# Genetic Architecture of genes coding for RNA-binding proteins

***Fernando Andrade***

# Genetic Architecture of genes coding for RNA-binding proteins

**Fernando Andrade**

***Orientador:*** **Prof. Dr. André Fujita**

Doctoral dissertation submitted to the *Institute of Mathematics and Statistics* – IME-USP, in partial fulfillment of the requirements for the degree of the Doctorate Bioinformatics Graduate Program. *FINAL VERSION*.

**USP – São Paulo**
**October 2017**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Andrade, Fernando
Ag Genetic Architecture of genes coding for RNA-binding proteins / Fernando Andrade; orientador André Fujita. - São Paulo - SP, 2017.
69 p.

Doctoral dissertation (Doctorate Candidate - Programa de Pós-Graduação em Bioinformática) - Instituto Matemática e Estatística, Universidade de São Paulo, 2017.

1. QTL. 2. Pleiotropy. 3. RBP. 4. Gene expression. I. Fujita, André, orient. II. Título.

# ACKNOWLEDGEMENTS

Agradeço à CAPES e à FAPESP pelo financiamento desse projeto. Agradeço também a todos aqueles que ajudaram durante esse período do doutorado com críticas, conversas e sugestões valiosas.

# ABSTRACT

ANDRADE, F.. **Genetic Architecture of genes coding for RNA-binding proteins**. 2017. 69 f. Doctoral dissertation (Doctorate Candidate em Bioinformatics Graduate Program) – Instituto Matemática e Estatística (IME/USP), São Paulo – SP.

RNA binding protein (RBP) is a class of proteins closely related to gene expression, directing RNA splicing, maturation, transportation, and degradation. Despite its importance in post transcriptional regulation, little is known about the genetic basis of the regulation of these genes. In this work we use quantitative genetic approaches to identify which genes or regulatory elements are associated with variation in the expression of RBP-coding genes. We show that the regulation of these genes is highly modular, and the expression of subgroups of RBP-coding genes share QTLs among themselves but not with other subgroups.

**Key-words:** QTL, Pleiotropy, RBP, Gene expression.

# LIST OF FIGURES

# LIST OF TABLES

# CONTENTS

CHAPTER

# 1

# INTRODUCTION

The study of how variations in genotype affect the traits has long been one of the main aims of genetics. In agricultural sciences, understanding this relation enables us to devise better breeding programs, leading to higher yielding crops. In medicine this knowledge can help the identification of disease markers, which could potentially lead to better treatments.

The first step in the relation between genotype and phenotype is gene expression. Anything that changes what is being expressed (which allele, for example), how much of this allele is being expressed and where this allele it is being expressed might have downstream effects, possibly changing the final phenotype. This makes the gene regulation one of the key links between genotypic and phenotypic changes.

RNA binding protein (RBP) is a class of molecules that plays a central role in gene regulation. Alterations in the activity of some of these genes (PRECISA VER NO TEXTO TODO. VC FALA THESE GENES MAS NAO FICA CLARO SE SAO AS RBP OU OS TARGETS DELAS.) have been linked to different diseases. Although these genes are central to the connection between genotype and phenotype, to the best of our knowledge, there is no genome-wide scale study about the regulation of RBP-coding genes. Thus, our study aims at better understanding the regulation of this class of genes, by describing potential biological factors underlying the regulation of expression of individual genes as well as the expression of subgroups of these genes.

In this section we will further present and discuss RBPs (DISCUTIR SOBRE RBP NA INTRO SOA ESTRANHO.) as well as the methods we will use PARA QUE???. In section 1.1 we describe the focus of our research, the RBP-coding genes. The methods used to describe how the genotype is related to the expression of RBP-coding genes is discussed in sections 1.2 and 1.3. Section 1.4 presents an overview of QTL studies in gene expression as well as the structure and degree of pleiotropy in gene expression regulation. All these sections are brought together in section 1.5, where we delineate the aims of this work. RE-ESCREVA

TODO ESSE PARAGRAFO. FALE ASSIM: NA SECAO BLABLA FALAREMOS SOBRE BLABLABLA. NA SECAO BLABLA APRESENTAREMOS BLABLA. NA SECAO BLABLA DISCUTIREMOS BLABLA. E ASSIM POR DIANTE. TEM QUE SER SISTEMATICO.

## 1.1 RNA-binding proteins

RNA-binding proteins (RBP) are a group of proteins that share a RNA-binding domain and are essential to RNA metabolism. RBPs may present more than one binding domain per molecule and, in some cases, these can be even different from each other (NAO ENTENDI. O QUE PODE SER DIFF?). More than 600 different domains are distributed among 1,542 RBPs(GERSTBERGER; HAFNER; TUSCHL, 2014), and the five most frequent ones are the RNA recognition motif (RRM), the K homology (KH) domain, the DEAD box, the double-stranded RNA-binding motif (DSRM), and the zinc finger domain (GERSTBERGER; HAFNER; TUSCHL, 2014).

RBPs participate in all processes involving RNA, such as alternative splicing, maturation, edition, polyadenilation, and transportation and degradation (GLISOVIC *et al.*, 2008), making these proteins central (VC GOSTA DESSE CENTRAL NE? KKKK TROQUE POR SINONIMOS. EU VEJO UM CENTRAL POR PARAGRAFO!) to the post transcriptional regulation of gene expression. RBPs work by interacting with coding or non-coding RNAs through untranslated sequence elements for regulation, which are short sequences in the RNA that determine which RBP can be bound, therefore determining which RNP can be formed (KEENE, 2007). In addition to RNAs, RBPs can also bind other proteins, forming ribonucleoprotein (RNP) complexes. One example of RNP is the exon-junction complex which interacts with newly spliced exons, binding them together to form a mature mRNA (GLISOVIC *et al.*, 2008).

The interaction between RBPs and mRNAs, forming the RNPs, is a process that depends on the function of the protein encoded by the gene (KEENE, 2007; HASAN *et al.*, 2014). In yeast, for example, each gene of the PUF family of RBPs interacts with groups of functionally related mRNAs (GERBER; HERSCHLAG; BROWN, 2004). In neuronal tissues, of all interactions of the RBP NOVA, 75% bind RNAs that code for synapse proteins (ULE *et al.*, 2003). The fact that RNPs are formed among mRNAs with similar functions emphasizes one more relevant role of RBPs, the coordination of translation.

RNPs are formed right after the nascent mRNA begins to emerge, protecting the RNA (MANIATIS; REED, 2002). Concomitantly, RBPs related to splicing start binding to exons and introns, preparing the mRNA for the splicing events (MOORE *et al.*, 2006). After splicing, the exon junction complex (EJC) also binds the spliced exons to form the mature mRNA (GLISOVIC *et al.*, 2008). With the EJC also bound to the RNP, the complex is then transported to the cytoplasm through the nucleolar pore complex. In the cytoplasm the RNP can then contact ribosomal subunits ant proceed to translation (MOORE, 2005). But transcription may not proceed

immediately, some RNPs may have their translation held until proper time and others have their mRNA targeted for degradation (MOORE, 2005). Throughout all these steps the RNP is being remodeled, changing some RBPs that are part of the complex (MOORE, 2005).

One final point to show the relevance of RBPs is that mutations in these genes can lead to severe diseases and other impairments, for example, mutations in FMR1 lead to hypermetilation and therefore silencing the gene, leading to the fragile X syndrome (CHELLY; MANDEL, 2001; LUKONG *et al.*, 2008). A host of neuromuscular disorders is also associated with mutations in RBP-coding genes, for example, spinal muscular atrophy, myotonic distrophy and oculopharyngeal muscular dystrophy (LUKONG *et al.*, 2008). Mutations, chromosomal alterations and alterations in gene expression of RBP-coding genes are also associated with different types of cancer (LUKONG *et al.*, 2008). A compendium of diseases associated with alteration in RBP activity can be found in (COOPER; WAN; DREYFUSS, 2009).

Given the relevance of these genes and the lack of knowledge about their regulation on a genome-wide scale, we sought to identify the elements that affect the expression of this group of genes.

## 1.2 Quantitative genetics

Quantitative genetics is a field of inquiry that focuses on the relation between the phenotype, the genotype and the environment. It has originated from the reconciliation of Mendelian genetics and the study of continuously varying traits (biometry), based on the idea that these traits are underlined by numerous *loci* following the Mendelian law of segregation, but each possessing infinitesimal effect on the trait. The first proponents of these ideas were Fisher (FISHER, 1919), Wright (WRIGHT, 1921) and Haldane (HALDANE, 1932), who also developed statistical methods (analysis of variance and path analysis) in order to split the phenotypic variance into different components, such as variation due to genetic and environmental factors.

The idea resulting from these first models is that some of the phenotypic variation ($V_P$) is related to variation in the additive effects of genotype ($V_A$), which are effects that are intrinsic to each *locus*, independent of other factors. This relation is called narrow sense heritability ($h^2$, see section 3.4.1) and is defined as the proportion of the phenotypic variance explained by the variance of additive effects:

$$h^2 = \frac{V_P}{V_A} \tag{1.1}$$

The concept of heritability then implies that it is possible to estimate the extent to which phenotypes are amenable to selection, artificial selection in breeding programs and natural selection in wild populations. Based on this idea, researches could then to some extent predict the response of a population to a selection regime. This advancement allowed further development of evolutionary biology and the origin of breeding programs. The relation between directional

selection and the heritability of a trait can be seen in the breeder's equation:

$$\Delta z = h^2 s \tag{1.2}$$

, where $\Delta z$ is the difference in the phenotype mean across generation and $s$ is the selection coefficient.

Despite of success in breeding programs (see for example (DUDLEY; LAMBERT, 2010)), the use of heritability in other areas is more restricted. In order to estimate heritability some assumptions are made, like random mating, Hardy Weinberg equilibrium (HWE) and linkage equilibrium, and these assumptions do not always hold in natural populations. HWE assumption is specially relevant when studying the heritability of a trait in two different populations, or two different generations of the same population. In these cases, the frequency of the alleles can be different among the situations so the heritability of the trait might not be comparable. Another limitation of heritability studies, specially in wild populations, is the effect of the interaction between the genotype and the environment, which is generally neglected. So, with all these caveats we should take care about interpretations of heritability (KEMPTHORNE, 1978; KEMPTHORNE, 1997).

Another useful concept is genetic correlation (see section 3.4.1), defined as the correlation of additive effects between two phenotypes. This way, if two traits have appreciable genetic correlation it would imply that they have a shared genetic basis, either due to shared, or linked genes. The opposite is not the case however, traits without genetic correlation may nevertheless share genetic basis, because the additive effects of different shared genes may have different signs and cancel out (HOULE, 1991; GROMKO, 1995)).

These two concepts, heritability of traits and genetic correlation between traits, are useful to characterize the relation between genotype and phenotype in a population, but they treat the organisms as "black boxes". Little information about the genetic basis of a trait can be inferred by using either heritability or genetic correlation, so questions regarding which genes are underlying a given trait or how are the *loci* effect sizes distributed cannot be addressed.

Quantitative trait *loci* (QTL) mapping is a methodology that is related to heritability and genetic correlation, since it also studies the relation among phenotypic, genotypic and environmental variances. Aside from being related, QTLs are also complementary to the concepts put forward by Fisher and Wright, since it aims to map which *loci* are contributing to phenotypic variance. QTL has its roots in the same period as heritability (SAX, 1923), but until the development of molecular biology its application was very limited since it depends on knowing the genotype of at least some positions in the genome.

The idea that underlies the QTL mapping is that unknown biological effects on traits can be mapped if they are linked to known variants (markers) in the genome. For example, consider that the trait $y$ is affected by allelic changes in the unknown gene $g$, which is closely linked to the marker $m$. If we test the association of all markers and the variation of $y$, we can then observe

that it is associated with $m$, indicating the region on the genome where the true biological effect $g$ lies. The precision of QTL mapping, then, depends on many factors such as the linkage between the biological effect and the marker, the genome coverage (the number of markers and their positions) as well as the allele frequencies on the population being studied.

The association between the QTL and the phenotype can be tested, in its simplest case where there is only additive effects and one bi-allelic marker, by fitting a linear model. In this linear model, the phenotype $y$ of each individual $i$ is related to the number of marker's minor allele $g$ (0, 1 or 2 for each individual) plus a random normal error term $e$:

$$y_i = \mu + \beta_i g_i + e_i \tag{1.3}$$

where $\beta$ is the additive effect on $y$ of changing one allele on $g$. Also from this linear model we can obtain the coefficient of determination ($R^2$), that estimates the proportion of phenotypic variance explained by the additive effect of each *locus*.

Taking a close look at equations (1.1) and (1.3) we can see how heritability and QTL relate to and complement each other. Using heritability we are able to estimate, for a given population in a given condition, what is the proportion of the variance of a phenotype that can be accounted for by variance in additive effects, which, in turn, can be studied using QTL mapping approaches. Although the effects of QTLs would have to sum up to the estimated heritability this is not the case in actual conditions, in most cases the QTLs explain less that what is expected given the heritability of a trait, a feature that is called "missing heritability". This missing heritability might be explained by the effect of interactions between genes, which are generally not tested and also by QTLs of small effect sizes, which require greater sample sizes.

Developments in molecular biology not only allowed widespread usage of QTL mapping, they also opened new ways of estimating heritability and genetic correlation. Heritability is estimated from the degree of resemblance between relatives, so early works used families or pedigrees for estimation. The development of high density SNP arrays and high throughput sequencing techniques allowed the estimation of the genotypic similarity across individuals even without known pedigrees (HAYES; VISSCHER; GODDARD, 2009; YANG *et al.*, 2010),allowing the estimation of heritability in natural populations.

## 1.3 Pleiotropy

Pleiotropy is a central concept in biology that dates back to the beginning of the 20th century (reviewed by (STEARNS, 2010)). It is defined as one gene, or one mutation, affecting multiple traits, and is a relevant concept when studying the genetic basis of multiple traits.

Studies of pleiotropy encounter technical and conceptual difficulties (PAABY; ROCKMAN, 2013; WAGNER; ZHANG, 2011). One of the relevant conceptual difficulties relevant to this work is that different biological mechanisms can cause pleiotropy. For example, let

us consider two cases, the first one where an enzyme performs a single function in different situations and the second one where another enzyme performs two distinct functions in the same situation. One can argue whether these two enzymes are pleiotropic but in both cases mutations in these enzymes will affect multiple traits and will be classified as pleiotropic. For this reason, in this work we consider pleiotropy as a property of polymorphisms, not genes.

Technical difficulties with pleiotropy involve, mainly, indirect pleiotropy and linkage. Indirect pleiotropy is the situation where variation in one gene affects one trait, which, in turn, affects a second trait and is most relevant when dealing with characters in different developmental stages. In this situation, the gene is not affecting both traits, it affects only the first one but, depending on how we assess the effects, we might get an association of the gene with both traits. The situation described here can be mostly resolved by ways of statistical modeling (STEPHENS, 2013).

The second methodological difficulty when dealing with pleiotropy is linkage, a situation where two different polymorphic *loci*, that affect one trait each, are both segregate with the same marker. In this situation, depending on the number of markers included in the analysis, it is possible that both traits will be associated to the same marker, which will be considered pleiotropic, contrary to the true effects of polymorphisms. The issue of linkage can only be mitigated by increasing the number of markers used to dissect the *loci* and also increasing the sample size.

Despite the conceptual and technical challenges, pleiotropy has been studied in different areas, in evolutionary biology, for example, it plays a central role in defining constraints on possible evolutionary outcomes (MITCHELL-OLDS, 1996). In health sciences pleiotropy can help establish the connection among different diseases (SIVAKUMARAN *et al.*, 2011), explain syndromes (VALENTE *et al.*, 2006) and also senescence (WILLIAMS, 1957).

In gene expression pleiotropy is ubiquitous. In an experiment using yeast, more than 95% of single gene mutations lead to significant changes in at least one gene beside the focal one (HUGHES *et al.*, 2000). QTL mapping studies also found substantive frequency of pleiotropic effects. For example, in yeast some QTLs were found to affect the expression of up to 94 genes (BREM *et al.*, 2002). In mice, a study found seven key regulators that were associated with variation in expression of more than 1.500 genes (CHESLER *et al.*, 2005). In humans, pleiotropic *loci* were also mapped, although to a lesser degree, with *loci* affecting up to 31 phenotypes (out of 984) (MORLEY *et al.*, 2004). The RNA operons can also be a source of pleiotropy in gene expression since RNPs are formed coupling mRNAs that have similar functions together (KEENE, 2007), this way, mutations in RBPs that are part of one RNP might impact the expression of all the related mRNAs.

