of a co-expression network. MR is a statistical method commonly used in epidemiology that allows us to infer causal directionality between two phenotypes using instrumental variables (typically genetic variants) and their relationship with the analyzed phenotypes, provided that certain conditions meet. While the classical scope of MR are high-order, one-dimensional phenotypes, here we propose its massive, parallel application to resolve the direction of the hypothetical causal effects underlying co-variation in thousands of gene co-expression pairs simultaneously.

In this TFG, the method will be applied to a panel of broadly genotyped individuals, from whose blood, our collaborators derived macrophages and collected RNA-seq gene expression data.

From a scientific training perspective, the main objective I pursued during this TFG was to learn the basic concepts underlying some of the main experimental methodologies in contemporary genomics, especially transcriptomics and genotypic characterization of genetic variants, as well as to gain relevant experience in the implementation of complex computational pipelines for their analyses. As a result, we have proposed a pipeline that allowed us to infer significant causality profiles for 69154 gene pairs (47.3 % of all interactions that we analyze through MR) connecting as many as 815 genes expressed in human macrophages. Thanks to the implementation of two different selection strategies for the instrumental variables to use, our results are essentially unaffected by the heterogeneous quality of the instrumental variables that are available to analyze causality among the different genes under study, as we show from the comparison to an empirical null model of directional bias.

2. Introduction

2.1. Mendelian Randomization and its epidemiological applications

Observational epidemiological studies are prone to confounding, reverse causation and various biases and have generated findings that have proved to be unreliable indicators of the causal effects of modifiable exposures on disease outcomes. (4) Mendelian randomization (MR) is an statistical technique commonly used in epidemiology for the study of causal effects between observational data in the presence of confounding factors. It uses the measured variation of genetic factors of known function on one of the phenotypic variables under study in a population, as instrumental variables to ascertain the presence of a causal effect from the phenotype that is impacted by the instrumental variable (phenotype X) on the secondary phenotype (referred to as Y). In many examples, one of these phenotypes is associated to an environmental exposure (X), while the other one is often related to the population risk associated to a pathological phenotyope or a disease (Y). (5).

The approach leans on the idea that the genotype is randomly assigned due to meiosis and can thus be considered as an instrumental variable (G). Intuitively, this means that, if the correlation found in the population between phenotypes Y and X is due to a causal effect on X over Y, then, the effect of G over X should also propagate onto Y. Instead, should the X-Y correlation arise from a confounding factor, the effect of G over Y should not be significant. For a genetic variable to be considered a valid instrument for determining the causality from X to Y during MR it must verify a series of conditions. First, these variables (G) should not be directly related to the outcome Y (disease), but only eventually through an indirect path involving the phenotype X. In turn, both phenotypes Y and X (but not G) may be influenced by confounders U. This network of relationships between the ingredients of the MR is described in the following DAG (Direct Acyclic Graph) [1]:

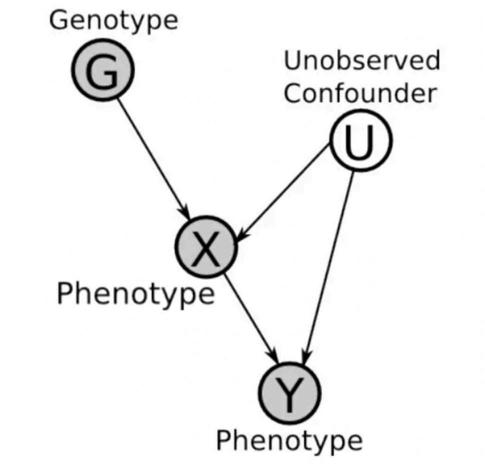


Figura 1: Diagram of Mendelian Randomization Instruments.

For MR to constitute a valid methodological approach to test for the existence of a causal link from X to Y, there are three conditions that must be met at once. First, the genetic variants G must explain a significant fraction of variance of phenotype X (relevance assumption). Second, while variables X and Y may be affected by common confounders, none of these shall affect the genetic variant itself (independence assumption). Last, no causal relation from G to Y must exist, except for the indirect pathway involving X (no horizontal pleiotropy assumption).

To understand the potential of MR in epidemiology, let us consider a practical example. Suppose that we wanted to investigate the effects of alcohol consumption (”phenotype”X, in this case an environmental exposition) on blood pressure (phenotype Y) as a means to understand the overall relationship of alcohol with risk of coronary heart disease (6).

One source of evidence is the association between alcohol and blood pressure in observational studies. This association may be a poor indicator of the causal effects of alcohol if there are other factors —“confounders” (U)— that influence both alcohol intake and blood pressure. People who consume more alcohol may also have other risk factors for cardiovascular disease, such as smoking more heavily than those with lower alcohol consumption, follow less healthy diet habits, and/or be subject to stronger levels of social adversity and higher stress. In this hypothetical scenario, the confounding factor U (e.g. smoking) induces a positive association between the risk factor X (alcohol) and an outcome Y (blood pressure). In this case, interpreting this association as evidence of a causal relation between alcohol consumption and blood pressure would be misleading. To deal with this problem, many epidemiological methods attempt to correct for, or minimise, observed differences in confounders between study participants. These methods can give useful evidence about causal relations if we measure enough confounders so that, after adjustment or matching, study participants who consume different amounts of alcohol become comparable. However, this type of approach presents many problems. On the one hand, it relies on our ability to identify all the relevant confounders, and even when the confounders have been identified, measurement errors may translate into residual confounding, even after apparent statistical adjustment.

In these cases, MR constitutes a highly instrumental approach instead, both to detect direct (from X to Y) and reverse (Y to X) causation, provided that valid instrumental variables are available for both correlated phenotypes X and Y. In our example, reverse causality would be observed if people with symptoms of cardiovascular disease (i.e. high blood pressure), for example, may reduce their alcohol consumption with respect to those without symptoms, as a response to a high blood pressure diagnosis. This would lead to a negative association between alcohol consumption and blood pressure levels, which, again should not be interpreted as evidence that consuming alcohol causes a reduction in blood pressure levels.

In the mentioned example, an adequate genetic instrumental variable is available for testing the causality link from X to Y: the SNP variant rs671. rs671 is a genetic variant of the ALDH2 gene found in East Asian populations whose less frequent allele (A) slows the metabolism of acetaldehyde, causing an adverse response to alcohol consumption. In a study of selected individuals from the general population, male who carried two copies of the A allele drank an average of 1.1 g of alcohol per day, while those with no copies drank 23.7 g. Therefore, if high alcohol consumption is really the cause of high blood pressure, patients with two copies of the A allele should present, in average, lower blood pressure. If this were not the case, the levels would be similar in both groups. Similarly, any genetic variant physiologically related to physiological regulation of blood pressure unrelated to alcohol metabolism (e.g. the SNP rs13107325, a non-synonimous variant in the SLC39A8 gene, encoding a transporter protein of divalent metal cations such as cadmium, whose accumulation has been linked to increased blood pressure (7)) could be used to run MR in the opposite direction.

If MR is a statistical approach used in epidemiology to characterize causal relations between phenotypes under partial genetic control, in this TFG we plan to apply it to the characterization of gene regulatory networks as described in the following section.

2.2. Leveraging MR to characterize Gene Regulatory Networks.

In this work, we propose using MR as a useful resource to infer causal regulatory relations between the cross-sectional expression patterns of pairs of genes. To do that, we need to estimate gene expression from RNA-seq transcriptomic data on a large cohort of individuals and analyze the pairwise co-expression matrix of the genes under scrutiny. Whilst this operation is useful to characterize large amounts of X-Y phenotype pairs (each genotype is the expression of a gene), the instrumental variables will be regulatory genetic variables (in our case, SNPs), affecting the expression levels of each of these genes individually (instrumental variables G).

In this context, the genetic variables chosen will be the so-called expression quantitative trait loci (eQTL), which are genetic variants whose alleles modify significantly the expression of a target gene. The vast majority of eQTLs that can be identified through computational analysis act only locally (that is: the active allele only modifies the expression of the gene in the same chromosome) on its target, which is why these are referred to as cis-eQTL. The reason why cis-eQTL are easier to identify is that they typically lie in the vicinity of the target genes, which means that for their identification the testing space can be restricted to the nearby region of each target gene. In these cases, cis-eQTLs are usually associated to sequence modification on regulatory regions of the target gene, paradigmatically, binding sites of transcription factors.

Considering this, cis-eQTLs are highly valuable instrumental variables for MR, since they comply with the main assumptions underlying MR, providing that only adequate gene pairs candidates are analyzed. First, concerning the relevance assumption (that the eQTL explains sufficient variance of the target gene under analysis), it is possible to restrict our method to genes with at least one nearby SNP fullfilling this condition (that is, at least one eQTL). Second, concerning the independence assumption, since genotypes are randomly distributed in a population, there is no reason to expect strong common confounding factors affecting the genetic variables and gene expression at once. The most troublesome assumption in our case may arguably be the assumption of no horizontal pleiotropy, that states that any genetic variant affecting gene X should not affect independently any conjugated gene Y that is to be analyzed. Even so, we can be sure that our variants are mainly no pleiotropic, at least at first order, by considering only sufficiently far away gene pairs and restricting the selection of instrumental variables to cis-eQTLs in the vicinity of target genes only. This is because the probability of finding strong genetic effects on gene expression decays quickly with the distance between the variant and the target gene. Finally, the fact that we can typically identify multiple eQTLs per gene further reduces the risk of considering instrumental variables that are significantly pleiotropic, for the probability that all cis-eQTLs of a gene may show a common pleiotropic effect on other gene are reduced, specially if their genotypes are not strongly correlated.

These selection criteria, considered together, allow us to restrict the testing space only to gene pairs and instrumental variables for which horizontal pleiotropy can be reasonably discarded.

Summing up, for the development of this strategy, three types of independent analyses are performed to find the directionality between the pairs of genes studied. First, we need to define the co-expression networks. Second, we need to perform eQTL mapping to identify all putative instrumental variables, and select only those which will be of use. Finally, we Apply the Mendelian Randomization analysis to the results of the previous steps. The latter analysis (RM) aims to characterize what fraction of the correlation levels found between various pairs of genes is explained by genetic effects on each of the genes involved in the interaction, or in other words, what is the component of these correlations that can be traced back to a genetic effect.

A broad description of the basic concepts related to the building blocks of the work, including the types of experimental data we will deal with (RNA-seq and genotyping chips), as well as the analytic modules that we plan to combine in an integrated pipeline (co-expression networks, eQTL mapping and MR itself) can be found in the appendix.

2.3. Context and scope of this TFG: further requirements for MR to infer GRNs

Applying MR in parallel to thousands of highly correlated phenotypes such as the expression levels of all genes detected in a given transcriptome imposes additional challenges with respect to simpler applications of the MR method.

Specifically, since the main outcome of the analyses is the distinction of the direction of causality (i.e. forward vs. backward) in multiple gene pairs at once, it is extremely important to ensure that the instrumental variables chosen in both directions do not introduce any significant bias in the results.

This observation constitutes, in fact, the starting point in my involving in this research project, which finds its most direct precedent in the TFG by my colleague Regina Santesteban Azanza (8) who submitted her own TFG, titled *”Métodos computacionales para la caracterización de relaciones causales entre genotipo y fenotipo: Heredabilidad de la expresión y coexpresión genética”*, in 2020 at the University of Zaragoza, under the supervision of Joaquín Sanz and Pierpaolo Bruscolini. In her work, Regina produced an implementation of the main analytic modules needed to infer the GRNs (coexpression network inference, eQTL mapping and MR); which lead her to produce a first preliminary version of an algorithm for GRN inference based of MR. The evaluation of the results from her work led Regina to identify a crucial limitation of the approach, according to which the heterogeneity in the amount of instrumental variables used in MR in each direction affected the relative levels of statistical significance found for each link. According to this first version of the algorithm, our ability to detect significant link directions was strongly dependent, more than anything else, on the number of instrumental variables available, which obviously introduces an external bias to our link direction inferences, based on MR significance assessments.

The work now presented is but a continuation of the previously mentioned work, while applying several improvements to solve the problem that was faced at its conclusion. Specifically, after observing Regina’s results, indicating that the number of instrumental variables bias our ability to detect significant causal fluxes, we have completed two essential analytic tasks. On the one hand, we designed and implemented two different algorithms to down-select the genetic variants to ensure that the number of instrumental variables used in both genes within each MR pair is balanced. On the second hand, we designed a computational strategy to evaluate the efficacy of our two methods to remove directional bias from our results. This is necessary, because, even if our algorithms ensure that the number of instrumental variables is leveled, this may yet feature heterogeneous relevance in explaining expression variability of the genes that they target. To do that, we designed a stochastic algorithm to simulate the distribution of percentages of links inferred from gene X to gene Y and viceversa, under an analogous scenario to the real case, while assuming no statistical bias present.

3. Methods

In this section I will describe the main steps necessary to complete the analytic pipeline proposed, starting from the raw RNA-seq and genotyping data, to the characterization of gene regulatory pairs.

Concerning these, I have capitalized on code initially available from (8) for data pre-treatment (section 3.1), and co-expression networks (appendix section 2.1), which I have slightly modified to adapt them to our criteria. Instead, the implementations of the analyses corresponding to the last steps (eQTL mapping -appendix section (2.2)-, MR -section (3.4)-, computational strategies for down-selection of structural variants -section (3.5)-, and empiric evaluation of directional bias (section 3.6), have been essentially written from scratch, to implement and evaluate the new algorithms for the selection of instrumental variables here proposed.

3.1. Pre-treatment of RNA-seq data

For the realization of this project, we started from a raw gene expression matrix where the results of mapping the raw sequencing data to the reference human genome are stored. In this matrix, featuring 58051 rows and 90 columns, the entry (i, j) captures the number of cDNA fragments that have been mapped to the i-th gene in the j-th individual. The samples correspond to the transcriptomes of human macrophages derived from peripheral blood cells extracted from a panel of 90 genetically diverse donors.

From this raw-data object, we first select those genes that encode proteins (19651) and are also located on autosomal genes (18727).

From the filtered version of the raw RNA-seq data matrix, the implementation of analyses that involve comparisons between expression levels across samples cannot be done directly, since, obviously, the number of fragments mapped to a given gene in a given sample depends both on the sampling depth at which that sample has been sequenced as well as on its complexity (i.e. distribution of gene relative frequencies in the population of cDNA fragments sampled). Since the popularization of RNA-seq as the technology of choice to measure genome-wide gene expression levels, different approaches have been proposed to tackle this question, among which, the subsequent usage of the so-called "Trimmed M Means algorithm"(TMM), followed by the transformation of raw data to extract (normalized) logarithmic counts per million has become one of the most widely adopted, specially for analyzing data with a large number of samples. For additional information, the procedure followed by the TMM algorithm is described in the appendix.

Once the data has been properly normalized and transformed, we seek to identify, and remove lowly expressed genes. To do so, we calculate the median logarithmic counts per millions, as provided by the voom transformation, that are observed for each gene across samples, and selected only those genes verifying $median(log(cpm)) > 1$. This reduces the number of genes to be considered to just 10586.

Remove batch effects

In our experiment, we are analyzing an essentially homogeneous set of samples, the only variable in our design is technical, and it corresponds to the experimental batches in which the samples were processed, since, due to the large sample size, it was not possible to process all the data at once. In biology, and especially in genomics, taking into account the variation in experimental results that is verified between batches is an absolutely central issue.(10)

To eliminate these technical batch effects, a regression model is run in which the differences in the averages of the expression levels associated with each batch are subtracted. To do this, we will use the following equation to find the matrix B that represents the variation of expression of each gene in relation to the model variables. We will also find the matrix E that corresponds to the residual matrix.

$$Y^T = D \cdot B + E \tag{1}$$

In this equation, matrix Y or modeling matrix contains gene expression data (rows) for each individual (column). The D matrix captures the experimental design, that is, in our case, the information about which samples were processed in which batch, while B contains the coefficients inferred for each gene for the effect of each of the n=9 batch levels different from the reference. Last E corresponds to the residuals of our data.

Specifically, our matrix Y or modeling matrix is defined by:

$$y_{ij} = log_2\left(\frac{r_{ij} + 0{,}5}{R_j + 1} \cdot 10^6\right) \tag{2}$$

Where $R_j = N_j * TMM_j$. This equation is formed via the *voom* tool in charge of performing the TMM normalization (Trimmed Mean of M values), as explained above.

Once the linear model is completed, and the coefficients B, ,as well as the residuals E are inferred, it is interesting to see for each gene what percentage of variance is explained by the batch effects, captured by the coefficient of determination R^2. The density plot (Figure 2) of this coefficient in our data is shown below.

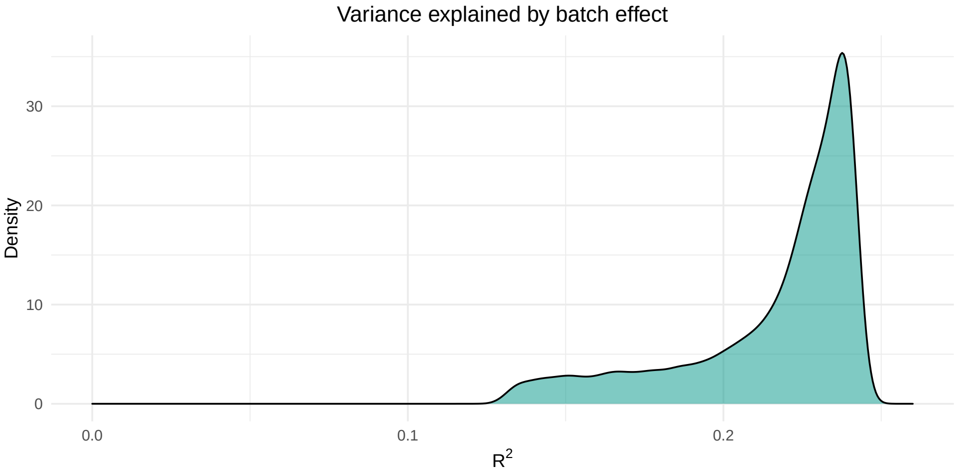


Figura 2: Variance explained by Batch effect.

As we see in 2, batch effects do exert non-neglectable effects on gene expression, explaining an avearge of circa 25 % of the variance in the data, and therefore, they must be corrected.