One caveat of QTL mapping in gene expression is that the majority of these studies do not explicitly deal with indirect pleiotropy (also discussed in (STEPHENS, 2013)). This fact can have different impacts, the first one being that the estimation of the frequency of pleiotropy

might be inflated, since some of these pleiotropic effects can be indirect pleiotropy, not direct ones. The second factor argues the contrary, since multivariate models tend to be more powerful than their univariate counterparts (SCHMITZ; CHERNY; FULKER, 1998; STEPHENS, 2013), we could expect more pleiotropy being mapped when using proper analyses.

The distribution of pleiotropic effects is also a relevant characteristic. The first model of how pleiotropy is distributed comes from Fisher's geometric model (FGM) (FISHER, 1930), and it assumes that one mutation affects all traits. One implication of this model is that the rate of adaptation decreases with the complexity of the species (ORR, 2000). A modular distribution of pleiotropic effects would avoid this limitation by restricting pleiotropic effects of mutations to a single module, preventing it to affect all other traits (WAGNER; ALTENBERG, 1996).

Modularity is indeed a ubiquitous feature in biological systems, and can be observed across different complexity levels (WAGNER; PAVLICEV; CHEVERUD, 2007). At the molecular level, the modularity of the relation between the phenotype and the genotype might stem from the fact that some genes were selected for acting in concert, such as, for example, in the case of HoxD gene cluster. The HoxD cluster is a group of 9 genes known to act in the development of limbs. These genes act in coordinate fashion the set patterns in the limb development (TARCHINI; DUBOULE, 2006) and are known to be under the regulation of some key regulators (SPITZ; GONZALEZ; DUBOULE, 2003). The regulation of these genes is hierarchical, key regulators have to be activated before the expression of these transcription factor encoding genes can be individually regulated (MONTAVON *et al.*, 2011). Growing data on chromatin topology and interaction shows that the activation of genes, even when these genes are not in close vicinity, in groups might well be the norm rather than exception (SEXTON; CAVALLI, 2015).

Based on these theoretical and empirical considerations, we expect that the regulation of the expression of RBP-coding genes is modular, where modules are composed by genes functionally related that share QTL among themselves, but also have unique QTLs.

## 1.4 QTL mapping in gene expression traits

The new molecular techniques from the beginning of the 21th century (like micro-arrays and high-throughput sequencing), together with the concepts and methods from quantitative genetics allowed the study of gene expression in a genome-wide scale (CAVALIERI; TOWNSEND; HARTL, 2000; JIN *et al.*, 2001; BREM *et al.*, 2002). These studies mapping QTLs associated with gene expression traits became known as eQTL and were the first to systematically shed light on the genetic basis of gene expression. Surprisingly, the expression of many genes in *Drosophila melanogaster*, were shown to be strongly affected by sex, age and interaction between sex and environment, the genotype also explained the phenotypic variation, but to a lesser extent (JIN *et al.*, 2001). In yeast, (BREM *et al.*, 2002) showed one of the earliest evidences of polygenic effects (when one trait is affected by many *loci*) and also pleiotropy in gene expression. A similar

pattern can be seen in a study covering different organisms, the traits had ubiquitous polygenic basis and shared pleiotropic QTLs (SCHADT *et al.*, 2003). The distribution of explained variance is highly variable, with few QTLs able to explain almost 50% of phenotypic variance on their own (SCHADT *et al.*, 2003).

Gene-targeted mutation, RNA interference, chromosome conformation capture and mRNA over-expression and silencing are also advancements that contributed to the QTL study, since these methods help in the validation of the findings. The results of QTL mapping have also been functionally validated (TESLOVICH *et al.*, 2010; WEN *et al.*, 2011; PARÉ-BRUNET *et al.*, 2014; LAWRENSON *et al.*, 2015), showing that this method is indeed powerful to uncover the genetic basis of gene expression.

Studies mapping eQTLs have validated many biological causes for the association between genotypic polymorphisms and gene expression variance, but pleiotropic effects remain neglected by most works. In spite of th recognition of pleiotropy as relevant feature of the relation between genotype and phenotype since the first studies of eQTLs (e.g. (BREM *et al.*, 2002)), a search in the database Science Direct with the term “eQTL” returns 1406 entries, if we add the term “pleiotropic” this number falls to 175 and if, instead, we use “pleiotropy” this number fall to 92 entries. Using the PubMed data base yields a similar pattern, with 1198, 27 and 17 entries respectively (searches conducted in August/2017). This lack of pleiotropic QTLs would be acceptable in morphological traits, since measuring multiple trai

I‘d point out that this is understandable for mapping of the phenotypic raits, because studying pleiotropy requires that one measures more than one trait. But in the case of eQTL this wouldn‘t be a problem.

Of those studies that do mention pleiotropic QTLs in gene expression, few use appropriated methods, raising doubts about the generality of their findings. As discussed in section 1.3, not taking into consideration the possibility of indirect effects might lead to some biases on the QTL mapping results. Studies like (MORLEY *et al.*, 2004; WEST *et al.*, 2007; SCHADT *et al.*, 2008), for example, refer to “master regulators” or “eQTL hotspots” but do not take into account indirect pleiotropy, a feature that should be relevant in gene expression traits. Indirect effect are relevant in gene expression data since the expression of one gene (a transcription factor, a rRNA, a RBP or a miRNA, for example) might affect the expression of other genes.

## 1.5 Overview

In section 1.1 we discuss the relevance of RBP-coding genes to the regulation of gene expression in different levels and the lack of studies regarding the regulation of these these in a genome-wide scale.The methods and concepts presented in section 1.2 then can be used to dissect the genetic basis, the regulation of RBP-coding genes, pinpointing the *loci* linked to the genes or regulatory elements associated with each phenotype. As seen in section 1.4, given

sufficient sample size and effect size, these methods are capable of identifying the biological causes of phenotypic variability.

Section 1.3 discuss the theoretical and empirical evidences for the pleiotropic nature of the genetic basis of traits, being the traits morphological or molecular. Together with the discussion in section 1.4, we see that the pleiotropy is pervasive, but under appreciated in studies of gene expression.

With the aim of studying the regulation of RBP-coding genes, the proposed scenario then leads to two main objectives. The first one is the identification of genes related to the regulation of RBP-coding genes, and this can be further divided in the identification of individual and the pleiotropic effects, QTLs affecting single and multiple phenotypes. The second main objective is to study the extent of pleiotropy among RBP-coding genes, which can be dived into studying the frequency and the distribution of effects sizes of pleiotropic QTLs and how pleiotropy is organized in the relation between phenotypes and genotypes.

CHAPTER

# 2

# OBJECTIVES

The first aim of this study is to identify the genetic causes of variation in the expression of RBP-coding genes, this is described by two specific objectives:

- map and characterize QTL affecting individual gene expression traits,
- map and characterize QTL affecting groups of gene expression traits.

The second general objective can be described as understanding how pleiotropy is organized and distributed across RBP-coding genes, how hey, genetically, relate to each other. In more precise terms, this can be divided into two specific aims:

- describe the frequency distribution of pleiotropic effects as well as their magnitudes,
- test the hypothesis of modular organization of the genotype-to-phenotype relation.

CHAPTER

# 3

# METHODOLOGY

In this section we will describe the data obtained for the study (section 3.1), the pre-processing of this raw data (section 3.2), the simulations we built in order to assess the chosen methods (section 3.3) and the approaches used to analyze the pre-processed data (section 3.4)

## 3.1 Obtaining the Data

The raw gene expression data consists consists of duplicates of mRNA expression assessments in lymphoblastoid cell lines (LCL) of 717 individuals from eight HapMap populations (GIBBS *et al.*, 2003) (now part of the "1000 Genomes Project" (CONSORTIUM *et al.*, 2015)). We obtained the mRNA data used to study the genetic architecture of RBP gene expression in (STRANGER *et al.*, 2012). This data-set was obtained from the "ArrayExpress" data base under the ID "E-MTAB-198" and "E-MTAB-264".

## 3.2 Raw data pre-processing

### *3.2.1 Gene Expression*

The first step on the pre-processing of gene expression data is the mapping of the each probe from the microarray chip onto the human genome reference (left panel of figure 3.2). This step was done to ensure that each probe maps correctly onto the target gene and no other. For this step we used the package "illuminaHumanv2.db" (DUNNING; LYNCH; ELDRIDGE, ; GENTLEMAN *et al.*, 2004), which contains the information of the nucleotide sequence of each probe.

This filtering step was done using the 50 nucleotide sequence of each probe. These sequences were aligned to the human genome (hg38) using the "Blat" algorithm (KENT, 2002) and only probes that mapped to a single position on the genome with at least 75% of identity

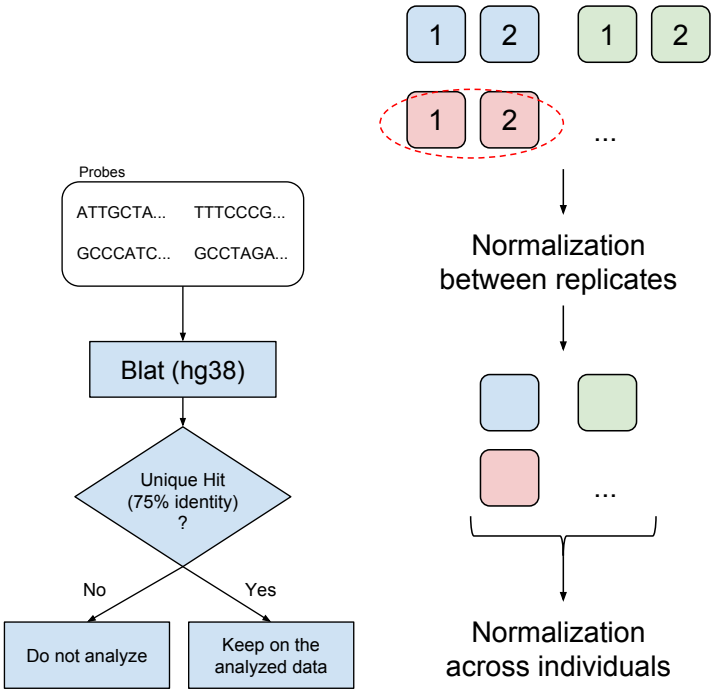



Figure 1 – Overview of the pre-processing steps. The left panel depicts the process of selecting probes that map correctly to their respective genes in the human genome reference "hg38". The right panel depicts the two-step procedure for normalizing the data used in this work.

were kept for further analyses. This way, of the more than 47 thousand probes, about 21 thousand were kept.

It is widely known that microarray technology suffers from technical biases, i.e. variation in gene expression measurements that is not due to biological reasons, but may be attributed to differences in which reagents were used, differences in handling and other sources of technical variation. In order to correct for possible technical biases and to make a gene expression comparable across individuals, we performed normalization of the filtered raw data. This normalization procedure involved three different steps.

The first step in the normalization procedure is to take the log2 of the raw expression value, which ensures the normal distribution of these expression levels. Having gene expression levels normally distributed is a desired characteristic, since many statistical analyses are based on the supposition that variables are normally distributed.

The normalization between replicates of the same individuals is the second step in the normalization procedure (right panel of figure 1). As both replicates are assessing the same characteristics of the same individual, large differences between them can be interpreted as being caused by technical factors, not biological ones. The two duplicates of each individual were normalized using quantile normalization from the "limma" Bioconductor package (RITCHIE *et al.*, 2015; GENTLEMAN *et al.*, 2004). Following this inter-replicate normalization, we performed a LOESS regression of one replicate onto another and the predicted values were

used as the expression values of a given sample. This procedure was done to summarize the expression values of both replicates into a single value for each gene for a given individual.

The final step is a quantile normalization across individual, also using the "limma" Bioconductor package (RITCHIE *et al.*, 2015; GENTLEMAN *et al.*, 2004). The reason here is to correct for technical variation while still maintaining the biological variation across different individuals. Both normalization steps (inter-replicates - step two; among individuals - step three) are based on a quantile normalization without background correction. This choice was based on the evaluation of different normalization procedures and their impact on data (SCHMID *et al.*, 2010).

### 3.2.2 SNPs

The matching genotype data was obtained from the former HapMap (GIBBS *et al.*, 2003) (now part of the "1000 Genomes Project" (CONSORTIUM *et al.*, 2015)). We included markers with minor allele frequency (MAF) greater than 5% and no evidence of deviation from Hardy-Weinberg Equilibrium (HWE) ($p\text{-}values \geq 0.01$). In order to use only independent markers we included only those markers which did not show Linkage Disequilibrium (LD) $r^2 \geq 0.5$. Besides MAF, HWE and LD filters, we included only autosomal SNPs. In the end, about 140 thousand SNPs were used in this study out of the initial more than 1 million SNPs.

### 3.2.3 Populational stratification

The individuals used in this study comes from eight different human populations:

- Northern and Western European ancestry in Utah, USA (CEU);
- Han Chinese in Beijing, China (CHB);
- Gujarati Indian in Houston, USA (GIH);
- Japanese in Tokyo, Japan (JPT);
- Luhya in Webuye, Kenya (LWK);
- Mexican ancestry in Los Angeles, USA (MEX);
- Maasai in Kinyawa, Kenya (MKK);
- Yoruba in Ibadan, Nigeria (YRI).

This clustering of individuals in different populations is known to affect the results, and can cause spurious association between genotypes and phenotypes (LI, 1969; PRICE *et al.*, 2010).

In order to deal with this populational stratification we used Principal Components Analysis (PCA) on genotypic data to, first, identify possible differences in allele frequencies

among different populations and then used the Principal Components (PCs) as covariates in order to correct this effect (PRICE *et al.*, 2010).To determine the number of principal components that is sufficient to represent the population structure we used a clustering analysis based on the silhouette statistic (ROUSSEEUW, 1987).

The silhouette statistic measures the distance between the value of a given PC for a given individual to the mean of this individual true group, as well as the distance to the other groups means. If the distance of this individual to its true group is lesser than the distance to any other group, the silhouette gets positive values. If the distance to this individual true group is greater than the distance to any other group, the silhouette gets negative values. Given that we know beforehand the true number of clusters (populations) as well as which cluster a given sample belongs to, the silhouette statistic can help defining how many PCs better describe the data.

In our approach, we first applied the silhouette on the first PC and got the silhouette values for each individual and a median value for each population and a median value for all individuals. We, then, subsequently added one more PC to the analysis then got the median values again. So, in our first analysis, we only used the first PC, on the second, we used the first two PCs, in the third the first three PCS, and so on. The result of this approach can be seen in figure 2, where the first three PCs are the ones that cluster better the individuals into their populations, and these are the PCs that will be used to control for population stratification in following analyses. In figure 2 only the first ten PCs are represented since it only decays to when we add more PCs.

Another indicator of bias in genotypic data is the genomic inflation factor ($\lambda$, (DEVLIN; ROEDER, 1999)). Briefly, $\lambda$ indicates whether the allele frequencies have departed from HWE, so when $\lambda \approx 1$, indicates that the influence of data stratification is negligible, values $\lambda > 1$ indicate that the results will be biased by the data stratification. Applying tis indicator in our corrected data we see that $\lambda$ assumes the value of 1.0278, whereas the raw data suggested a bias ($\lambda = 1.2294$)

## 3.3 Simulations

In this section we describe simulations that will guide the analyses on real data. As our approaches to real data will depend on the results of these simulations, we will describe and discuss their results in this section.

Simulations described here have been used to assess and compare different approaches to estimate heritability (section 3.3.1), univariate QTL mapping (section 3.3.2) and bivariate QTL mapping (section 3.3.4). All simulated phenotypes were created using the software LDAK (SPEED *et al.*, 2012) and were based on genotypes from the former HapMap Project (GIBBS *et al.*, 2003; CONSORTIUM *et al.*, 2015) that were described in section 3.2.2.

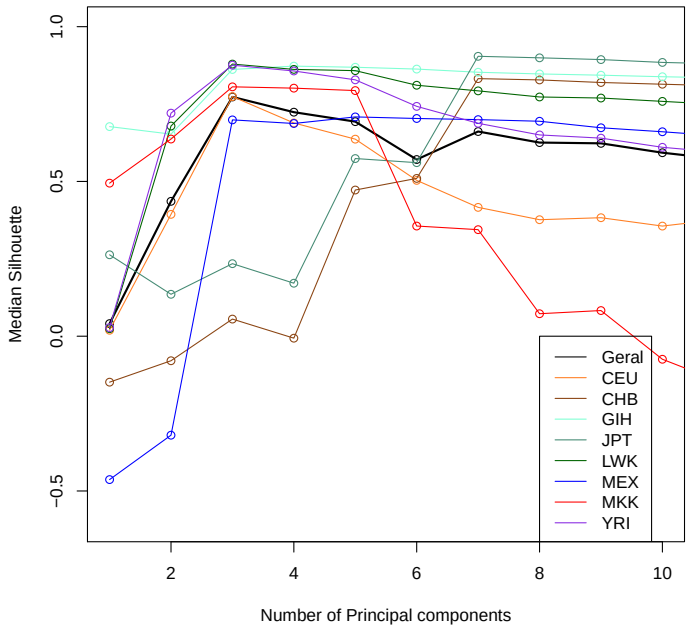



Figure 2 – The 'y' axis represents the mean silhouette value and the 'x' axis represents the number of Principal Components (PC) used to obtain the mean silhouette. The colors represents the mean silhouette in a given population and, in black, the mean silhouette for all populations. Only the first ten PCs are shown given that the mean silhouette after these components only diminishes.