Variance stabilization

Before starting with the analysis of the correlations among genes, and between SNPs and gene expression, we must take into account that there is a systematic relationship between the average number of reads that we can detect for a given gene and the technical dispersion (noise) subject to the measurements of each individual for each gene. This fact, commonly referred to as a form of heteroskedasticity, represents an important technical problem, as it can introduce biases and increase the noise in the data, preventing us from characterizing real co-expression relationships, and spuriously inducing others. To solve this problem, one of the most straightforward approaches consists of performing a variance stabilization procedure. (11)

This procedure consists of studying the empiric relation that is typically found in RNA-seq data between the residual variance of the genes (variance of residuals after regressing out any relevant, technical and biological effects) and mean, raw expression. This approach is implemented in a number of software libraries dealing with NGS data, both at bulk and single cell resolution (e.g. *limma*, *scran*, *variancePartition*), and is widely used in the transcriptomics literature.

Operationally, the method consists of modeling the total residual variance as the sum of a technical (which depends on mean expression), and a biological component $\sigma^2(\epsilon) = \sigma^2_{tech} + \sigma^2_{bio}$, where the technical component σ^2_{tech} is inferred as a function of mean raw expression upon a local regression algorithm. More precisely, once the residuals of our experimental design (after removing batch effects) are available, we represent their dispersion (more precisely, a power σ^β with $\beta < 1$), against the mean raw expression, as we do in Figure 3, and infer, using a local regression algorithm the systematic trend (red line in Figure 3), commonly referred to as the technical dispersion trend and denoted $\hat{s}^\beta$. From the result of that fit, we infer $\sigma^2_{tech} = (\hat{s})^{(2/\beta)}$. The rationale underlying the usage of an exponent β to transform the response variable prior the local regression, instead of modeling directly the expression variance -which would be cleaner from an analytical perspective-, is to reduce the weight of variance outliers (e.g. highly variable genes) on the trend fit.

Once that is done, the residual vectors, by construction centered at zero, are re-scaled by the technical trend as follows: $\hat{\epsilon} = \epsilon/\sigma_{tech}$, which leads to a new vector of variance-stabilized expression values where the dependence between technical noise and mean expression is abrogated, as we represent in Figure 3.

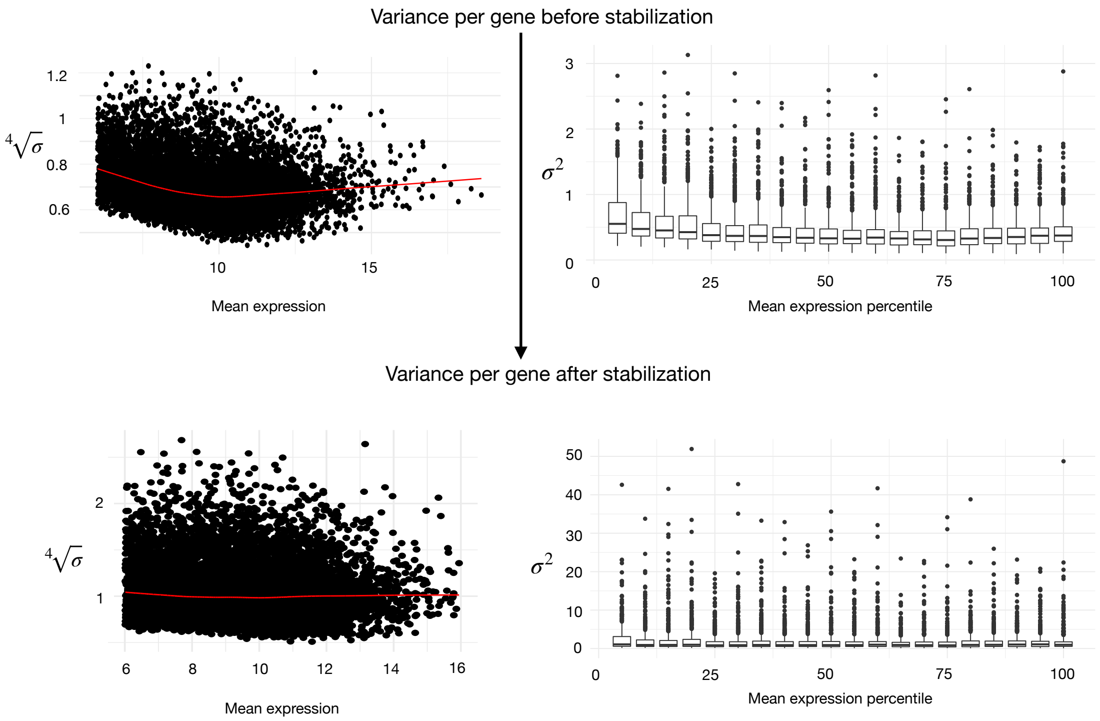


Figura 3: Variance per gene before and after stabilization.

SNPs selection

The second main raw data object used in my project corresponds to the genotype data obtained from the same cohort. The data object consists of a matrix of 6159736 rows and 90 samples, where the entry (i, j) captures the genotype found for the i-th SNP variant the j-th in individual. For the sake of our analysis, the data captures the number of the less frequent allele (0, 1 or 2), for each SNP and individual.

SNPs are subject to one filtering step. To do this, we obtain the MAF (Minor Allele Frequency) value, keeping the SNPs with MAF$>$ 0,05.

3.2. Correlation networks

Partial Pearson correlation and nominal significance tests.

Once the expression data have been preprocessed, the correlation matrix between pairs of genes must be obtained. Having removed batch effects, the variables are not completely independent. If they were, we could make use of the standard correlation between independent expression levels of a gene pair. Instead, in our case we will have to correlate the residuals of each of these variables with respect to the "batch" factor, which is known as a partial correlation test.

In practice, we calculate calculate the Pearson's correlation r for each pair of residue vectors, belonging to each pair of genes, as follows:

$$r_{gs} = cor(s,g) = \frac{\sum(s_i - \bar{s})(g_i - \bar{g})}{\sqrt{\sum(s_i - \bar{s}^2)(g_i - \bar{g})^2}} = \sum s_i g_i =< s, g > \tag{3}$$

Where $<s, g>$ represents the inner product of the two vectors and determine the nominal statistical significance of each (partial) correlation from the distributional assumption that the statistic:

$$t_{n-k-2} = r\sqrt{\frac{n-k-2}{1-r^2}} \tag{4}$$

behaves, under the null hypothesis of no (partial) correlation, as a t-student distribution with $n - k - 2$ degreees of freedom, where n is the degrees of freedom and k is the number of covariates in our batch-effects model.

Here, for performing a regular correlation test of uncorrected gene expression, the significance test should be the same, only considering $k = 0$. Instead, in our case, when taking into account batch effects, we find a difference of $\Delta k = 8$, corresponding to the degrees of freedom invested in fitting the effects of the N=9 batch levels. We compare both distributions in Figure 4 to see what is the effect of considering these nuances when removing batch effects.

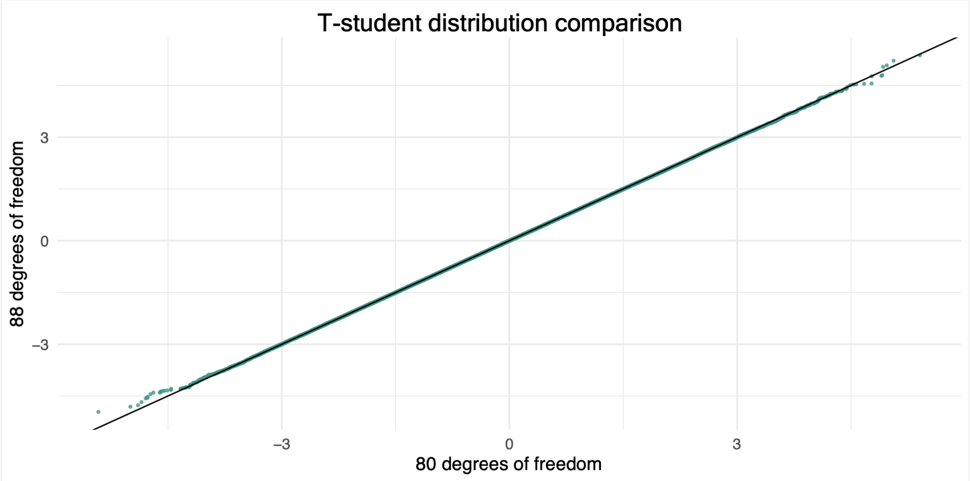


Figura 4: T-student for 80 and 88 degrees of freedom.

The conclusion to be drawn from the Figure above is that the precaution taken of adjusting the degrees of freedom is done since it is the proper approach, even though the data would be practically unaffected if it were not. A small divergence associated with noise at the extremes can be observed, otherwise, both distributions match.

In order to obtain some qualitative insight on the significance levels that we can achieve from some example values for the correlation coefficients, we represent, in Figure 5, the values of the t-statistics associated to values of r of 0,25, 0,5, and 0,75 within the context of a t student distribution with 80 degrees of freedom.

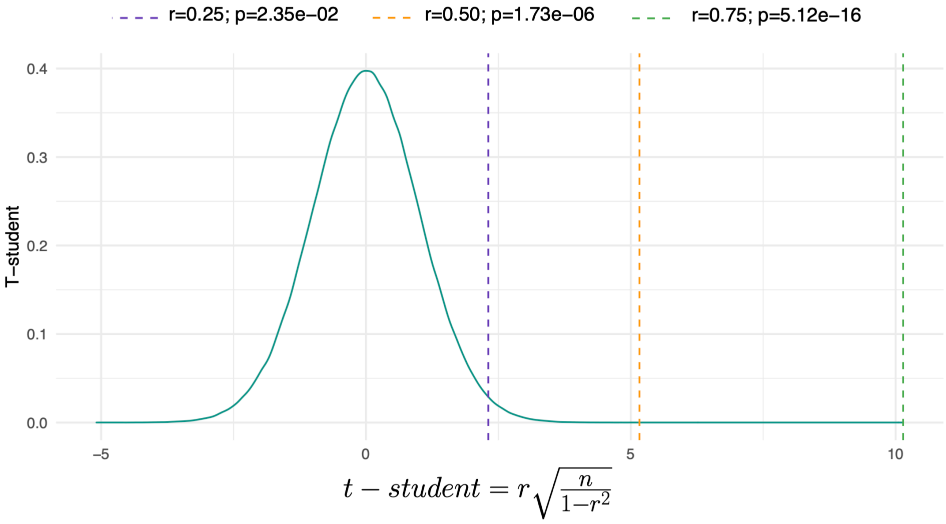


Figura 5: T-student distribution with 80 DoFs.

Multiple testing corrections.

In statistics, when a given hypothesis contrast test is repeated many times, further corrections beyond the estimation of nominal significance scores (i.e. p-values) are needed. By making multiple measurements, the probability of finding at least one rare event increases, and consequently, the expected number of times that the null hypothesis will be rejected erroneously also increases: this is called Type I error.

In this context, a meaningful approach to control for the statistical significance of multiple tests at once is to estimate the so-called False Discovery Rate (FDR), defined as the expected fraction of type I errors among a set of tests all of which have been deemed nominally significant under a given p value threshold.

The approach used in this work for the estimation of FDRs is the so-called Storey and Tibshirani method (16), which starts by obtaining the distribution of p-values of the set of tests analyzed (Figure 6), $f(p)$, and comparing it to the one we would have if the null hypothesis were true, which would be uniform, $f_0(p)$.

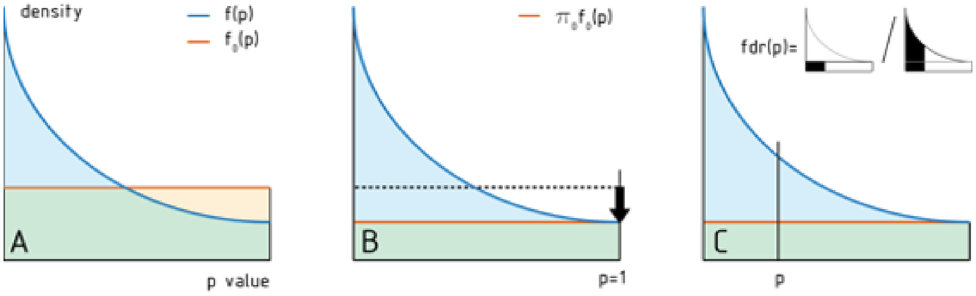


Figura 6: Storey-Tibshirani p-value distribution.

Then, the actual p value distribution is described according to a two component model, as a mixture of a fraction π_0 of tests behaving according to the null hypothesis, and a complementary fraction $(1 - \pi_0)$ distributed according a priori unknown distribution of alternative-hypothesis tests: $f(p) = \pi_0 f_0(p) + (1 - \pi_0)f_A(p)$. Since the null distribution is flat, and the alternative distribution is enriched in low p-values, a numerical procedure can be designed to fit the only parameter of the model, the fraction of true nulls, π_0. Once π_0 is estimated, we can obtain the

FDR associated to a p value p as $fdr(p) = F_0(p)/F(p)$, where $F_o(p)$ and $F(p)$ are the cumulated counterparts of the null $f_o(p)$ and empiric tests distribution $f(p)$, capturing, respectively, the fraction of tests with p value lower than p 1) should the null hypothesis be true ($F_o(p) = p$), or 2) in the observed data ($F(p)$). The procedure for inferring π_0 in order to calculate FDRs according to this method is sketched in Figure 6, while the observed p-value distribution in our co-expression network, containing the p values of the 56026405 gene pairs under analysis, is represented in Figure 7

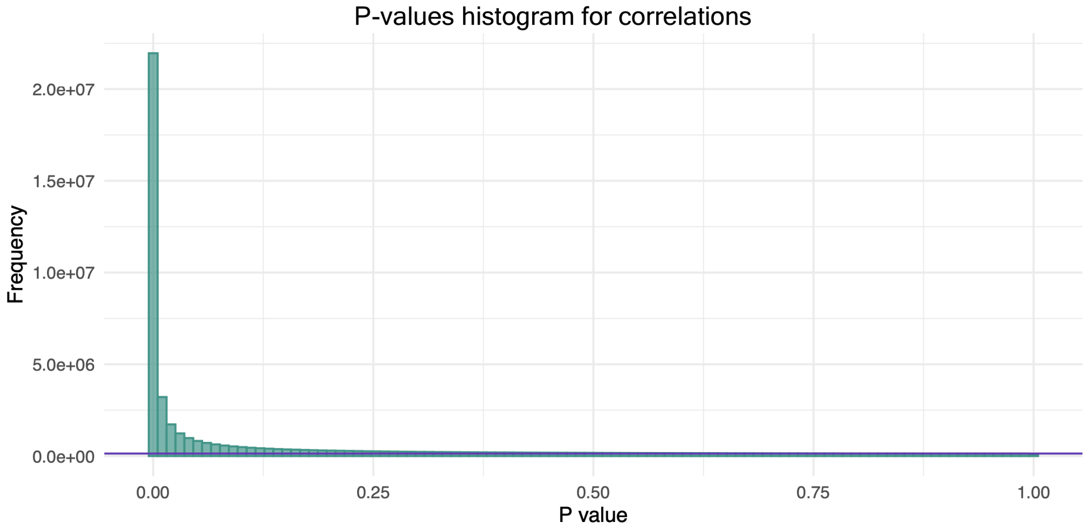


Figura 7: Distribution of p-value of our data.

As seen in Figure 7, we do observe an extreme enrichment in low p-values in the p-values distribution, corresponding to an estimated fraction of null hypothesis of only $\pi_0 = 0{,}27$. This corresponds to a total number of 25897510 gene pairs interactions deemed significantly co-expressed at a FDR of 1 %, more than 46 % of all links tested.

A complementary way to visualize how strongly correlated our variables of interest are is to represent the distribution of average absolute values of the t-statistics across gene pairs, which we see in 8. As we see there, the distribution of the average t-statistic per node has its mode between $t = 3$ and $t = 4$, coherent with the strong and frequently significant correlations we have in our system.

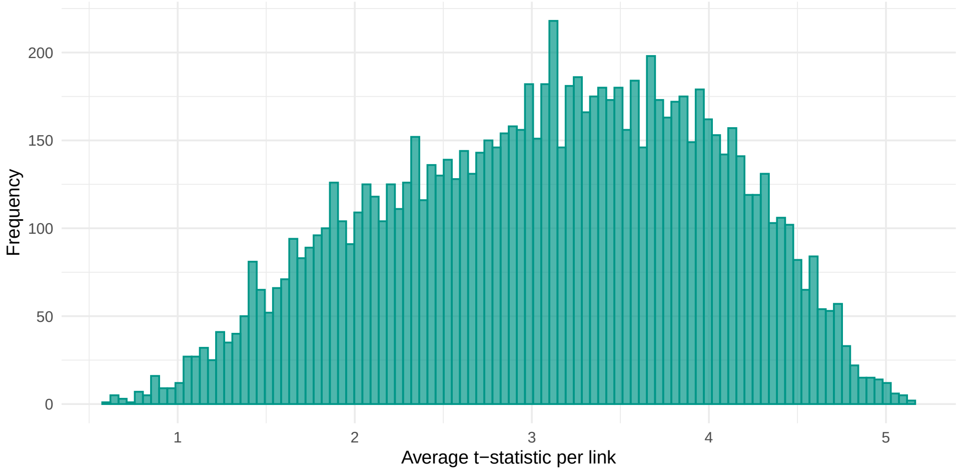


Figura 8: Distribution of gene-wise average absolute value of the t-statistic per gene.

Finally, it is important to note that a key advantage of the ST method is that it also allows a straightforward definition for the False Non Discovery Rate: $Fndr(p) = 1 - \pi_0/\hat{\pi}_0(p)$, with $\hat{\pi}_0(p) = \frac{1-F(p)}{1-F_0(p)}$. While the FDR is defined as the expected fraction of type I errors one makes after calling significant everything under a give p-value threshold, the FnDR is the fraction of type II errors one makes after calling non-significant everything above a given p-value threshold.

3.3. Obtaining eQTLs

The eQTL analysis uses, to identify associations between SNPs and the expression values of their target genes, a simple linear model where the independent variable is the number of less frequent alleles, and the dependent variable is gene expression. The models can contain covariates to account for factors such as sex, population structure or clinical variables. The analysis admits heterocedastic errors (where the variance of these does not remain constant for all observations) and correlated errors to take into account the relationship of the samples. As this is a data analysis that may involve performing 10^9 - 10^{10} statistical tests, using optimized regression algorithms is key. Below is the algorithm for the simple linear regression model, which does not include covariates and assumes homoscedastic, uncorrelated errors, which is the one we will use in this TFG, since the data has been previously transformed to ensure that this easy setup can be used.