### 3.3.1 Heritability

In brief, narrow sense heritability ($h^2$) is the proportion of the phenotypic variance that can be statistically explained by genotypic variation and will be used in this work as a criteria to select which phenotypes will be included in the study (a detailed discussion is given in section 3.4.1). So, to choose the best method to estimate the heritability of the phenotypes (gene expression) from SNP data we compared two approaches, one based on Mixed Linear Models (MLM) and another one based on a Bayesian models. The first approach, based on MLM, is a combination of two softwares (GCTA (YANG *et al.*, 2011) and LDAK (SPEED *et al.*, 2012)) and the second is implemented in the BSLMM software ((ZHOU; CARBONETTO; STEPHENS, 2013)).

we used heritability in order be able to only focus on the expression phenotypes in which we were powered to detect appreciable genetic basis

To assess and compare both approaches, ten different phenotypic conditions were created, and for each condition, 1000 phenotypes were simulated. These phenotypes were simulated using LDAK and the genotypes of the same 717 individuals used in this study (section 3.2.2) and we used the following conditions:

- three levels of heritability: 0.2, 0.5 and 0.8;

- three numbers of SNPs directly affecting the phenotypes: 10, 100 and 1000;
- one condition where the phenotypes have no relation with the genotypes.

This results in one set of 1000 phenotypes with heritability of 0.2 being directly affected by 10 SNPs, another set of 1000 phenotypes with heritability of 0.5 being directly affected by 10 SNPs and so on, and also one condition without relation between the variation in the phenotype and in the genotype.

We then applied both methods on these simulated phenotypes and the results are summarized in figures 3 and 4. Figure 3 depicts the heritability estimates given by the Bayesian approach for each condition. It is worth noting that this figure does not represents all ten situations. Given the enormous amount of time needed for each round of the Bayesian analysis, heritability estimates of conditions using 10.000 direct effect QTLs could not be assessed. The results of MLM approach can be seen in figure 4, where we have the estimates of heritability for each trait in each condition.

In both figures, 3 and 4, we can see that phenotypes in the zero condition (no relation with the genotypic variation) are indeed very close to zero and do not overlap much with phenotypes with modest and high heritability (0.5 and 0.8). Although there are some overlaps, we can select a hard threshold on the estimated heritability to filter out most of the false positives without compromising the results for phenotypes with modest to high heritability. Another similar pattern between the Bayesian and MLM approaches is the wide spread of estimations in any situation. Both these characteristics leads us to reason that, although both approaches are useful in discerning phenotypes that can be associated with the genotypes used in these analyses, the actual estimates should be interpreted with more caution.

Differences between the Bayesian and MLM approaches arise when we consider the precision of the estimates. Figures 3 and 4 show clearly that the estimates in BSLMM are generally inflated,they tend to be higher than the simulated value for that given condition. This inflation is not seen in MLM, where the estimates are centered on the simulated heritability.

Given these results and also the increased computational burden of the Bayesian approach, we chose to use MLM to assess the heritability of the the phenotypes studied here.

Figure 5 depicts the p-values (in -log10 scale) of each estimate and the horizontal lines represent commonly used thresholds (0.1 - green, 0.05 - blue and 0.01 - red). We can see that, even with a threshold of 0.1 we can safely reject more than 99% of phenotypes simulated to have zero heritability. A 0.1 threshold will lead to few false positives, although phenotypes with small heritability (0.2) will be mostly likely regarded as not having significant heritability.

To summarize this section, in the comparison between the two approaches Bayesian and MLM, we decided to use the former in our analysis of heritability in real data. We will describe this analysis in detail in section 3.4.1. Also, we will use $p - value \leq 0.1$ as a criterion to define

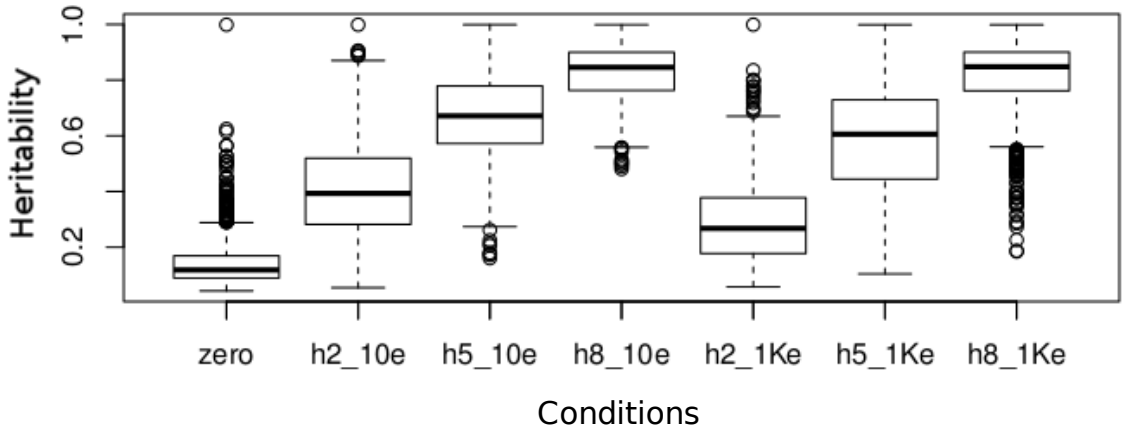



Figure 3 – BSLMM heritability estimates in each simulated condition. From left to right we have the condition with no true heritability, conditions with ten true QTLs ("10e") and conditions with a thousand true QTLs ("1Ke"). In each condition with true simulated heritability, we have three different levels of heritability: 0.2 ("h2"), 0.5 ("h5") and 0.8 ("h8").

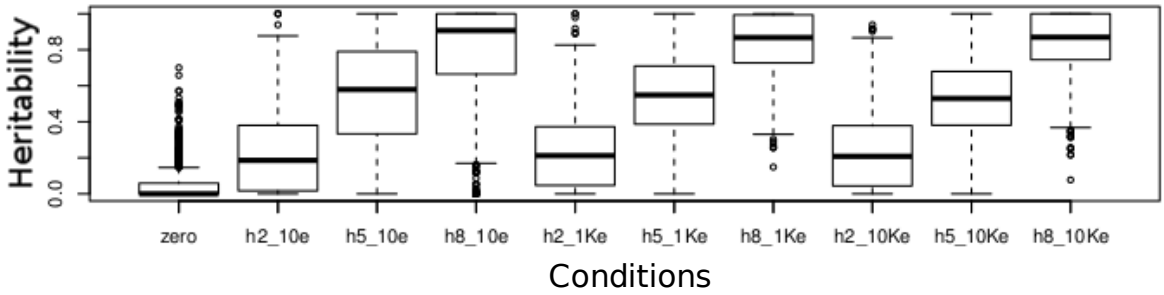



Figure 4 – GCTA heritability estimates in each simulated condition. From left to right we have the condition with no true heritability, conditions with ten true QTLs ("10e") and conditions with a thousand true QTLs ("1Ke") and conditions with a ten thousand true QTLs ("10ke"). In each condition with true simulated heritability, we have three different levels of heritability: 0.2 ("h2"), 0.5 ("h5") and 0.8 ("h8").

whether the phenotype will be included or not in our further analyses.

### 3.3.2 Univariate QTL mapping

Many QTL mapping methods are available, so, in order to select one that fits better the conditions of this study (structured population and 717 individuals), we conduced a assessment of three different QTL mapping methods. The first method is the simplest one, where each trait is regressed onto one *locus* at a time by fitting a linear model (LM) and it has been implemented in the software PLINK (PURCELL *et al.*, 2007). This LM has the advantage of being very fast to perform, even with large genotypic data (about one million *loci*, for example). One drawback

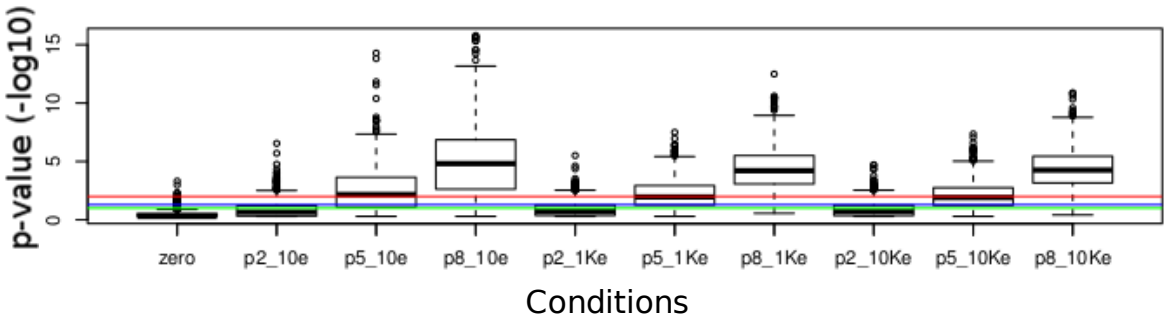



Figure 5 – GCTA heritability estimates p-values (in -log10 scale) in each simulated condition. Horizontal lines represent commonly used significance thresholds: 0.1 (green), 0.05 (blue) and 0.01 (red). From left to right we have the condition with no true heritability, conditions with ten true QTLs ("10e") and conditions with a thousand true QTLs("1Ke") and conditions with a ten thousand true QTLs ("10ke"). In each condition with true simulated heritability, we have three different levels of heritability: 0.2 ("h2"), 0.5 ("h5") and 0.8 ("h8").

of LM is that it performs one statistical test for each *loci* studied, leading to the multiple test problem, where one has to correct the significance threshold to account for testing multiple times the same data. Due to multiple testing, the current general consensus in the GWAS studies is to use a threshold of $10^{-8}$ to declare one association significant.

The second method that was assessed is similar to the LM, but it also takes into account the relationship between samples. This mixed linear model (MLM) approach is implemented through GCTA (YANG *et al.*, 2011). The advantage of this approach is that it explicitly models the interdependence between individuals, where in LM they are considered as independent. The drawback here is similar to the LM, as this model only tests one *locus* at a time.

The third, approach tested is implemented through BayesR software (BR). This approach is different from the previous one in many aspects, but the most relevant one is that it tests all *loci* simultaneously, which solves the multiple tests problems. One major drawback of BR is the computation time for each analysis, which is much greater than LM and MLM.

In order to evaluate these three approaches, we simulated three sets of hundred phenotypes each. These three sets were simulated using genotypes from the individuals included in this study (n = 717) and the phenotypes were simulated using LDAK (SPEED *et al.*, 2012) in order to have narrow sense heritabilities equal to 0.1, 0.5 and 0.9 respectively. All phenotypes had five true QTLs. We then applied all three methods to all three sets of phenotypes, and we recorded the p-values for each for each *locus* for both LM and MLM, and the posterior inclusion probability (PIP) for each *locus* in the BR results.

As is known the true QTLs of each simulated phenotype (and their respective positions in the genome), we recorded the p-values and PIPs for three sets of markers: the true QTLs ("TP"), the SNPs that are in a window of 10kb from each TP (we will call them "WTP") and markers that are at least 100kb away from the TP (we will call them "TN"). With these three distributions

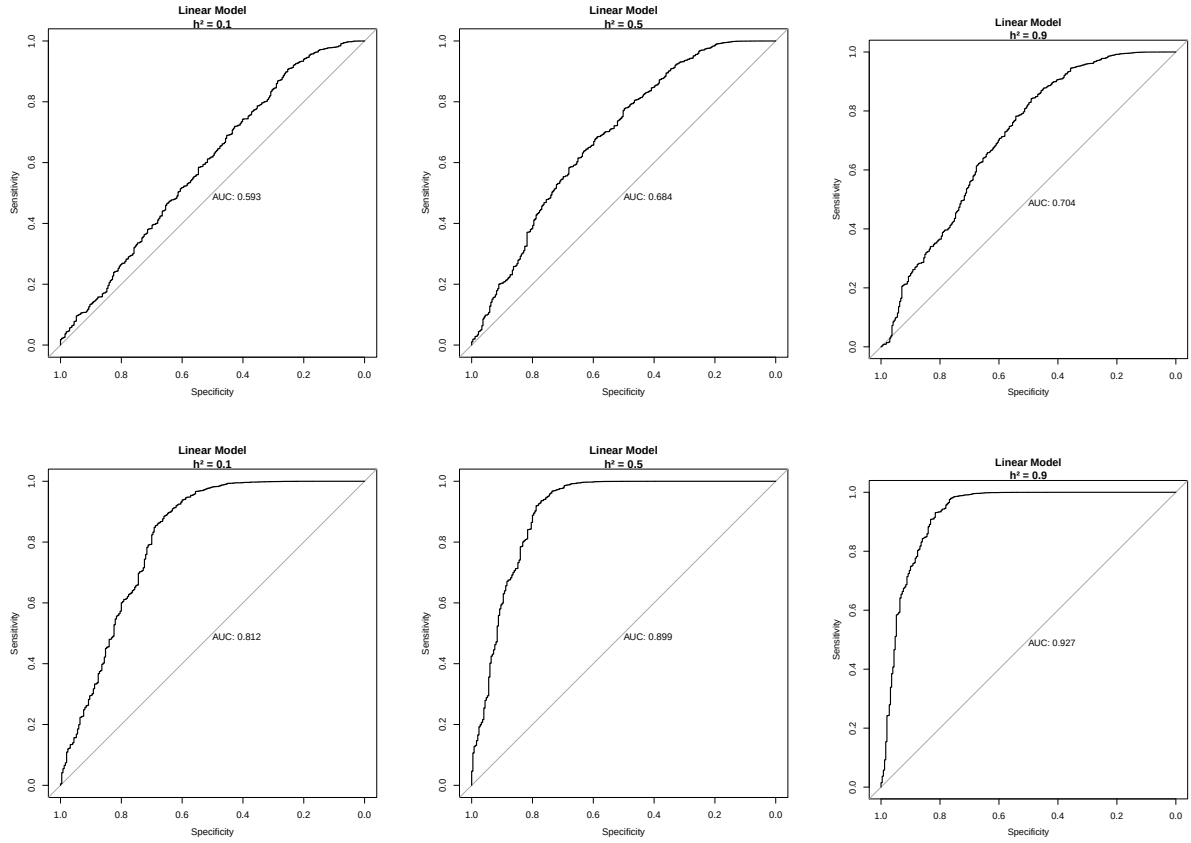



Figure 6 – ROC curves for QTL mapping using Linear Models.

(of either p-values or PIPs) for each condition and each approach, we used the area under the ROC curves (AUC) to compare the performance across conditions and approaches (HANLEY; MCNEIL, 1982), this method was implemented in R package "pROC" (ROBIN *et al.*, 2011).

ROC curves assess the ratio between true positive rate and false positive rate of a given analysis at different threshold values. This means that, if the studied analysis is good at recognizing true positives while avoiding accepting false positives the ROC curve will be skewed to the upper left corner of the plot. If the ROC curves lies on the diagonal, the studied analysis cannot identify true positives without accepting the same number of false negatives, i.e., the analysis is no better than chance. One way to evaluate the ROC curve is the area under the curve (AUC) (HANLEY; MCNEIL, 1982), the greater the AUC, the better the analysis is at recognizing true from false positives.

The ROC curves for LM, MLM and BR can be seen, respectively, in figures 6, 7 and 8. The upper row of these figures are the ROC curves for WTP and in the lower row for TP. Also, in all these figures, from left to right we have conditions with simulated heritability of 0.1, 0.5 and 0.9, respectively. The AUC of each ROC is represented on table 1.