We have each SNP coded by 0, 1 and 2 according to the frequency of the minor allele, 0 would correspond to a homozygous genotype of the most common allele in the population, 1 to the heterozygous and 2 to the homozygous of the least common allele. We take the association between gene expression g and genotype s to be linear of the form:

$$g = \alpha + \beta s + \epsilon \tag{5}$$

To perform this analysis, we use the *Matrix eQTL* software library. *MatrixEQTL* is a widely used tool for eQTL mapping that allows an extremely fast characterization of both cis and trans genome-wide eQTLs under different scenarios from RNA-seq and SNP chip data. In spite to the optimized software performance, and due to the extremely large testing space, before using the *MatrixEQTL*, we must filter our co-expression matrix so that we are left with only the significantly correlated gene pairs that will be later considered, a priori, for MR.

The result of the first run of the *matrixEQTL* algorithm conducts a total amount of 5766136 tests, of which, 14566 cis-eQTLs are found to be significant at a FDR threshold of 5 %, involving a total of 815 genes which we will refer to as genetically controlled genes (GCGs), since at least a fraction of its expression variance depend on genetic variation. Importantly, in the estimation of this number, we have already excluded all the instances where 2 or more SNPs in perfect linkage disequilibrium (cases where the genotype of one SNP perfectly predicts the genotype of the other one) are eQTLs of the same gene, for those tests are statistically identical. The number of cis-eQTLs found significant at different FDR threshold levels in each chromosome can be found in Figure 9, while Figure 10 represents the positional distribution vs. the significance of the most significant SNP tested for each gene: in this plot, only the SNPs above the horizontal line marking FDR=0.05 will be considered as cis-eQTLs. As we see in these figures, eQTLs are not homogeneously distributed in the genome, since there are hot spots (which are more relevant in trans- than in cis- eQTLs) accumulating a higher density of SNPs associated to changes in expression of target genes. In order to interpret Figure 9 results, it is important to note that not all chromosomes are the same size (as also seen in Figure 10), and typically the number of eQTLs is proportional to the size of the chromosome. Finally, it is important to note that levels of LD are also variable across the genome. There are areas of the genome that tend to inherit together more often than others and therefore are more likely to have more variables labelled as eQTL, all of them strongly correlated due to LD.

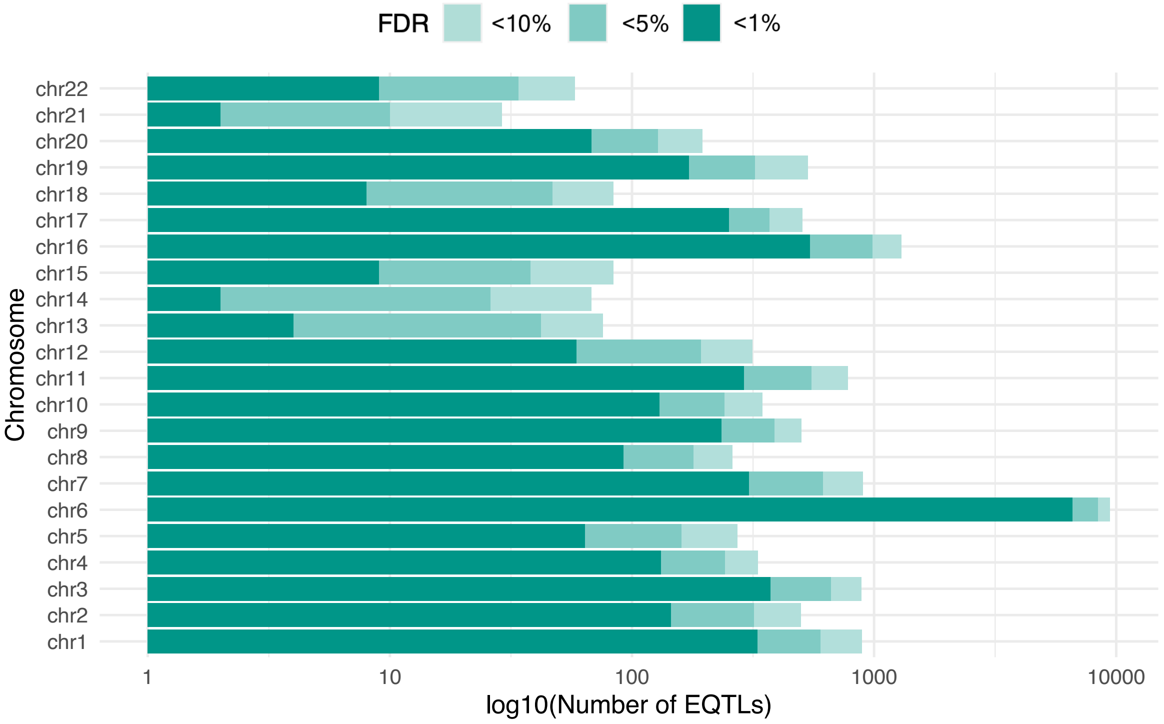


Figura 9: $\log_{10}$(Number of eQTLs) per chromosome under 10, 5 and 1 %.

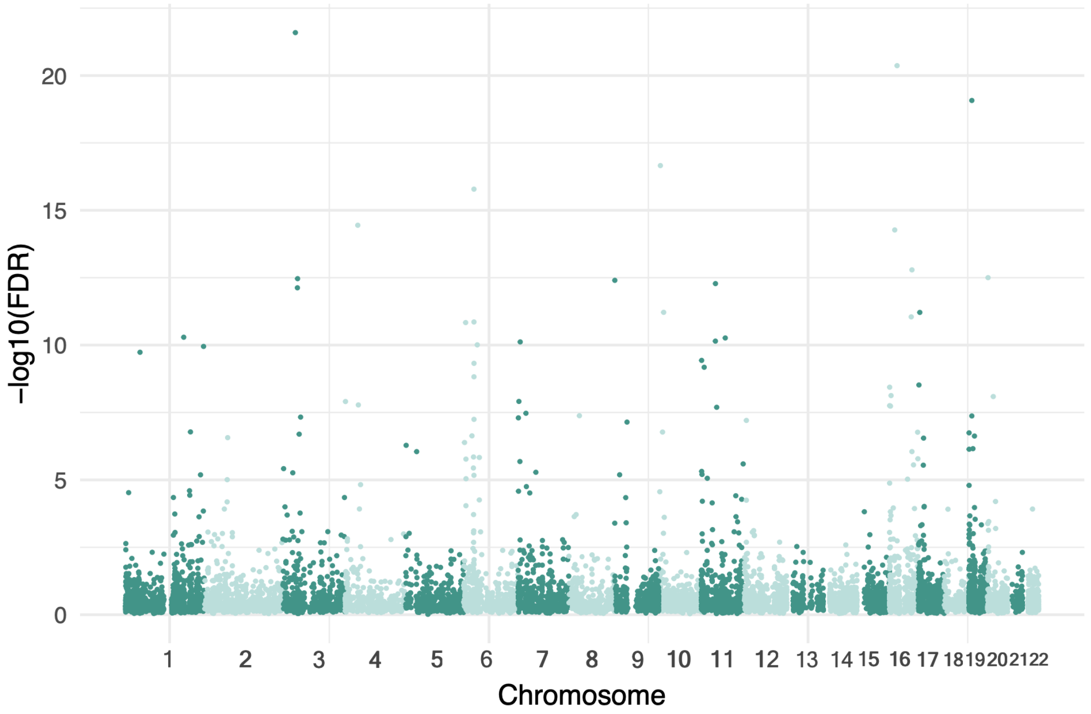


Figura 10: Manhattan plot. -$\log_{10}$(FDR) of the top SNP per gene, as a function of the SNP location, colored by chromosome.

When we focus on those 815 GCGs alone, we find that they are connected through 131644 significant co-expression links. Finally, one additional filtering step is done, whereby we exclude, from the set of significant co-expression links between GCGs, those connecting genes that are separated by less than 1.000.000 base pairs in the genome. After this final step, we stick with a set of 815 GCGs connected through 131230 co-expression links, the co-expression network that we now select for running MR on it, since, for each of the nodes involved in it, we count with at least one relevant instrumental variable (significant cis-eQTL), whilst the distance between the genes are long enough to exclude horizontal pleiotropy links between the SNPs of co-expressed pairs emerging as a consequence of parallel cis effects.

3.4. Mendelian Randomization

To perform the MR analysis we make use of the *MendelianRandomization* (12) library. *MendelianRandomization* is a package developed to carry out various Mendelian randomization analyses on summarized genetic data in R. The package uses various methods to assess whether a risk factor (also called an exposure, in our case, a gene) has a causal effect on an outcome (in our case, another gene). The package uses a special class called *MRInput* within the analyses in order to pass in all necessary information through one simple structure rather than inserting the object in parts. The *MRInput* object has the following components:
- β and its standard error: They are both numeric vectors describing the associations of the genetic variants (eQTLs) with both genes.
- The correlation matrix between genotypes.
- Identification of both genes that will be tested.
- Chosen N SNPs for testing of each gene.