One surprising caveat of the results shown in figures 6, 7 and 8 and also table 1 is that, although whole-genome models are generally regarded as more powerful models than their

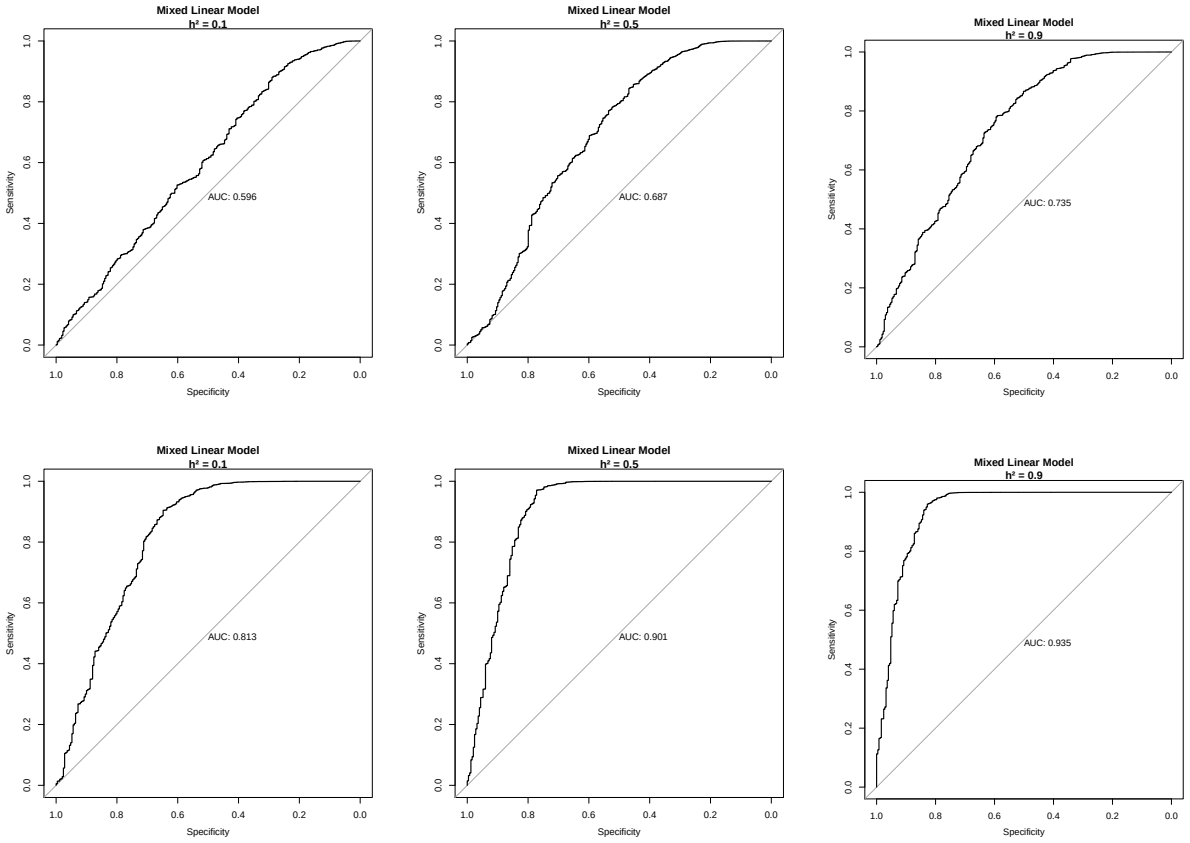



Figure 7 – ROC curves for QTL mapping using Mixed Linear Models.

Table 1 – Area under ROC curves (AUC) of each simulated condition.

| | Linear Model | | Mixed Linear Model | | BayesR | |
|---|---|---|---|---|---|---|
| Heritability | WTP x TN | TP x TN | WTP x TN | TP x TN | WTP x TN | TP x TN |
| 0.1 | 0.593 | 0.812 | 0.596 | 0.813 | 0.543 | 0.670 |
| 0.5 | 0.684 | 0.899 | 0.687 | 0.901 | 0.575 | 0.901 |
| 0.9 | 0.704 | 0.927 | 0.735 | 0.935 | 0.550 | 0.953 |

single *locus* counterparts, the AUCs from the three models are similar (at least in the conditions that were considered here).

One possible advantage of BR can be seen when we analyze closely why, different from LM and MLM, the AUCs from WTP x TN do not increase with heritability. This happens because BR can separate the effects of the true QTLs from the SNPs that are in the window around this QTL. This way, the PIPs of true QTLs are very close to one while the PIPs from the other SNPs are close to zero, which leads to a bimodal PIP distribution, affecting the ROC and AUC for these conditions. While this advantage of BR is relevant in simulated data, when analyzing real data we might not benefit from that, since the QTL *per se* are rarely known, and we have to consider that they might not even have been sequenced. When analyzing real data, it is more probable that the signals of association we are getting in each analysis are due to LD between the SNP and the true QTL.

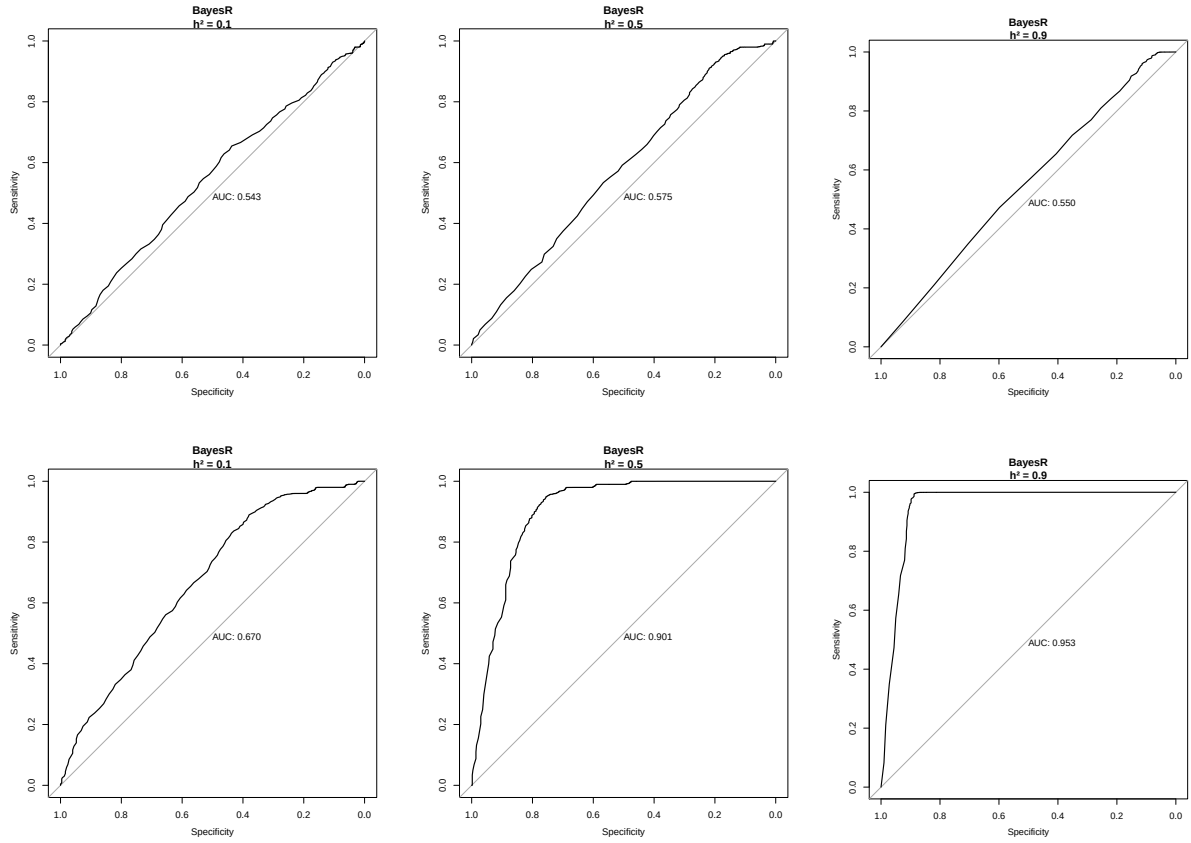



Figure 8 – ROC curves for QTL mapping using BayesR.

Considering the scenario described in this section, we decided to use the simple LM in all further QTL mapping analyses. A further explanation of this model and the steps used to perfor this anlysis will be discussed in section 3.4.2. Since there is not much difference between this model and the others, together with the facts that this model is fast to run and easy to implement, we judge that there is no need to use a more complex model.

Now that we defined the approach which will be used in order to map the QTLs, we should also define which will be the threshold value to define which associations are statistically significant and which are not. In order to do so, we will use the same simulated phenotypes as for assessing the three models (LM, MLM, and BR). We will vary the threshold for significance and check the proportion of WTP in that are significant among all the significant associations (all positives) as well as the proportion of WTP that are significant in each threshold (table 2 and figure 9).

In figure 9 we can see that, at least for the situations we simulated, if we want to guarantee any reasonable proportion of true positives in our results, we will inevitably have a great number of false negatives. For example, if we define that having 80% of our results being true positives is good enough, only about 9% of all true positives will be indeed recognized in the results (see table 2).

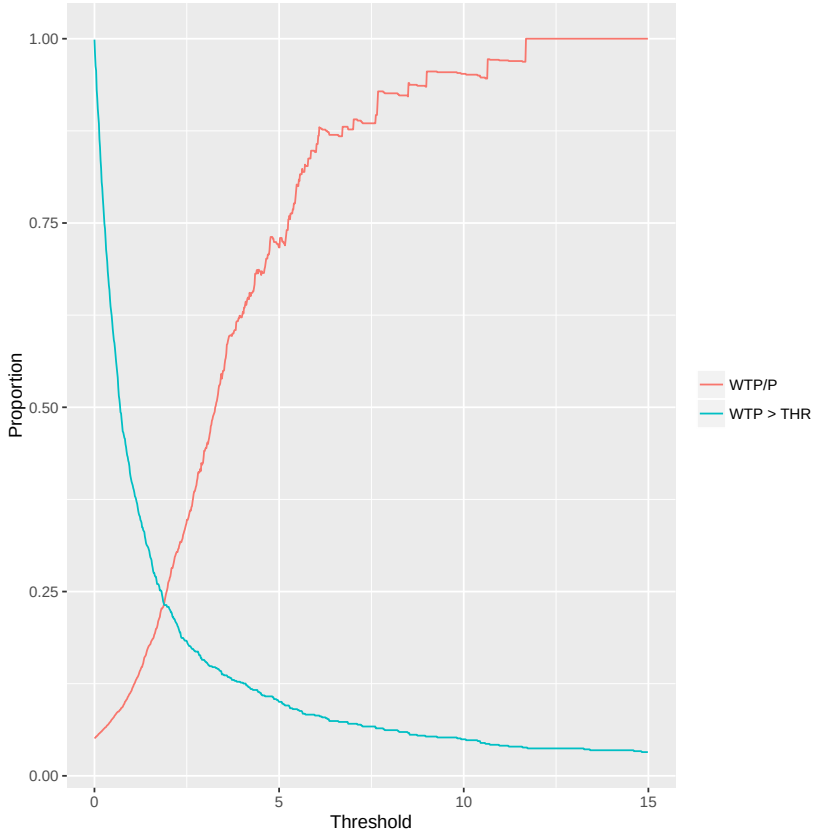



Figure 9 – "WTP" is proportion of true positives which are in a window of 10kb from the true QTL. "P" refers to all markers that are greater than a given threshold. In red we have the proportion of WTP in all P. In blue we have the proportion of all WTP greater than any given threshold ("THR").

Table 2 – Proportion of true positives and their respective thresholds.

| Threshold | Proportion of True Positives over all Positives | Proportion of all True Positives above the threshold |
|---|---|---|
| 5.46 | 0.80 | 0.0904 |
| 5.79 | 0.85 | 0.0817 |
| 7.02 | 0.90 | 0.0644 |
| 8.16 | 0.95 | 0.0532 |
| 11.68 | 0.99 | 0.0384 |

Given that the choice of significance thresholds are always arbitrary and may impact on the results, figure 9 and table 2 will help us understanding the impact of these choices on our results. Also, we will analyze our real data using three different thresholds, those that guarantee 85%, 90% and 95% of true positives in the simulated data (6.015, 7.665, and 9.000, respectively). These selected thresholds will be called "c85", "c90" and "c95" henceforth.

### 3.3.3 Genetic Correlation

As far as we know, GCTA (YANG *et al.*, 2011) is the only software to estimate the genetic correlation between two traits using SNP data. A more detailed description of how GCTA estimates the genetic correlation is given in section 3.4.1. In order to assess its performance

on the conditions we have in this work, we created two sets of five hundred phenotypes each, where the first is composed of pairs of related phenotypes and the second one of totally unrelated phenotypes. In the first set, for each pair, each phenotype had ten true QTLs, but out of these ten, at least two were shared between the pair of phenotypes. The second group of phenotype pairs were simulated having no common SNPs (not even from same chromosomes). The heritability of each phenotype, in both sets, was allowed to vary. We then applied GCTA on these phenotype pairs and the result can be seen in figure 10.

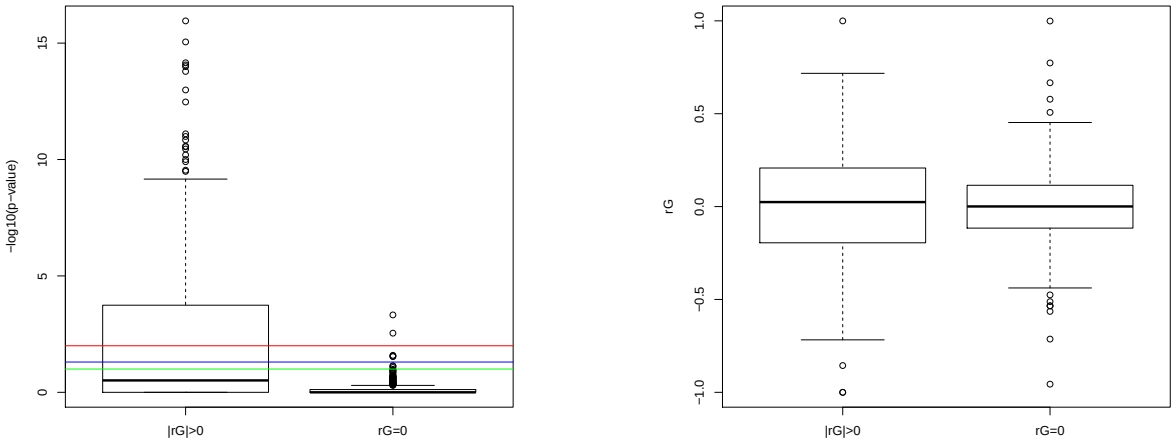



Figure 10 – Results of estimating genetic correlation using GCTA (YANG *et al.*, 2011) in simulated data. The left panel is showing the distribution of -log10 of p-values for each test and the right panel is showing the genetic correlation estimate. In both panels $|rG| > 0$ represents the pairs of phenotypes with true genetic correlation, while $rG = 0$ represents the phenotypes with no genetic relation between them. In the right panel, the horizontal lines represents commonly used significance thresholds, 0.1 (green), 0.05 (blue) and 0.01 (red).

Based on figure 10, specially on the left panel of this figure, we can see that, at least for the condition we simulated, GCTA can identify few false positives. On the other hand, when we take the horizontal lines from the left panel of figure 10 as guides, we see that the number of false negatives is high, even when considering not strict thresholds. For these reasons, we choose a loose threshold of 0.1 that still guarantees the rejection of more than 99% of false positives.

### 3.3.4 Pleiotropic QTL mapping

Based on the pairs of related phenotypes generated in section 3.3.3, we will assess the performance of a multivariate approach (STEPHENS, 2013) for mapping pleiotropic QTLs implemented in BIMBAM (SERVIN; STEPHENS, 2007). We will give a complete description of this approach in section 3.4.3.

In figure 11 we have, in the left panel, the ROC curve for pleiotropic association and its respective AUC. On the left panel is depicted the proportion of all true positives that are

Table 3 – Desired proportion of true positives and their respective thresholds for bivariate QTL mapping

| Threshold | Proportion of True Positives over all Positives | Proportion of all True Positives above the threshold |
|---|---|---|
| 4.771 | 0.80 | 0.1708 |
| 5.224 | 0.85 | 0.1602 |
| 5.847 | 0.90 | 0.1437 |
| 6.867 | 0.95 | 0.1234 |
| 12.361 | 0.99 | 0.0558 |

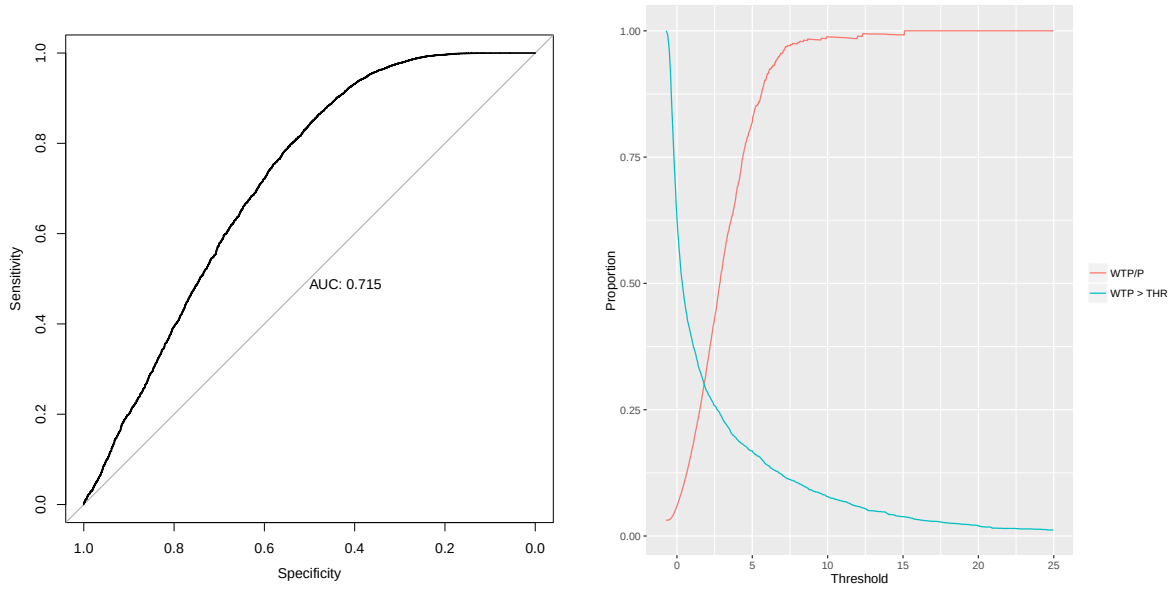



Figure 11 – Left panel, the ROC curve for pleiotropic association and its respective AUC. On the left panel is depicted the proportion of all true positives (WTCP) that are above a threshold (blue curve) and the proportion of true positives over all positives for a given threshold (red curve)

above a threshold (blue curve) and the proportion of true positives over all positives for a given threshold (red curve). Table 3 shows the threshold levels that guarantees that the proportion of true positives in all positives is above a desired level, which is directly related to the red curve of the right panel of figure 11.

As discussed in the end of section 3.3.2, the choice of thresholds are arbitrary, but these results might help us in understanding the impact of these choices in our results. Here, we will also choose three different significance thresholds, those that correspond to 0.85, 0.90 and 0.95 of true positives in the final results (table3).

## 3.4 Analyses of real data

In this section we will describe the analyses that were chosen in section 3.3 to be applied in our data set. Here the analysis used to study each characteristic of the genetic architecture of RBP-coding genes will be further described.

### 3.4.1 Heritability and Genetic Correlation

In section 3.3.1 we tested two different approaches to estimate heritability from SNP data and concluded that the approach based on MLM was faster and more precise than the Bayesian for the conditions we were able to explore. In this section we discuss the idea behind this approach as well as the procedures and steps taken in the analysis.

Considering that the phenotypic variance ($P$) can be explained by the variance in the genotype ($G$) and the variance in the environment ($E$), we have that $P = G + E$. The broad sense heritability ($H^2$) is the proportion of the phenotypic variance explained by the total variance in the genotype, $H^2 = G/P$. The genotype variation can be further separated in the variation attributed to additive effects ($A$) and the variation attributed to non-additive effects ($D$), $G = A + D$. In brief, the additive effects are the effects one allele has irrespective on the status of the other allele. In contrast, the non-additive effect effect of one allele is dependent on the status of the other allele, one example is dominance, where the effects of the allele $a_1$ on the phenotype is dependent on whether the other allele is $a_1$ or $a_2$. The proportion of phenotypic variance explained by the variance on the additive effects is called narrow sense heritability ($h^2$), $h^2 = A/P$. Both heritabilities, broad and narrow sense, assume, among other things, that the interaction of genotype and environment is negligible. This assumption poses no problem to this work since the gene expression data comes from immortalized cell lines, which have been cultivated all in the same media.

One of the main points of the narrow sense heritability is that it gives the upper bound of phenotypic variability that can be explained by the QTL mapping for additive affects. As the main goal of this work is to map QTLs associated with RBP-coding genes expression, $h^2$ is used to select those phenotypes that will be used in further analyses.

Narrow sense heritability is estimated based on the genetic resemblance of individuals. Briefly, a trait has $h^2 > 0$ when individuals more closely related have more similar phenotypes between themselves than with individuals more distantly related (accounting for the effect of the environment and other assumptions). To estimate the genetic resemblance across individuals from SNP data one has to determine the genetic variation of each individual and also the genetic covariation between individuals, this genetic variance and covariance matrix will be termed genetic similarity matrix (GSM). LDAK (SPEED *et al.*, 2012) builds the GSM from SNP data, just like GCTA (YANG *et al.*, 2011), but the difference between them is that LDAK takes into account the LD between SNPs. Since SNPs in hight LD carry similar information, inclusion of both SNPs would inflate the similarity between individuals. To deal with this problem LDAK weights SNPs based on the LD (SPEED *et al.*, 2012), the higher the LD across a given region, lesser will be the weights of each SNP in that segment. The GSM was estimated using all individuals from the former HapMap (GIBBS *et al.*, 2003; CONSORTIUM *et al.*, 2015) using all autosomal SNPs. The author recommends, on LDAK manual, that, for genotypes with more than 700 thousand markers, all the steps for the generation of the GSM to be run twice, and we

followed this recommendation.

After estimating the GSM, we used GCTA (YANG *et al.*, 2011) to estimate the heritability of each trait using the PCs obtained in section 3.2.3 and individuals sex as covariates. The estimation of variance explained used in GCTA is based on Restricted Maximum Likelihood (REML) where the variance/covariance components are estimated after taking into account fixed effects. The idea underlying GCTA is check whether the variation structure of the phenotype resembles the variation/covariation structure of the GSM.

Genetic correlation is defined as the correlation of additive effects ($A$) between two traits and can be useful to shed light into pleiotropy. Genetic correlation can assume values between -1 and 1, where -1 imply that both traits share the same "regulation", but the effects are in opposing directions and 1 means that the effects are in the same direction. We should also note that the genetic correlation between two traits should be interpreted with caution, since different genetic basis can lead to similar values of genetic correlation (HOULE, 1991; GROMKO, 1995). As heritability will be used to select which phenotypes we have enough power map QTLs, genetic correlation will be used to select which pairs of traits will be used to map pleiotropic QTLs.

The estimation of genetic correlation was carried out using the same GSM as for heritability estimation and can also be performed by GCTA (LEE *et al.*, 2012) using the "Bivariate" option. The idea underlying genetic correlation is trying to see whether the variance/covariance matrix of two phenotypes can be explained by the variance/covariance matrix of genotypes (GSM).

The results of these analyses will be presented and discussed in section 4.1 for heritability and section 4.2 for genetic correlation.

### 3.4.2 Univariate QTL mapping

In order to map the associations between *loci* and phenotypes of interest (the gene expression of RBP-coding genes), we applied univariate QTL mapping using PLINK (PURCELL *et al.*, 2007), through its "Quantitative trait association" option. The statistical test applied by PLINK is a simple linear regression where the phenotype is the response variable and the state of each *locus* is the explanatory variable. The genotype of each *locus* was encoded as 0, 1 or 2, representing the number of minor alleles for that individual in a given *locus*. As covariates, we used the PCs obtained in section 3.2.3 together with the gender of each individual. The significance thresholds were determined in section 3.3.2 and no correction in p-values were done.

The results of this analysis will be presented and discussed in section 4.3

### 3.4.3 Pleiotropic QTL mapping

To map the *loci* that have pleiotropic effects, i.e. affect multiple phenotypes, we used the approach implemented in mvBIMBAM (SERVIN; STEPHENS, 2007) (STEPHENS, 2013).

Mapping pleiotropic QTLs has two major difficulties. The first one is due to linkage, where two different genes are tightly linked to one SNP which, in turn is associated to two phenotypes. In this situation we cannot tell whether the two phenotypes are indeed affected by the same gene. This difficulty can only be solved by increasing the number of markers analyzed (which would unfortunately increase the multiple tests problem). The second major difficulty is indirect pleiotropy, a situation where one gene affects one trait which, in turn, affects a second trait. In this situation if we test the association between the genotype and both traits we will have that this gene is affecting both phenotypes, when it is not the case.

In order to deal with indirect pleiotropy we used one approach implemented in the mvBIMBAM software (STEPHENS, 2013; SERVIN; STEPHENS, 2007). This approach compares different models relating the genotype of one *locus* to two traits. For example, let us suppose that we aim to test whether the *locus* $G$ is associated with two phenotypes $P_1$ and $P_2$. In this scenario there are four possible outcomes: $G$ could be associated with $P_1$ but not with $P_2$, it could be associated with $P_2$ but not with $P_1$, it could be associated with both phenotypes and it could be associated with none. The idea underlying mvBIMBAM is to compare all models that could describe the data (STEPHENS, 2013), in order to differentiate between the possible situations. The advantage of this approach is that it makes the interpretation of the results easier in the sense that we can separate direct statistical effects from indirect ones without further tests.

The approach used in mvBIMBAM has two limitations. First, it does not accept covariates in the test, thus each phenotype is regressed ontothe covariates and the residuals are used as the corrected phenotypes. The second limitation is the assumption of normal distribution of the phenotypes. In order to check if the normality assumption holds for our corrected phenotypes, we regressed each of them onto a normal distribution with the same size and same parameters (zero mean and variance of one) and then recorded the coefficient of determination ($r^2$) for each regression. With $r^2$ we can estimate the proportion of phenotypic variance explained by the normal distribution (figure 12, which shows us how similar to a normal distribution a given corrected phenotype is.

In figure 12, we can see that some corrected phenotypes are not well explained by normal distribution, which could impact on the results obtained using mvBIMBAM. We therefore avoided all corrected phenotypes with $r^2 < 0.95$, which amount to 49 phenotypes.

mvBIMBAM is not a genome-wide association tool, it tests the association between a multivariate phenotype and one *locus* at a time. To apply mvBIMBAM to the data used in this work, we tested every possible pair of gene expression that had significant genetic correlation (according to sections 3.3.3 and 3.4.1). Because analysis of pairwise trait associations amounted to more than 600,000 tests, we have refrained from testing higher degrees of pleiotropy.

The results of this analysis will be presented and discussed in section 4.4.

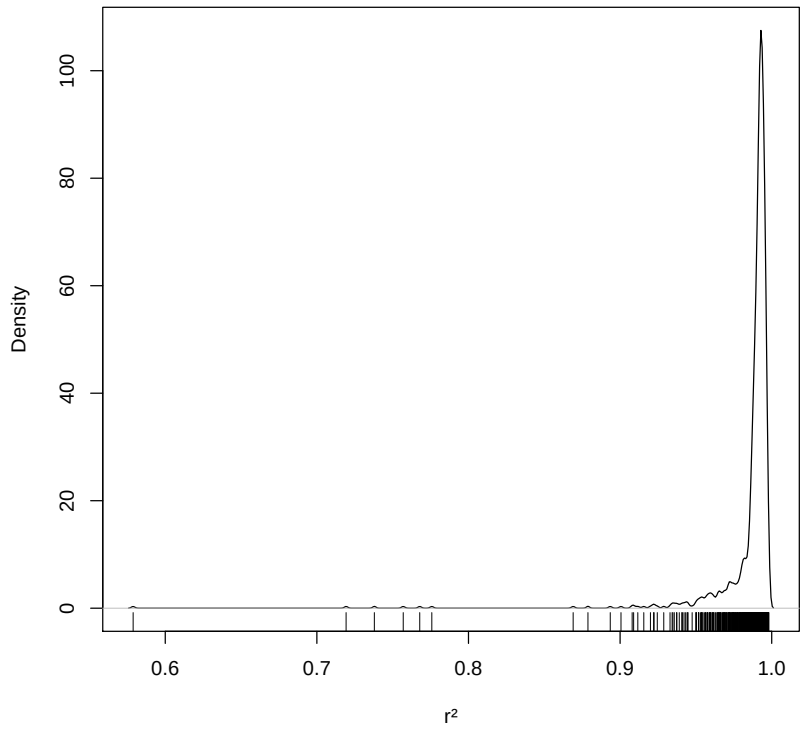



Figure 12 – Distribution of $r^2$ estimates for the regression of the corrected phenotypes and a normal distribution.

CHAPTER

# 4

# RESULTS AND DISCUSSION

## 4.1 Heritability

As discussed in section 3.4.1, heritability is used here to define the lower threshold for genotype-phenotype correlations that will be discussed in the study. Consequently, only the phenotypes with sufficient evidence for the genetic basis of variation will be included in our analyses. This way, only phenotypes with significant heritability at 0.1 threshold (see section 3.3.1) will be selected. The results of heritability estimates can be seen in figure 13.

Figure 13 shows that the majority of phenotypes show significant heritability, which means that there is enough phenotypic variance that is linearly correlated with the genotypic variance. This linear correlation of genotypic and phenotypic variability (heritability) indicate that it is possible to use QTL mapping approaches to try to identify possible elements that regulate the gene expression of these RBP-coding genes.

As expected from the results of section 3.3.1, the estimates of heritability show large standard errors ("y" axis in figure 13), so, although these estimates give us intuition about what portion of phenotypic variance that is statistically explained by the genotypic variance, the actual heritability might differ from the estimate. Also in this figure, we can see that only few phenotypes had low estimates (less than 0.25, for example). This lack of phenotypes with low heritability might be explained by the pattern seen in figure 5, where simulated phenotypes with heritability of 0.2 could not be efficiently separated from phenotypes with no heritability in any of the significance thresholds. Consequently, some of the estimates not significant in figure 13 can be false negatives.

To summarize the results from the heritability analysis, we can conclude that, considering the sample size and the SNPs we used in this study, we could find significant linear correlation between genotype and phenotype (heritability) for the majority of phenotypes (1114 out of 1380). These phenotypes with significant heritability will be used in further analyses.

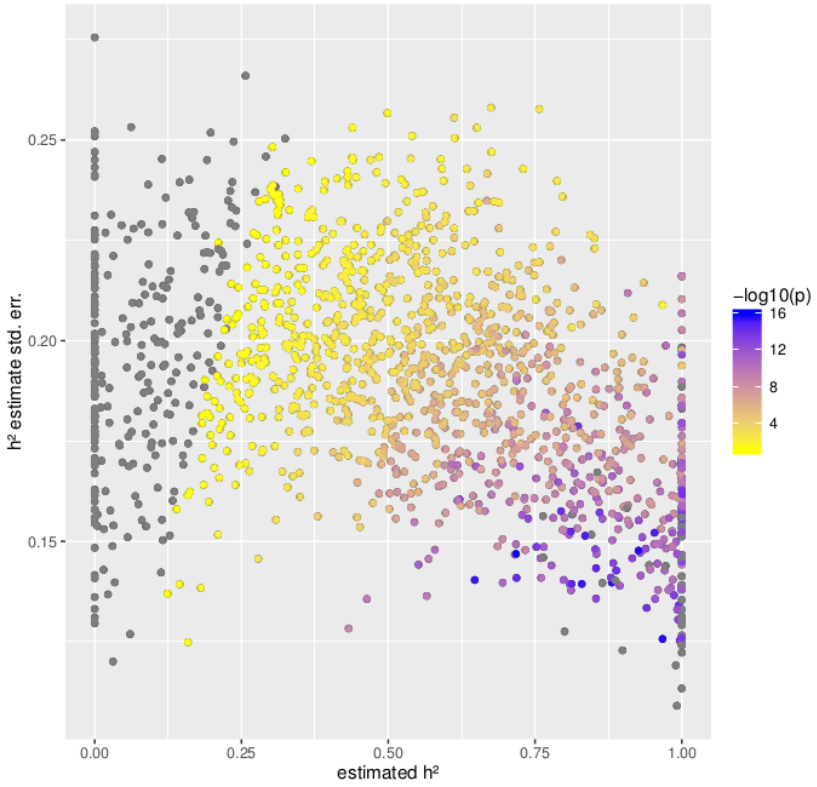



Figure 13 – The estimated heritability for each phenotype and the standard error for each estimate. The colors from yellow to blue represent the p-value of each estimate in -log10 scale. The data points in gray represent the estimates which had p-values above the threshold of 0.1, therefore not significant in our study.