Initially we assume that the scenario is two-sample Mendelian randomization and all the genetic variants considered are uncorrelated (in linkage equilibrium)(13). In this first scenario, for each of K genetic variants (eQTLs) ($k = 1, ..., K$), we represent the estimate of the genetic association with the gene X as X_k with standard error σ_{Xk}, and the estimate of the genetic association with gene Y as Y_k with standard error σ_{Yk}. For the development of this analysis, the inverse variance weighted method will be used. The causal estimate from this method ($\hat{\beta}_{IVW}$) is:

$$\hat{\beta}_{IVW} = \frac{\sum_{k=1}^{K} X_k Y_k \sigma_{Yk}^{-2}}{\sum_{k=1}^{K} X_k^2 \sigma_{Yk}^{-2}} \tag{6}$$

And its standard error:

$$se(\hat{\beta}_{IVW}) = \sqrt{\frac{1}{\sum_{k=1}^{K} X_k^2 \sigma_{Yk}^{-2}}} \tag{7}$$

The inverse-variance weighted estimator can be motivated as a weighted average of the ratio estimates $\frac{Y_k}{X_k}$ for each SNP k, weighted using the reciprocal of an approximate expression for their asymptotic variance $\frac{\sigma_{Yk}^2}{X_k^2}$. The estimate $\hat{\beta}_{IVW}$ expresses the causal increase in the gene Y per unit change in gene X, whose relationship is assumed to be linear, being the sum of a direct (pleiotropic) effect and an indirect (causal) effect.

$$\beta_{Yj} = \alpha_j + \theta\beta_{Xj} \tag{8}$$

being θ the causal effect of the risk factor (gene X) on the outcome (gene Y), which is what we seek.

If, like in our case, the genetic variants are correlated (linkage disequilibrium)(14), the *IVW* estimate can be extended to account for correlated variants by fitting the regression model of 8 using generalized weighted linear regression. Rather than the simple weights $se(\hat{\beta}_{Yj})^{-2}$, we use a weighting matrix Ω^{-1}, where Ω has elements $\Omega_{j1,j2} = se(\hat{\beta}_{Yj1})se(\hat{\beta}_{Yj2})\rho_{j1,j2}$ and $\rho_{j1,j2}$ is the correlation between the $j_1 - th$ and $j_2 - th$ genetic variants (eQTLs).

Considering that, it is important to note that, to run MR, the identification of the valid, significant cis-eQTLs available for each of these genes, in not enough. For example, if we have a gene A significantly co-expressed with another gene B, to run MR on the putative link $A \to B$ we must also test the effect of the SNP$_A$ on the expression levels of B. Since the distance between genes A and B must be, at least, larger than 100 kB (they may in general lie even on different chromosomes), this means that the SNPs involved in the cis-eQTLs should be tested, against the conjugated genes that are co-expressed with their targets.

This is why we re-run the *Matrix eQTL* function to estimate the conjugate effects in trans. We now obtain cis- and trans-conjugated eQTLs depending on whether the SNP is closer to 10^5 bases of the gene or further away. It would not really be necessary to keep the conjugates labelled as -cis by *matrixEQTL*, since previously, a gene-gene vicinity test is performed that removes gene pairs within 10^6 bases of proximity. It is therefore impossible for us to find a SNP that may be labelled simultaneously as cis-eQTL of two co-expressed genes (i.e. being closer than 1KB from the gene body of both genes), since all gene pairs closer than 1Mb are excluded from the analysis to ensure that pleiotropic effects are unimportant.

Therefore, as stated before, at this point we count with 1) 815 genes 2) connected through 131230 co-expression patterns susceptible to be tested for directionality, 3) 14566 significant cis-eQTLs that can be used as instrumental variables. Testing all the SNPs involved in these cis-eQTLs against each of the 815 GCG in trans, we compiled the conjugated tests, completing the set of inputs needed to run MR on all the pairs twice, once for each direction: from gene A to gene B and vice-versa.

The result, for each gene pair, is an effect size and a significance estimate of the effect per direction, which we later correct for multiple testing discovery using the Storey and Tibshirani method. From such analyses we retrieve an FDR (and an FnDR) associated with the statistical plausibility of the existence (or absence of such) of a causal relationship from one gene to the other of the pair, and vice versa.

Therefore, after one of these bi-directional tests, we can classify each pair according to which one is the most plausible causal model, among the following:

- Unidirectional causal relationship: $FDR < 0{,}1$ for one direction and $FnDR < 0{,}1$ for the other. ($\rightarrow$ or $\leftarrow$)
- Mutual causal relationship (or feedback): $FDR < 0{,}1$ & $FnDR < 0{,}1$ in both directions. ($\leftrightarrow$)
- Non-causal relationship (confounding): $FDR > 0{,}1$ & $FnDR > 0{,}1$ in both directions. ($-$)

We must also include the possibility that a studied pair does not comply with any of the 3 cases seen above and therefore its directionality is undetermined by our approach.

It is therefore thanks to the two established FDR and FnDR thresholds that we can infer, with statistical rigor, whether a $<\rightarrow>$, $<\leftrightarrow>$, $<\leftarrow>$, $<->$ relationship is present in a given pair, or if we do not have enough information to ensure causality in one or both directions.

		<←>		
		$FDR < 0,1$		$FnDR < 0,1$
	$FDR < 0,1$	↔	→/?	→
<→>		←/?	?/?	-/?
	$FnDR < 0,1$	←	?/-	-

Tabla 1: Diagram of possible pair relationships

3.5. Down-sampling strategies of the instrumental variables

An aspect that turns to be highly problematic of this approach is the fact that the amount, and relevance of the instrumental variables (eQTLs) available for each gene is not homogeneous (as we saw in Figures 9 and 10). A first strategy that could be chosen would consist of using all the instrumental variables possible whose impact on gene expression is significant enough, that is, all significant cis-eQTL under, let us say 5 % FDR; and its conjugated pairs. This was exactly the strategy followed in ref. (8), which, as mentioned above, led to highly biased results, since it was the relative number of instrumental variables used in cis for each gene in a pair the main predictor of the inferred direction between them.

Since that is obviously a result that is essentially meaningless from a conceptual point of view, our main aim from here on has been that of finding a valid algorithmic strategy to select the cis-eQTLs to be considered in MR in a way that balances the amount of information in the two directions for each link.

The first strategy that we have explored in this TFG is based on selecting only the most significant N SNPs of each gene in each pair, regardless their precise significance levels. Even if this approach allows, in principle, the inclusion of some non-significant eQTLs (that is, marginally relevant instrumental variables), by construction, the fact that we only deal with the set of 815 GCG guarantees that at least one of the SNPs of each gene will in fact be a significant cis-eQTL. According to this strategy, each gene pair will be tested within MR considering the exact same number of N instrumental variables in each direction, eliminating the bias due to considering a different amount of eQTLs in both directions observed in (8).

However, even if that approach guarantees balancing the number of instrumental variables per direction, it does not guarantee that their relevance (i.e. statistical significance) is equally balanced. In the event that this represents a problem, we propose an additional strategy to go further and, in addition to leveling the number of instrumental variables per gene (eQTLs), also balances, as much as possible, their statistical significance.

The latter is a more complex perspective, that requires considering the possibility that different cis-eQTLs of the same gene are chosen as instrumental variables of different pairs in which the gene participates. Therefore, for each test between a gene A and a gene B, each of its N SNPs will depend on the balance made for this pair. While choosing the most significant top N SNPs of each gene is not taking into account the difference in magnitude that may exist with the SNPs of another gene against which they are tested.

The process of balancing the FDR of the N SNPs per gene necessary for each pair is described in the appendix.

3.6. Implementation of empirical null models to quantify directional bias in link prediction

The next step consists of the evaluation of the strategies designed to remove bias in the inference of link directions. Since both methods explored make use of the same number of variants in both directions, we only need to address whether the significance of the cis-eQTLs used in each direction affects our ability to detect directional causality. In other words, we want to address whether the two strategies described above tend to detect links more often going from the gene with the most significant instrumental variables (eQTLs) to the gene with the least significant ones (which we will call forward links) or vice-versa (backward). Under the null hypothesis of no bias, the number of inferred forward and backward links should be approximately the same.

But what does the word approximately mean in the last sentence is of course not trivial. In order to obtain a statistically sound estimate of how compatible our observations of the number of forward and backward links with the hypothesis of no bias, we need to evaluate the percentages obtained that meet our hypotheses based on an empirical null model. An empirical null model is a statistical procedure that generates random versions of the observed statistic as a result of a process where the null hypothesis is satisfied.

For this purpose, we build a stochastic model that generates random instances for the classification of links of different directional types (unidirectional-forward, unidirectional-backward, bi-directional, absent, and undetermined). The algorithm treats both possible link directions identically, and proposes, based on an stochastic procedure, empiric labels depending on whether the said direction is determined with statistical significance, (present link), is discarded with statistical significance (absent link), or it is declared undetermined. Combining the labels in both directions, the link directional profile is determined.

To do that with any of the two possible directions of each link, the parameters used by the model are the following, which will be extracted from the real analyses:

- N number of links to be classified, same as in the real data.
- N_p/N expected fraction of links under FDR=0.1 (link present), extracted from real data estimates.
- N_a/N expected fraction of links under FNDR=0.1 (link absent), extracted from real data estimates.
- π_0 Fraction of null hypothesis among the analyzed links (fraction of links where the causality flux in the direction under analysis does not exist).

In order not to include bias in the algorithm, the parameters N_p/N, N_a/N and π_0 are extracted from the real data, by averaging the corresponding real parameters in the directions forward and backward as follows:

$$N_p/N = (N_p^F + N_p^B)/(2N) \tag{9}$$

$$N_a/N = (N_a^F + N_a^B)/(2N) \tag{10}$$

$$\pi_0 = (\pi_0^F + \pi_0^B)/2 \tag{11}$$

$$\tag{12}$$

where N_p^F, and N_p^B are the number of links deemed significant in the forward and backward direction in the real data; N_a^F, and N_a^B, the number of links discarded, and π_0^F and π_0^B the forward and backward estimated fractions of null hypothesis, respectively.