## 4.2 Genetic Correlation

The genetic correlation estimate across all the phenotypes (described in section 3.4.1) are shown in figures 14 and 15. In figure 14 we can see that the genetic correlation across these phenotypes is ubiquitous and, in figure 15, that the magnitude of these correlations are generally higher than 0.5. Although this scenario leads us to expect hight levels of pleiotropic effects across all phenotypes, this might not be the case. In section 3.3.1 we saw that it is fairly easy to separate simulated phenotypes with zero heritability from the ones with heritability greater than zero, but actual estimates have high standard errors. This characteristic may also be relevant in genetic correlation estimation, since both approaches are based on the same methods. With this possibility in mind, we can infer that the genetic correlation is widespread among the expression of RBP-coding genes, but the magnitudes of these correlations cannot be precisely defined.

Just like heritability is being used in this work to filter out phenotypes that show no significant correlation with genetic variability, the genetic correlation is used to filter out pairs of phenotypes that show no significant correlation among their respective additive effects. Using the genetic correlation as a filter, out of 622.170 possible phenotype pairs, 185.469 had significant genetic correlation and will be used in the pleiotropic QTL mapping approach (section 4.4).

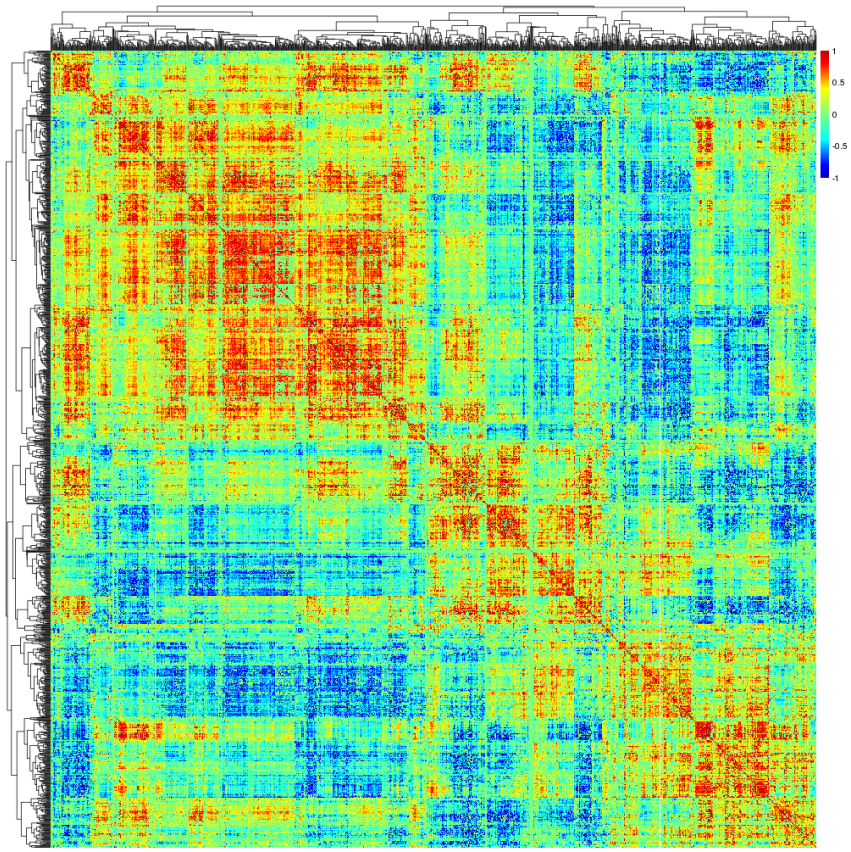



Figure 14 – The estimated genetic correlation for each pair of phenotypes with $p-value \leq 0.1$. Red colors represent positive genetic correlation, blue colors represent negative correlations.

## 4.3 Identifying univariate QTLs

In this section we will describe the results for the approaches described in section 3.4.2. In table 4 we can see the number of phenotypes which had at least one QTL in each threshold level. Figure 16a depicts the number of associations per phenotype, indicating that, dependent on the threshold level, multiple different *loci* are related to variation in the expression of each RBP-coding gene.

Figure 16b depicts the effect size of each association. It is possible to see that none of the identified QTLs had small, close to zero, effect sizes. This pattern is probably due to the conditions of this study, specially the sample size, that limits the effect size that can be confidently estimated.

Table 4 – Number of significant association per threshold level

| Threshold | Number of phenotypes with at least one significant association | Number of SNPs with at least one significant association |
|---|---|---|
| c85 | 771 | 1729 |
| c90 | 402 | 283 |
| c95 | 121 | 160 |
| c99 | 44 | 51 |

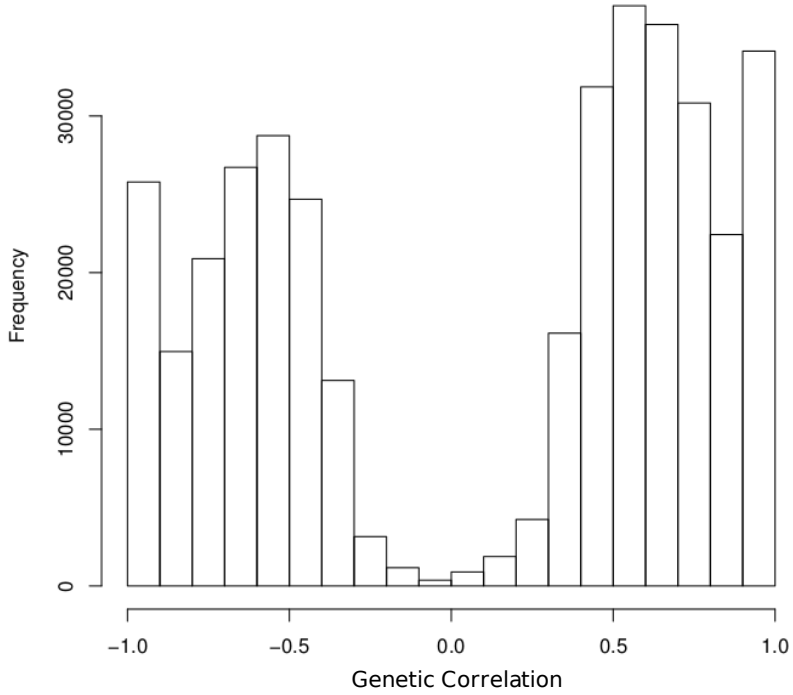
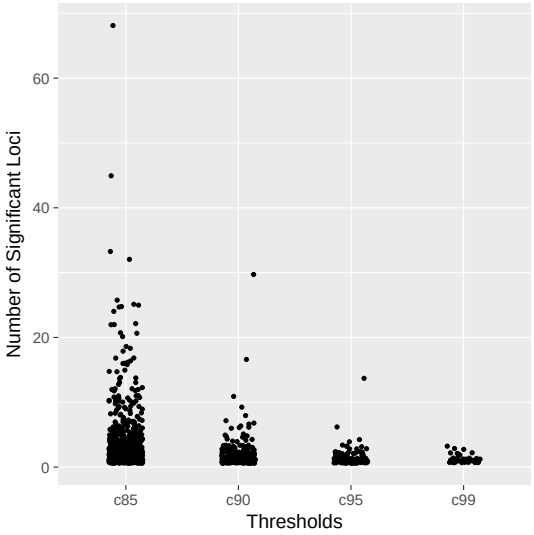



Figure 15 – The distribution of significant genetic correlations (at 0.1 threshold) for all pair of phenotypes.

In order to identify the possible genetic elements that underlie the effects of the QTLs found in the association mapping, we performed a three-step approach to annotate these *loci*. The first step is to locate each QTL on the human genome (hg38, figure 19)). The second step was done using the software Annovar (WANG; LI; HAKONARSON, 2010) to search for QTLs mapped onto exons, introns, UTRs (3' or 5'-UTR), non-coding RNAs and those QTLs that are up to 10kb away from any known gene. Besides these classes, we also searched for regulatory elements that are up to 10 kb away from the QTLs using the data from ENCODE specific for the cell line used in this work, GM12878 from the CEU population (ERNST; KELLIS, 2010). It is worthy emphasizing that there is some overlap across some of these classes, so for example, one QTL can be mapped onto an UTR region and also be less than 10kb away from another gene or from an enhancer. With this annotation step we were able to identify the genes that may be the biological effects underlying the QTLs.

The third annotation step focused on obtaining the genes and ncRNA that are mapped onto or close to the QTLs, and was done to classify genes into transcription factors, RNA-binding proteins and miRNA. These particular classes were chosen based on their role in different steps of gene regulation. Table 6 describes how many genes of each of these three classes could be found in the 10kb windows of univariate QTLs.

(a) Number of association per phenotype




(b) Effect size of each association

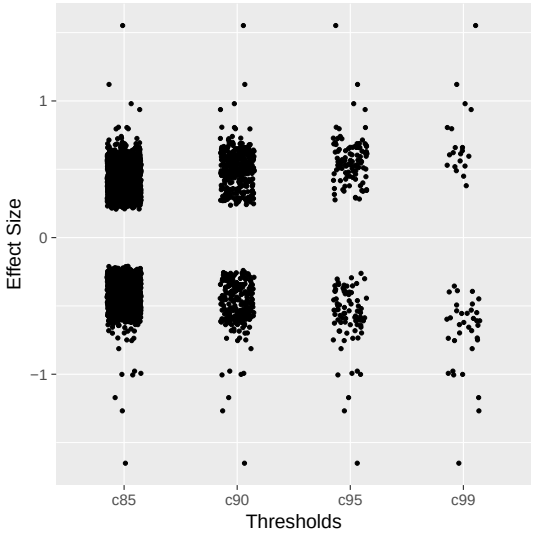



Figure 16 – Left panels depicts the number of *loci* associated with each phenotype that has at least one significant association. The right panel depicts the number of different phenotypes associated with a given *locus*. In both panels, the results are shown relative to the selects threshold.

Table 5 – Annotation of genetic elements up to 10kb from univariate QTLs

| Annotation | $c85 \leq x < c90$ | | $c90 \leq x < c95$ | | $c95 \leq x < c99$ | | $\geq c99$ | |
|---|---|---|---|---|---|---|---|---|
| | n | % | n | % | n | % | n | % |
| *Protein-coding* | | | | | | | | |
| exon | 33 | 0.0248 | 5 | 0.020 | 6 | 0.055 | 4 | 0.078 |
| intron | 570 | 0.429 | 99 | 0.409 | 43 | 0.394 | 23 | 0.450 |
| UTR | 64 | 0.048 | 9 | 0.037 | 4 | 0.036 | 3 | 0.058 |
| up-downstream | 150 | 0.113 | 37 | 0.152 | 16 | 0.146 | 7 | 0.137 |
| *Non-coding RNAs* | 25 | 0.018 | 4 | 0.016 | 5 | 0.045 | 9 | 0.176 |
| *Regulatory* | | | | | | | | |
| enhancer | 511 | 0.385 | 102 | 0.421 | 45 | 0.412 | 23 | 0.450 |
| insulator | 270 | 0.203 | 59 | 0.243 | 29 | 0.266 | 12 | 0.235 |

Table 6 – Transcription factors, RBPs and miRNA close to univariate QTLs

| Class | $c85 \leq x < c90$ | | $c90 \leq x < c95$ | | $c95 \leq x < c99$ | | $\geq c99$ | |
|---|---|---|---|---|---|---|---|---|
| | n | % | n | % | n | % | n | % |
| RNA-Binding Protein | 39 | | 6 | | 12 | | 8 | |
| Transcription Factor | 83 | | 9 | | 5 | | 3 | |
| miRNA | 8 | | 1 | | 1 | | 0 | |

## 4.4 Identifying pleiotropic QTLs

In this section we will describe and discuss the results from the pleiotropic QTL mapping described in section 3.4.3. Table 7 summarizes the the number of pairs of phenotypes that share at least one QTL, the number of phenotypes associated with at least one pleiotropic QTL and the number of SNPs that are considered pleiotropic QTLs.

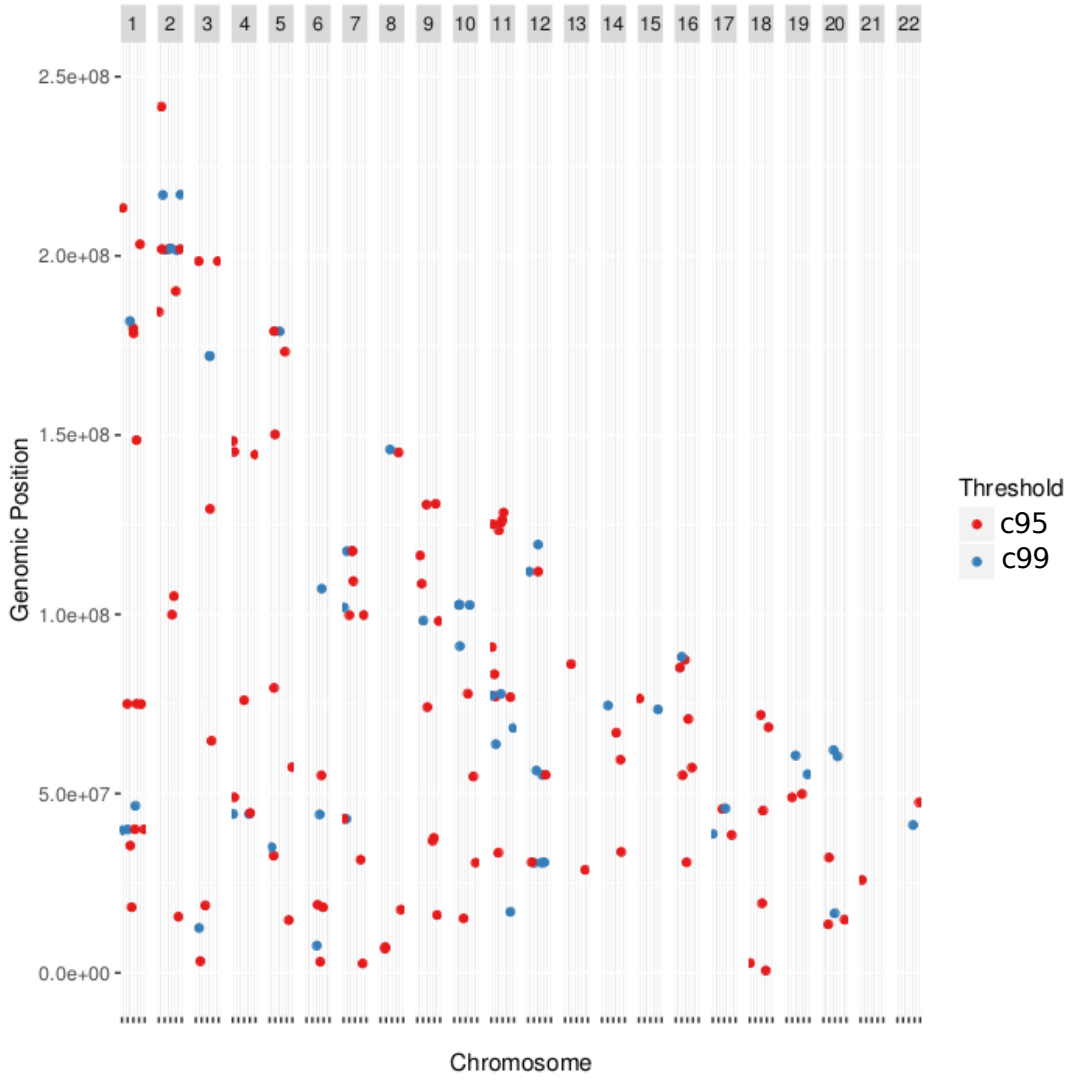



Figure 17 – The genomic position of positive association in each chromosome. Here only two thresholds are shown, c95 (red) and c99 (blue).

Given the initial number of 1.100 phenotypes with significant heritability (section 4.1), testing all possible pairwise phenotype combinations lead to more than six hundred thousand pairs of phenotypes, of which a little over one third had significant genetic correlation (section 4.2). Of all the pairs with significant genetic correlation, only a small fraction shared pleiotropic QTLs (2.893 considering c85 threshold; see table 7). Although few of all possible pairs shared pleiotropic QTLs, the majority (930) of phenotypes were found to have at least one pleiotropic QTL (table 7) which indicate that pleiotropic effects are organized in clusters, where phenotypes in one cluster share pleiotropic effects among the group, but not with other phenotypes.

Table 7 also shows, we can see that the number of pleiotropic QTLs is less than one third of the number of pairs affected by these QTLs. From this, we expect that one pleiotropic QTL is shared between more than one pair of phenotypes. In figure 18 we can see that this is indeed the case, but the distribution is not uniform across *loci*, while some QTLs are associated with many different phenotypes, the majority of QTLs affects few phenotypes.

Table 7 – Number of association in pleiotropic QTL mapping

| Threshold | Number of phenotype pairs | Number of phenotypes | Number of SNPs |
|---|---|---|---|
| c85 | 2893 | 930 | 767 |
| c90 | 1077 | 639 | 302 |
| c95 | 306 | 285 | 80 |
| c99 | 67 | 77 | 10 |

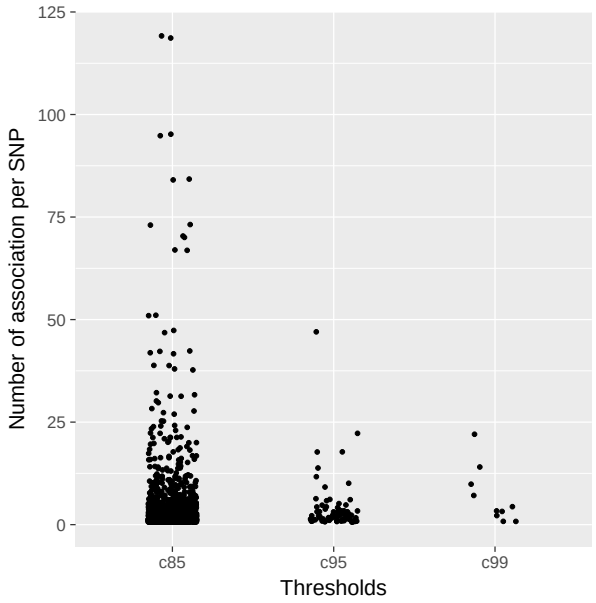



Figure 18 – Number of associations of each pleiotropic QTL.

The annotation of the pleiotropic QTLs follows the same steps as the univariate QTL annotation (section 4.3), and the results of each step are shown in figure 19 and tables 8 and 9.

Table 8 – Annotation of genetic elements up to 10kb from pleiotropic QTLs

| Annotation | $c85 \leq x < c90$ | | $c90 \leq x < c95$ | | $c95 \leq x < c99$ | | $\geq c99$ | |
|---|---|---|---|---|---|---|---|---|
| | n | % | n | % | n | % | n | % |
| *Protein-coding* | | | | | | | | |
| exon | 11 | 0.023 | 4 | 0.018 | 5 | 0.071 | 1 | 0.100 |
| intron | 192 | 0.412 | 106 | 0.477 | 36 | 0.514 | 7 | 0.700 |
| UTR | 21 | 0.045 | 11 | 0.049 | 3 | 0.042 | 2 | 0.200 |
| up-downstream | 47 | 0.101 | 24 | 0.108 | 9 | 0.128 | 1 | 0.100 |
| *Non-coding RNAs* | 5 | 0.010 | 6 | 0.027 | 4 | 0.057 | 1 | 0.100 |
| *Regulatory* | | | | | | | | |
| enhancer | 168 | 0.361 | 84 | 0.378 | 27 | 0.385 | 7 | 0.700 |
| insulator | 97 | 0.208 | 54 | 0.243 | 16 | 0.228 | 2 | 0.200 |

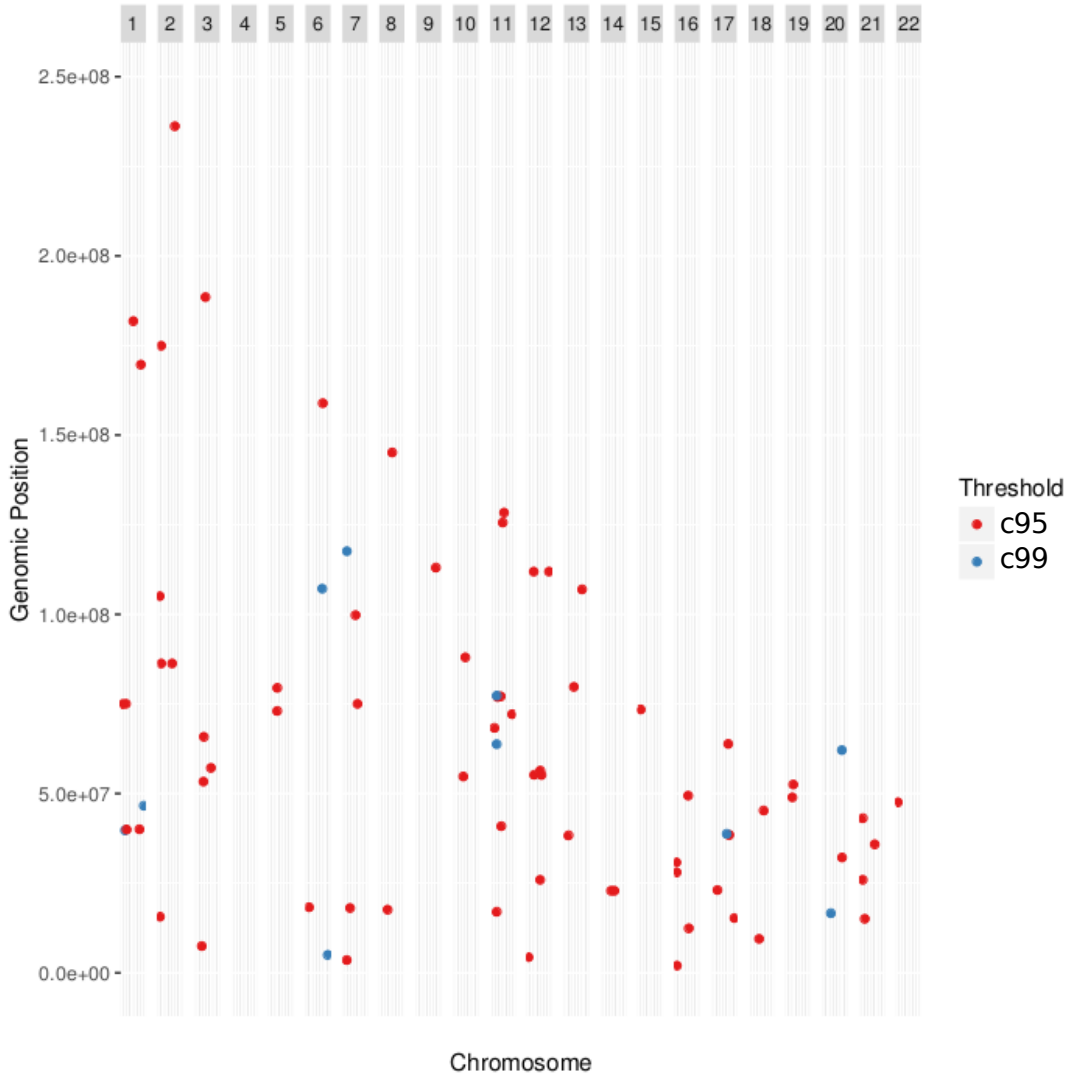



Figure 19 – The genomic position of positive association in each chromosome. Here only two thresholds are shown, c95 (red) and c99 (blue).

Table 9 – Transcription factors, RBPs and miRNA close to pleiotropic QTLs

| Class | $c85 \leq x < c90$ | | $c90 \leq x < c95$ | | $c95 \leq x < c99$ | | $\geq c99$ | |
|---|---|---|---|---|---|---|---|---|
| | n | % | n | % | n | % | n | % |
| RNA-Binding Protein | 19 | | 15 | | 12 | | 3 | |
| Transcription Factor | 31 | | 14 | | 5 | | 1 | |
| miRNA | 2 | | 4 | | 0 | | 0 | |

## 4.5 Comparison of univariate and pleiotropic QTLs

The results presented in section 4.3 show that QTLs affecting single phenotypes could be found for 771 expression values of RBP-coding genes (table 4), in contrast, section 4.4 showed us that pleiotropic QTLs could be found for 930 phenotypes (table 7). The intersection between these two groups shows that, from the 771 phenotypes with univariate QTLs, 677 of them also shared QTLs with other traits. These numbers might reflect a characteristic of the relation between genotype and phenotype, where pleiotropy is a ubiquitous feature in the regulation of

gene expression and at least as frequent as genomic variations affecting single traits. Earlier works mapping QTLs affecting gene expression also found substantial evidence of pleiotropic effects (BREM *et al.*, 2002; SCHADT *et al.*, 2003; MORLEY *et al.*, 2004; WEST *et al.*, 2007), although these earlier works did not account for possible indirect effects and did not use bivariate methods as we did (section 3.4.3).

Differences in the number of pleiotropic and univariate QTLs could be influenced by the differences in statistical power between the two association mapping approaches, since multivariate regression models have more power than univariate ones to detect additive effects (SCHMITZ; CHERNY; FULKER, 1998). Although we potentially underestimate the number of univariate QTLs in the face of pleiotropic QTLs, these "hidden" QTLs would have smaller effects than those already mapped (figure 16b) and would have small contributions to the phenotypic variation.

The effect of a pleiotropic QTL can be understood in two ways. First, we can see the effect of the pleiotropic QTL in each of the associated phenotypes separately. Looking at pleiotropic effects in this way some effects of pleiotropic QTL are indeed small, close to zero (figure 20a). Another way of looking at the effect size of pleiotropic QTLs is is regarding it as a vector in a multidimensional space, where each dimension in this space represents one phenotype affected by the pleiotropic QTL. Considering this multidimensional vector, its magnitude (and therefore the multidimensional effect size - $T$) is given by the square root of the squared individual effects ($A$) summed over all affected phenotypes ($N$) (WAGNER *et al.*, 2008):

$$T = \sqrt{\sum_{i=1}^{N} A_i^2} \tag{4.1}$$

Comparing the effect sizes of univariate QTL (figure 16b) with the size of pleiotropic effects (figure 20b), we can see that the magnitude of pleiotropic effects ($T$) is generally greater than the magnitude of univariate effects.

We should note that equation (4.1) used to calculate the total effect of a pleiotropic QTL assumes independence across all dimensions, and violations of these assumption would imply inflation of $T$. Inflation is expected in cases where the traits have any relation between themselves other than sharing the underlining genes. Since we explicitly accounted for indirect pleiotropy, inflation is not a relevant issue in this analysis.

There are two models that try to explain the expected effect size distribution of pleiotropic QTLs, the "limited" model (ORR, 2000; WINGREEN; MILLER; COX, 2003) and the "quadratic scaling" model (WAGNER, 1988). In brief, the limited model posits that the total effect of a QTL ($T$ from equation (4.1)) is constant, therefore the more traits one *locus* affects, the smaller the individual effects will be. The quadratic scaling model posits a positive relation between the number of phenotypes affected by a pleiotropic QTL and its total effect $T$, more specifically $T$ should scale with the square root of $N$.

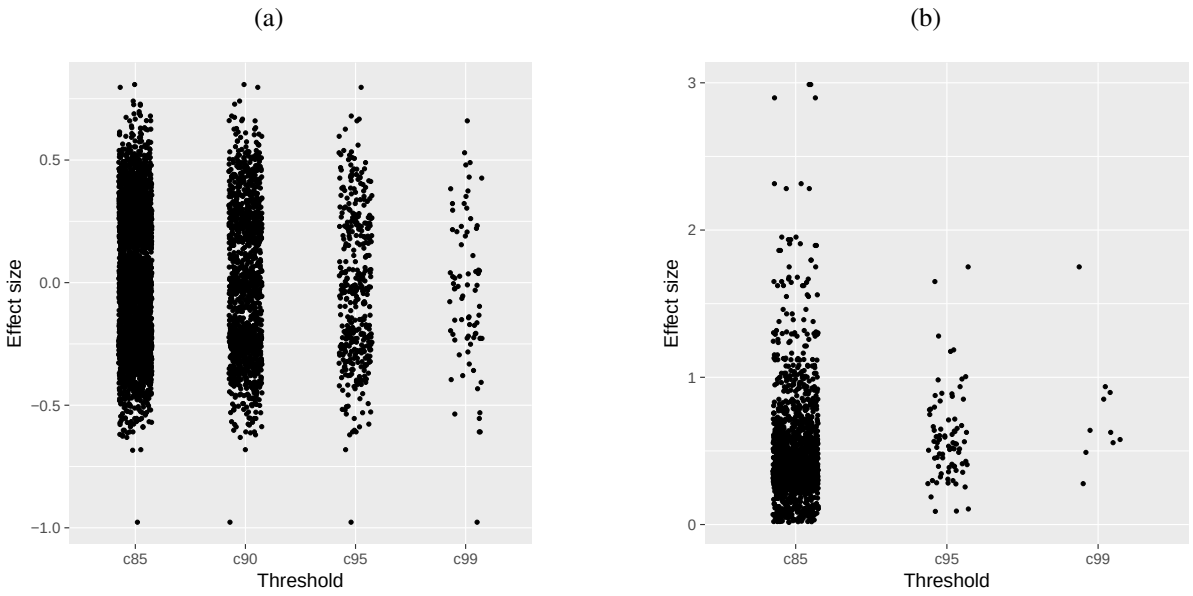



Figure 20 – Left panel depicts the effect of a pleiotropic QTL on each of its affected phenotypes alone. Right panel depicts the total effect size of a pleiotropic QTL (given by equation (4.1)).

Linear regression analyses of the relation between the number of affected phenotypes $N$ and the total pleiotropic effect $T$ clearly shows that, in this study, $T$ indeed scales with $N$, regardless of the chosen threshold (table 10). This result is in contrast to the expectation given by the the limited model. As for the quadratic scaling model, table 10 shows that regressions of $T$ on both, $N$ and $\sqrt{N}$ are significant, so the results presented in this work cannot decisively exclude or support the quadratic scaling model.

This is not the first study about the expected distributions of $T$ (WAGNER *et al.*, 2008). The results presented in figure 20b and in table 10 are similar to the results of earlier study, especially in terms of rejecting the limited model. An interesting feature is that this earlier study by Wagner *et al.* (WAGNER *et al.*, 2008) is based on the morphological characteristics of mice, and our study is based on human gene expression. If the distribution of pleiotropic size effects is indeed similar across different levels, would there be any common cause or biological explanation of this pattern across all complexity levels?

Table 10 – Total pleiotropic effect of a QTL - quadratic and non-quadratic model

| Threshold | $r^2$ | | beta | | p-value | |
|---|---|---|---|---|---|---|
| | $\sqrt{n}$ | n | $\sqrt{n}$ | n | $\sqrt{n}$ | n |
| c85 | 0.634 | 0.593 | 0.307 | 0.044 | <2e-16 | <2e-16 |
| c90 | 0.575 | 0.534 | 0.306 | 0.047 | <2e-16 | <2e-16 |
| c95 | 0.496 | 0.437 | 0.235 | 0.032 | 2.99e-13 | 2.49e-11 |
| c99 | 0.551 | 0.469 | 0.264 | 0.040 | 0.013 | 0.028 |

Pleiotropic and univariate QTL are also similar when we compare the classes of genetic elements they map to (exons, introns, enhancers and so on). We annotated all genomic elements

that fall up to 10kb away from these univariate QTLs just as we did in section 4.3 and then we compared against the results shown in table 8 using the $\chi^2$-test. In this analysis no p-value was smaller than 0.1, and we conclude that we do not have enough evidence to say that the frequency of annotated elements is different between pleiotropic and univariate QTLs.

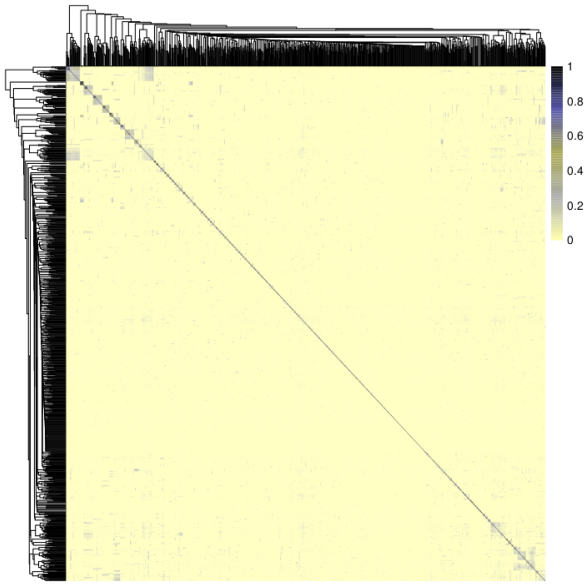
## 4.6 Modular structure of gene expression regulation

As we saw in section 4.4, although small percentage of all possible pairs indeed had detectable pleiotropic QTLs, the majority of phenotypes are affected by pleiotropic QTLs. To accommodate these two seemingly opposite results we reasoned that pleiotropy might be indeed structured in relatively small modules, where a group of phenotypes share pleiotropic QTLs among them but rarely so with a vast number of phenotypes belonging to other modules. This modular structure of the relation between phenotype and genotype has been suggested to underlie different biological traits (WAGNER; PAVLICEV; CHEVERUD, 2007).