To develop the algorithm we start from an ideal population of links (gene pairs), a fraction π_0 of which corresponds to gene pairs with no flux in the direction under analysis. From this population we draw N links. This is done by considering a vector of N elements, and assigning a $TRUE$ vs. $FALSE$ with probability $(1 - \pi_0)$ vs. π_0.

These labels capture the *real* status of each gene pair in regards to forward causation, an information that is not known to the analyst. Because of that reason, at this point we need a second step, consisting of assigning a second set of empiric labels describing whether, as a result of the analysis the gene pair will be labeled as $PRESENT$, $ABSENT$, or $UNDEFINED$.

Intuitively, gene pairs that were labeled as $TRUE$ vs. $FALSE$ in the first step will be labeled with $PRESENT$, $ABSENT$, or $UNDEFINED$ empirical labels at the second step with different probabilities. More specifically, we assign the following probabilities:

- p_1: probability for $TRUE$ links to be empirically labelled as $PRESENT$.
- p_2: probability for $TRUE$ links to be empirically (mis)labelled as $ABSENT$.
- $1 - p_1 - p_2$: probability for $TRUE$ links to be empirically labelled as $UNDEFINED$.
- p_3: probability for $FALSE$ links to be empirically (mis)labelled as $PRESENT$.
- p_4: probability for $FALSE$ links to be empirically labelled as $ABSENT$.
- $1 - p_3 - p_4$: probability for $FALSE$ links to be empirically labelled as $UNDEFINED$.

In order to infer the values of these probabilities, we can rebuild the expected fractions of gene pairs labelled as having, or lacking a forward direction in the link:

$$p_1 \times \pi_0 + p_4 \times (1 - \pi_0) = N_p/N \tag{13}$$

$$p_2 \times \pi_0 + p_3 \times (1 - \pi_0) = N_a/N \tag{14}$$

Additionally, we can express the fractions of false discoveries, and false non-discoveries, to obtain two further constraints:

$$\frac{p_4 \times (1 - \pi_0)}{p_1 \times \pi_0 + p_4 \times (1 - \pi_0)} = FDR = 0{,}1 \tag{15}$$

$$\frac{p_2 \times \pi_0}{p_2 \times \pi_0 + p_3 \times (1 - \pi_0)} = FnDR = 0{,}1 \tag{16}$$

Thanks to these equations we can obtain the values of the probabilities p_i (with $i = 1, 2, 3, 4$) as follows:

	Top N SNPs	Balanced N SNPs
p1	0.068	0.091
p2	0.014	0.019
p3	0.497	0.199
p4	0.003	0.012

Tabla 2: Probabilities obtained from equations 20 to 23.

And, from these, complete our computational model assigning an empirical label to each of the drawn N links stochastically.

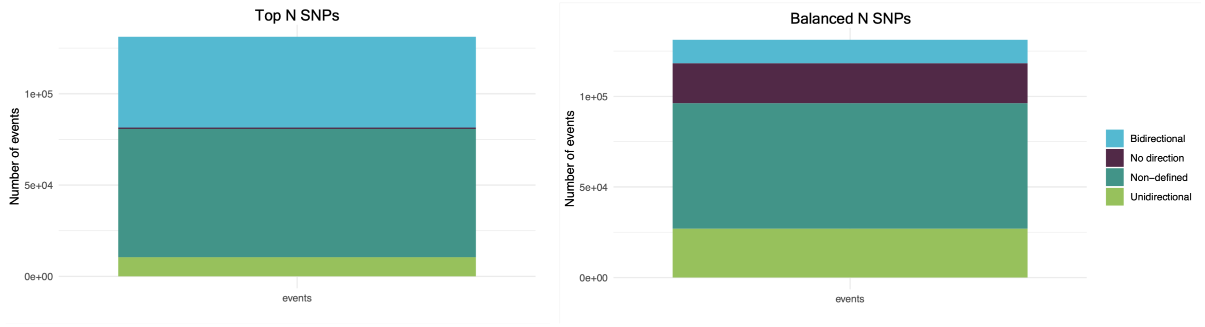
Repeating the procedure in an independent fashion in the conjugated direction direction, using the same parameters, we can combine the empirical labels of the forward and backward directions to reproduce the classification of the fraction of links of each directional profile: forward, backward, bidirectional, absent, or undefined. Repeating the procedure, for both directions a total number of times $I = 100{,}000$, we can obtain an empiric distribution of the relative fraction of forward vs. the sum of forward plus backward links.

Through this procedure, we capture, empirically, the stochastic dispersion that may be expected for this variable, and, since the procedure features absolutely zero bias by construction, it serves as an appropriate null model to evaluate whether our findings in the real data may be considered compatible with the null hypothesis of no bias, or not.

4. Results

4.1. Link direction inference under different genetic variable selection criteria

Once the strategies used to select the instrumental variables (Top-SNPs, and Balanced-SNPs methods) for MR were introduced, and the concepts of forward, and backward link directions (from gene with the more significant instrumental cis-eQTLs, to the gene with the less significant one; or vice-versa), we can classify the gene-pairs in our system according to each strategy. Figure 11 shows the results of both classification attempts done in our study.



(a) Top SNPs

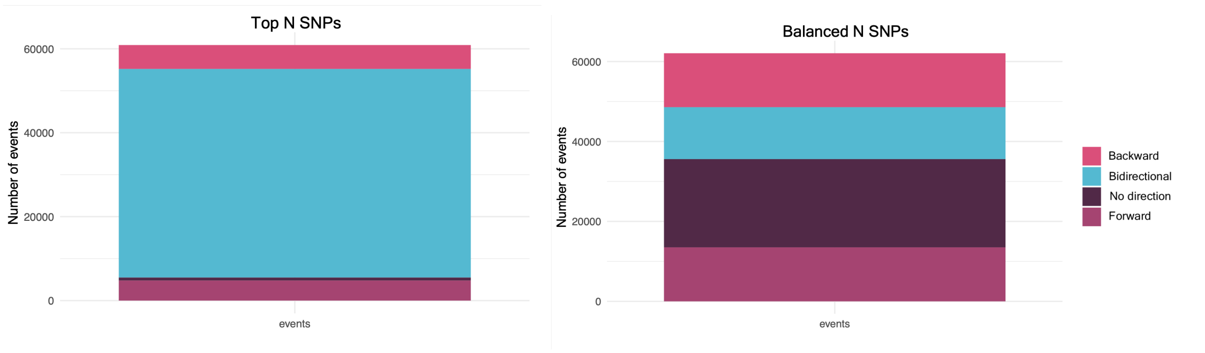
(b) Balanced SNPs

Figura 11: Fraction of tested link results.

It is interesting to note the discrepancies between the two results, since the same number of instrumental variables per pair and per gene are being analyzed, but the differences in discerning directionality are clear.

While, in principle, the balanced method should provide a better control of the directionality bias, the price to pay for this improved symmetry is the inclusion of potentially less relevant instrumental variables in the analyses than the Top-SNPs method. This is why we have more cases of non-directionality in the results of the Balanced-SNPs strategy than in Top-SNPs, since the FDRs of the SNPs used in Top-SNPs are more extreme and therefore more often lead to significant results in either direction. Similarly, we found more bidirectional links in the Top-SNPs strategy than in the balanced, while links without causal relations (i.e. confounded gene pairs) are more often identified after using the balanced strategy.

Since most of the links are undefined, in the following Figure 12 we get rid of them in order to see more clearly the proportions between defined directions. Furthermore, we divide the unidirectional links defined as forward and backward into two subsets. In this analysis, the evaluation of the gradient of FDRs to discern between a forward or backward directionality linkage is given by the comparison of the SNP of highest significance for each gene. Another strategy can be proposed by comparing the logarithmic sum of the N FDRs (one for each SNP) of each gene. Since the results are very similar to those reported in the initial methodology of using the most significant FDR, it has not been considered pertinent to present the figures of the results of this new proposal which, however, are included in the appendix to go into more detail.

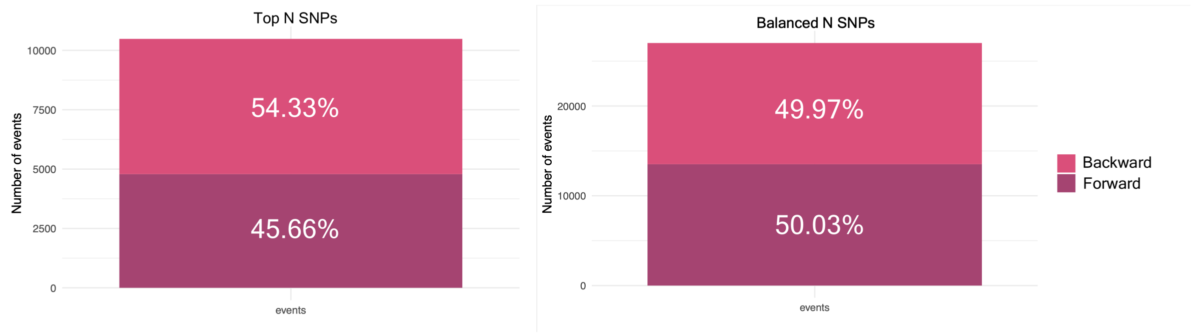


(a) Top SNPs

(b) Balanced SNPs

Figura 12: Fraction of links of defined direction results using the most significant of each N SNP to discern between forward and backward directionality.

Next, we evaluate individually the relative frequencies of forward and backward links, as represented in Figure 13. As we see in the figure, the Top-SNPs strategy presents a significant departure of more than 4 % of the links, from the expected equilibrium between forward and backward links; while the Balanced-SNPs strategy produces, as expected, a much more reduced bias of 0,03 %. (13515 links forward vs. 13505 backward).



(a) Top SNPs

(b) Balanced SNPs

Figura 13: Fraction of links of defined forward/backward using the most significant of each N SNP to discern directionality.

4.2. Statistical bias in the directionality assessment

Regarding the results found in the model proposed in section 3.6, the following distribution of percentages found as forward/backward.

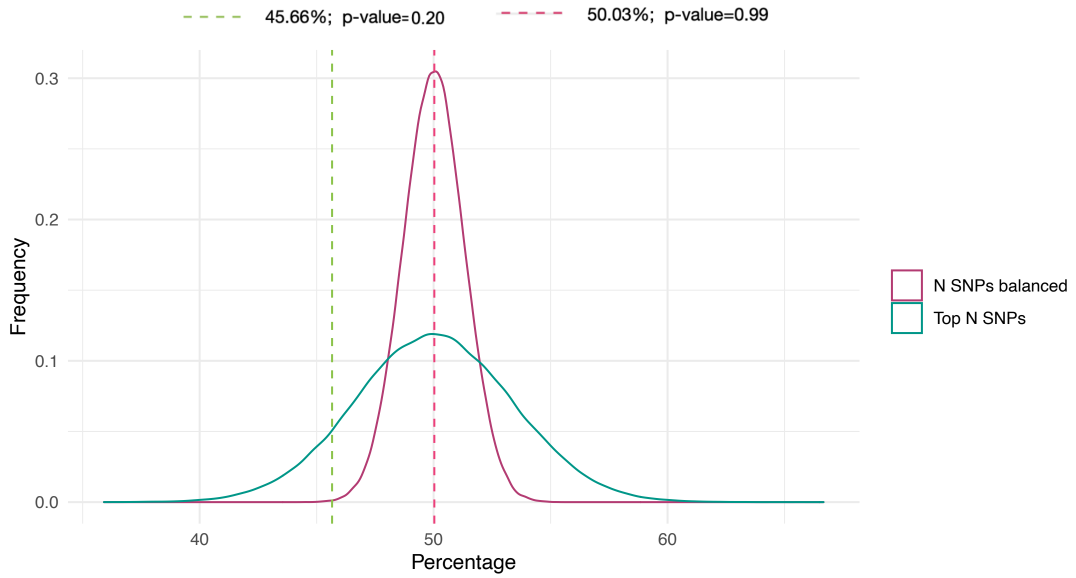


Figura 14: Percentages distribution of forward/backward founded links .

The dashed lines show the values of the percentages that have been found experimentally for Top-SNP and Balanced-SNP method together with their p-value. It is clearly noticeable that the variance of the curve for Balanced-SNP is much narrower than for Top-SNP, a fact that emphasizes again that the bias found in the results of the first method is notoriously lower and its experimental result is closer to the expected value.

5. Discussion

Upon completion of this report I can assure that I have gained significant practical expertise and theoretical knowledge of a number of key aspects in computational genomics. Especially on transcriptomics pre-treatment and analysis using RNA-seq, how basic genotyping techniques work, what are the main concepts around eQTLs mapping, co-expression networks and Mendelian Randomization in epidemiology. In addition, I have acquired programming skills in a language that was previously unknown to me, *R*, making use of it for the development of my research.

As a result, we have proven that our method is feasible, and that, with less than one hundred individuals, it allows the estimation of significant causal relationships in thousands of co-expression pairs. The problem of asymmetry in the results caused by the use of different number of instrumental variables per gene in the analysis of pairs through RM has been solved. For this purpose, two heuristic strategies have been developed. The first one is based on equalizing the number of instrumental variables per gene and the second one is based on (approximately) equalizing the significance of these variables. In both ways the bias is reduced, especially in the algorithm based on balanced variables, which is more conservative in assessing links directionality. This translates into a lesser number of links with at least one direction labeled as significant, a higher number of genes for which the existence of causality fluxes are discarded, and a lower number of bi-directional links. This provides an outlook that aligns better to what is typically found in gene regulatory networks, which are sparse systems where bi-directional interactions are scarce (17). Furthermore, the balanced algorithm provides a value for the fraction of links that align with, or against, the eQTL significance gradient (forward vs. backward directions) that is broadly compatible with the hypothesis of no-bias, as ascertained after the implementation of an empiric null model for this statistic.

In summary, the proposed methodology is promising, but it must be refined to become an effective tool for the inference of causal regulatory networks since the results obtained here are still preliminary and the development up to now has limitations.

The first imperfection found in this model is that the control of the RM assumptions in each link is not strict. Instead, certain filtering steps have been proposed on the selected SNPs and gene pairs under analysis to enrich the cases where they are verified. The selection of cis- or trans- eQTLs by means of the distance to the gene on which they act is introduced to reduce the number of cases where it is easier for the MR conditions to be violated, but this does not ensure categorically that all the analyzed links fulfill these conditions. This is why the most "meaningful" interpretation of the results in systemic, i.e., a global analysis in the directed network.

As a next step, we propose to explore the effect of the N-value reflected in the results. To this end, we propose to develop a strategy that allows N to be specific for each link, since the distribution of instrumental variables per gene is heterogeneous and the restriction to N variables could work against us. There are cases of gene pairs that show a higher number of significant eQTLs, part of which are discarded by limiting the study to N variables. At the same time, we can find the opposite situation, in which N eQTLs are studied even though they are not of relevant significance, and the number of study variables can be reduced to below N.

Another proposal is to study the results obtained in relation to external transcriptomic regulatory interaction data. As a general rule, it should be found that the nodes from which more outgoing direction links start function as transcription factors, since they are the ones in charge of gene regulation. A development tool for this proposal is provided by HOMER: a computational tool for the analysis of TFBS (Transcription Factor Binding Sites) enrichments. By providing this program with the list of genes that have been found to receive incoming links, HOMER presents information on their relationship with different transcription factors that can be compared with those found in our network as outgoing links to the genes studied. In this way it can be questioned whether the information in that database is consistent with the network structure found in our algorithm.

By integrating all these improvements we hope to achieve an effective characterization of causality in co-expression networks, at the genomic level, which will be useful to shed light on the role of individual genes and gene-modules on the emergence of global transcriptomic profiles significantly associated to assorted phenotypes of interest, such as the distinction between health and disease for a number of pathologies.

Referencias

[1] Banf, M. & Rhee, S. Y. (2007) "*Computational inference of gene regulatory networks: approaches, limitations and opportunities*". Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1860(1), 41-52.

[2] Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E & Guthke, R. (2009) "*Gene regulatory network inference: data integration in dynamic models*". Biosystems, 96(1), 86-103.

[3] Consulting about genotyping chips for SNPs sequencing in https://www.illumina.com/science/technology/microarray.html and through the video https://www.youtube.com/watch?v=lVG04dAAyvY

[4] Consulting about Mendelian Randomization in August 2022 from https://academic.oup.com/hmg/article/23/R1/R89/2900899?login=false

[5] Consulting about Mendelian Randomization in August 2022 from https://es.slideshare.net/jamesmcm03/mendelian-randomisation

[6] Consulting about Mendelian Randomization in August 2022 from https://www.bmj.com/content/362/bmj.k601

[7] ZHANG, Ruoxin, et al. "*A blood pressure-associated variant of the SLC39A8 gene influences cellular cadmium accumulation and toxicity. Human molecular genetics*", 2016, vol. 25, no 18, p. 4117-4126

[8] Regina Santesteban Azanza (2020) "*Métodos computacionales para la caracterización de relaciones causales entre genotipo y fenotipo: Heredabilidad de la expresión y coexpresión genética*"

[9] Lindsay I Smith. "*A tutorial on Principal Components Analysis*" February 26, 2002.

[10] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., & Irizarry, R. A. (2010) "*Tackling the widespread and critical impact of batch effects in high- throughput data.*" Nature Reviews Genetics, 11(10), 733-739.

[11] durbin, B., Hardin, J., Hawkins, D. & Rocke, D. (2002) "*A variance-stabilizing transformation for gene-expression microarray data.*" Bioinformatics, 18(Suppl 1), S105-S110

[12] Burgess, S. Yavorska, O. "*MendelianRandomization v0.3.0: an R package for performing Mendelian randomization analyses using summarized data*"

[13] Burgess, S., Scott, R. A., Timpson, N. J., Smith, G. D. Thompson, S. G. "*Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors*" (2015)

[14] Burgess, S. Thompson, S., G. "*Interpreting findings from Mendelian randomization using the MR-Egger method*" (2017)

[15] Benjamini, Y., Hochberg, Y (1995) "*Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*". Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300

[16] Storey JD, Tibshirani R (2003) "*Statistical significance for genomewide studies*". PNAS 100:9440–9445

[17] Sanz, J., Cozzo, E., Borge-Holthoefer, J. Moreno, Y. (2012). "*Topological effects of data incompleteness of gene regulatory networks. BMC systems biology*"., 6(1), 1-10