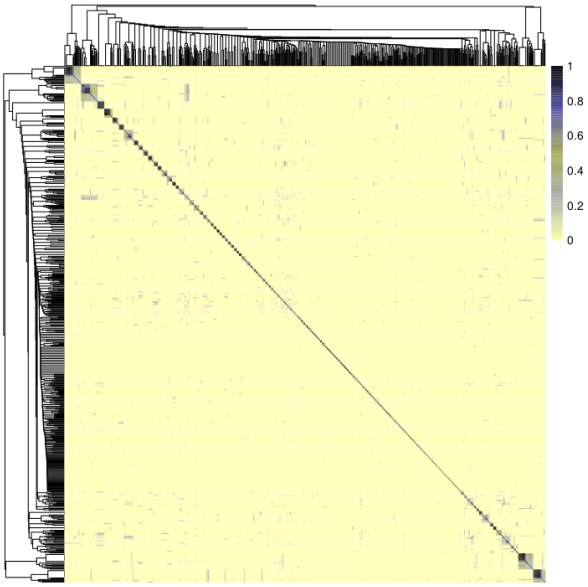
To test the possibility of such modular organization, we first obtained all the pleiotropic QTLs associated with each phenotype. For example, consider that the phenotypic pair $(X,Y)$ is associated with the marker $A$ and the phenotypic pair $(X,Z)$ is associated with the marker $B$, we have that $X$ is associated with both markers $A$ and $B$ (but that is not true for $Y$ and $Z$). Since the SNP associated with a phenotype is probably only correlated with the true biological effect, we used 20kb windows. This way, for every pleiotropic QTL affecting a given phenotype we also included every other SNP that is no more than 10kb away, up- or downstream, from the pleiotropic QTL. Now, instead of only having the pleiotropic QTLs for each phenotype, we have pleiotropic "regions" for each phenotype. We then compared the regions among all phenotypes that had at least one pleiotropic QTL. The similarity of pleiotropic regions between two phenotypes was calculated as the number of identical SNPs in the regions of both phenotypes over the number of SNPs in the union between the two regions. In this similarity index, zero means that the two phenotypes share no QTLs and one means that regions of two phenotypes are identical.

In the results shown in in figure 21 we can see that, independent of the threshold we consider, we detect a modular organization of the genotype-to-phenotype map. Modules in general tend to be small and most modules are composed by less than ten phenotypes (figure 21), although some can be composed by 30-32 phenotypes (upper left corner of figures 21c and 21d). The number of *loci* associated with a module is also highly variable, analyzing the first ten clusters from 21a we observe that some modules have only one pleiotropic QTL while others can have up to 80. With the exception of figure 21d, the actual number of phenotypes within some modules is not straightforward to determine, since it depend on arbitrary criteria. For example, let us consider either the third or forth module, from up left to bottom right in figure 21c. Both these modules have phenotypes that have identical genetic basis (black area), and

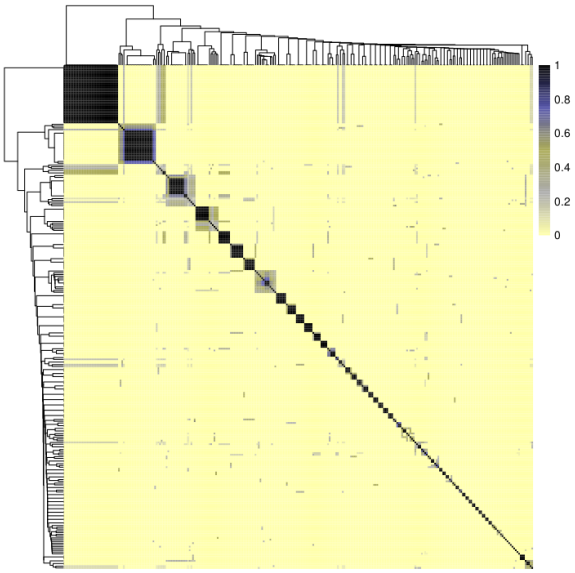
(a) c85




(b) c90




(c) c95




(d) c99

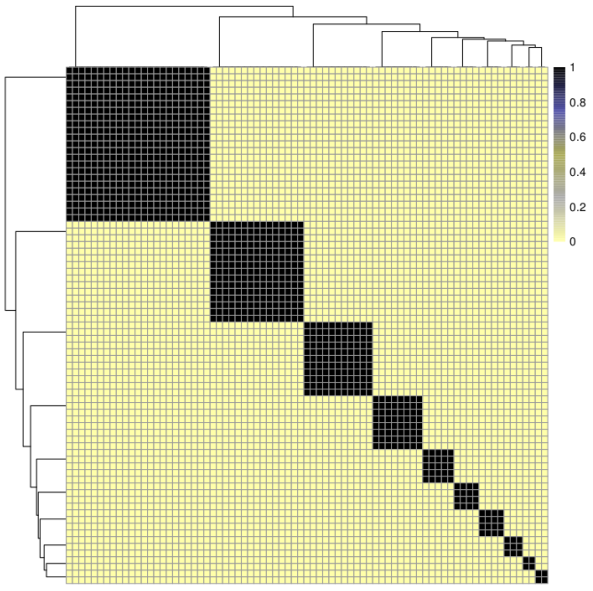



Figure 21

some that share some QTLs but not all of them are identical (gray area). The question is whether all these phenotypes should be considered as belonging to the same module or not, this choice would impact on both the number of phenotypes and QTLs in the module. As a consequence, the number of QTL per module also depends on the criteria used to define a module.

In order to investigate further the genetic basis of these modules of phenotypes, we first identified the genes belonging to each module from the c99 threshold (figure 21d), i.e., the genes that share their genetic basis. From the upper left corner to the bottom right:

- g1: DGCR8, ISG20L2, C17orf71, GTSE1, RUVBL2, ZNF579, DDX31, POLR2G, DHX34, IGHMBP2, CLEC3B, RPL37A, RBM25, CUGBP2, SMAD3, CWF19L2, PDCD4, MRPL55, TRNT1, SF3B5, JAKMIP1, INTS2, QRSL1;
- g2: NUDT16L1, PRPF40A, TRMT1, MBNL2, RBM16, MRPL21, THUMPD1, PPAN, KHDRBS1, ZC3HC1, SLTM, EIF5A2, NKRF, HSPC152, RPS19BP1;
- g3: NXT2, HNRPH2, GNL3, EFTUD2, PRPF6, UTP14A, CDK5RAP1, TDRD7, CIRBP, EIF2B1, ENDOG;
- g4: RBM17, RNASEH2A, PUM2, PAN3, NSUN4, RNMT, HNRPDL, MRPL11;
- g5: SNORD21, DDX42, SNORD43, RUVBL1, BRCA1;
- g6: DENR, HTATSF1, PAPOLB, SNRPB2;
- g7: MAZ, DAZAP1, MYST1, INTS4;
- g8: SFRS3, SAMHD1, PPIE;
- g9: C1orf25, LSM8;
- g10: RPP40, RRP15.

Over-representation analysis (conduced in the same way as sections 4.3 and 4.4) shows that these modules have genes that tend to play similar biological and molecular roles (table 11), although some of them did not show any significant enrichment given the small size of some modules. This evidence can lend support to the idea that the modules inferred in this work are indeed biologically meaningful units.

Each of these phenotypic modules has only a single pleiotropic QTL at the c99 threshold: rs7756087, rs28395880, rs3829704, rs6703669, rs11657004, rs2073055, rs672460, rs11488181, rs9655829 and rs707966, respectively. Nine of these *loci* had protein-coding genes in their respective 20kb windows (10kb up- and downstream) and seven of them had strong evidence of having active regulatory elements as well. Table 12 describes the protein coding genes found in the windows of each pleiotropic QTL of each phenotypic module.

## 4.7 Results overview

In this section we will briefly recapitulate the main findings and ideas presented throughout the section 4. The results described in section 4.1 and 4.2 are in accordance with previous results from the literature, showing that there is plenty of gene expression variance correlated with genotypic variance (SCHADT *et al.*, 2003; DIXON *et al.*, 2007) (figure 13) and also that there is substantial correlation of additive effects across all phenotypes (POWELL *et al.*, 2012) (figure 14).

Table 11 – Enrichment of phenotype modules

| GO Term | Observed | Expected | Fold Enrichment | Corrected p-value |
|---|---|---|---|---|
| *g1* | | | | |
| ncRNA processing | 6 | 0.42 | 14.32 | 2.35E-02 |
| ncRNA metabollic process | 8 | 0.59 | 13.66 | 4.96E-04 |
| ATP-dependent helicase activity | 4 | 0.10 | 38.57 | 1.05E-02 |
| *g2* | | | | |
| RNA processing | 7 | 0.63 | 11.18 | 9.10E-03 |
| *g3* | | | | |
| mRNA splicing, via spliceosome | 3 | 0.09 | 31.65 | 2.42E-02 |
| *g4* | | | | |
| mRNA processing | 3 | 0.09 | 32.15 | 2.05E-02 |
| *g5* | | | | |
| protein acetylation | 2 | 0.01 | >100 | 5.32E-03 |
| DNA repair | 2 | 0.02 | 85.90 | 4.33E-02 |
| helicase activity | 2 | 0.02 | >100 | 1.68E-02 |
| *g6* | | | | |
| RNA splicing, via transesterification reactions | 3 | 0.02 | >100 | 7.29E-05 |
| mRNA splicing via spliceosome | 3 | 0.03 | >100 | 1.54E-04 |

Table 12 – Protein coding genes up to 10kb away from pleiotropic QTLs (above c99 threshold)

| Module | *Locus* | Gene | GO classes |
|---|---|---|---|
| g1 | rs7756087 | RTN4IP1 | Zinc ion binding |
| g2 | rs28395880 | PLCB3, BAD | Phospholipase, phospholipid binding |
| g3 | rs3829704 | PRPF6,SOX18 | mRNA splicing, transcription factor |
| g4 | rs6703669 | FAAH | Amidase |
| g5 | rs11657004 | - | - |
| g6 | rs2073055 | SNRPB2 | mRNA splicing |
| g7 | rs672460 | INTS4 | snRNA processing |
| g8 | rs11488181 | BMP8A,MACF1 | Transcription factor, cytoskeleton |
| g9 | rs9655829 | NAA38 | NatC complex/transferase |
| g10 | rs707966 | RPP40 | rRNA processing |

Sections 4.3 and 4.4 provide the grounds to the discussion of two key issues in this work, the comparison between univariate and pleiotropic QTLs (4.5) and the structure of pleiotropy in gene expression (section 4.6). Here we also presented the annotations of the possible biological effects underlying these QTLs.

In section 4.5 we discuss evidence in support of the idea that pleiotropic QTLs are at least as common as the QTLs affecting only single traits (tables 4 and 7) and may also have comparable effect sizes (figures 16b and 20b).

Section 4.6 shows the modular organization of the regulation of the expression of RBP-

coding genes and that this modular organization in independent of the significance threshold chosen. Also, we show that the phenotypes in each module are enriched for some biological roles, which indicates that the genes found in each module are indeed biologically related to each other. Finally, we have annotated the pleiotropic QTLs common to ten modules (figure 21d), describing which genes, miRNA and regulatory elements could be found near these QTLs, so as to identify which are the possible elements underlining the regulation of each of these phenotypic modules (table 12).

CHAPTER

# 5

# CONCLUSION

In this section we present the major findings and contributions of this work. In sections 4.3 and 4.4, we were able to identify QTLs that are associated with variations in the expression of individual as well as groups of RBP-coding genes. We also characterized each *loci*, identifying the possible genes and regulatory elements that may be the biological cause of each QTL. In section 4.5 we show that pleiotropic QTLs are, at least, as frequent as univariate QTLs and their effect size is also comparable to the univariate QTLs. Finally, in section 4.6 we show that the regulation of the expression of RBP-coding genes is modular, this way subgroups of genes are regulated by similar elements but there is little or no similarity across different subgroups.

# BIBLIOGRAPHY

BREM, R. B.; YVERT, G.; CLINTON, R.; KRUGLYAK, L. Genetic dissection of transcriptional regulation in budding yeast. **Science**, American Association for the Advancement of Science, v. 296, n. 5568, p. 752–755, 2002. Cited 4 times on pages 20, 21, 22, and 53.

CAVALIERI, D.; TOWNSEND, J. P.; HARTL, D. L. Manifold anomalies in gene expression in a vineyard isolate of saccharomyces cerevisiae revealed by dna microarray analysis. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 97, n. 22, p. 12369–12374, 2000. Cited on page 21.

CHELLY, J.; MANDEL, J.-L. Monogenic causes of x-linked mental retardation. **Nature reviews. Genetics**, Nature Publishing Group, v. 2, n. 9, p. 669, 2001. Cited on page 17.

CHESLER, E. J.; LU, L.; SHOU, S.; QU, Y.; GU, J.; WANG, J.; HSU, H. C.; MOUNTZ, J. D.; BALDWIN, N. E.; LANGSTON, M. A. *et al.* Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. **Nature genetics**, Nature Publishing Group, v. 37, n. 3, p. 233, 2005. Cited on page 20.

CONSORTIUM, . G. P. *et al.* A global reference for human genetic variation. **Nature**, Nature Research, v. 526, n. 7571, p. 68–74, 2015. Cited 4 times on pages 27, 29, 30, and 41.

COOPER, T. A.; WAN, L.; DREYFUSS, G. Rna and disease. **Cell**, Elsevier, v. 136, n. 4, p. 777–793, 2009. Cited on page 17.

DEVLIN, B.; ROEDER, K. Genomic control for association studies. **Biometrics**, Wiley Online Library, v. 55, n. 4, p. 997–1004, 1999. Cited on page 30.

DIXON, A. L.; LIANG, L.; MOFFATT, M. F.; CHEN, W.; HEATH, S.; WONG, K. C.; TAYLOR, J.; BURNETT, E.; GUT, I.; FARRALL, M. *et al.* A genome-wide association study of global gene expression. **Nature genetics**, Nature Publishing Group, v. 39, n. 10, p. 1202, 2007. Cited on page 57.

DUDLEY, J.; LAMBERT, R. 100 generations of selection for oil and protein in corn. **Plant breeding reviews**, John Wiley & Sons, Inc., v. 24, p. 79–110, 2010. Cited on page 18.

DUNNING, M.; LYNCH, A.; ELDRIDGE, M. **illuminaHumanv2.db: Illumina HumanWG6v2 annotation data (chip illuminaHumanv2)**. [S.l.]. R package version 1.24.0. Cited on page 27.

ERNST, J.; KELLIS, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. **Nature biotechnology**, Nature Research, v. 28, n. 8, p. 817–825, 2010. Cited on page 48.

FISHER, R. The genetical theory of natural selection. Oxford University Press, 1930. Cited on page 21.

FISHER, R. A. The correlation between relatives on the supposition of mendelian inheritance. **Transactions of the Royal Society of Edinburgh**, Royal Society of Edinburgh Scotland Foundation, v. 52, n. 2, p. 399–433, 1919. Cited on page 17.

GENTLEMAN, R. C.; CAREY, V. J.; BATES, D. M.; others. Bioconductor: Open software development for computational biology and bioinformatics. **Genome Biology**, v. 5, p. R80, 2004. Disponível em: <http://genomebiology.com/2004/5/10/R80>. Cited 3 times on pages 27, 28, and 29.

GERBER, A. P.; HERSCHLAG, D.; BROWN, P. O. Extensive association of functionally and cytotopically related mrnas with puf family rna-binding proteins in yeast. **PLoS biology**, Public Library of Science, v. 2, n. 3, p. e79, 2004. Cited on page 16.

GERSTBERGER, S.; HAFNER, M.; TUSCHL, T. A census of human rna-binding proteins. **Nature Reviews Genetics**, Nature Publishing Group, 2014. Cited on page 16.

GIBBS, R. A.; BELMONT, J. W.; HARDENBOL, P.; WILLIS, T. D.; YU, F.; YANG, H.; CH'ANG, L.-Y.; HUANG, W.; LIU, B.; SHEN, Y. *et al.* The international hapmap project. **Nature**, Nature Publishing Group, v. 426, n. 6968, p. 789–796, 2003. Cited 4 times on pages 27, 29, 30, and 41.

GLISOVIC, T.; BACHORIK, J. L.; YONG, J.; DREYFUSS, G. Rna-binding proteins and post-transcriptional gene regulation. **FEBS letters**, Elsevier, v. 582, n. 14, p. 1977–1986, 2008. Cited on page 16.

GROMKO, M. H. Unpredictability of correlated response to selection: pleiotropy and sampling interact. **Evolution**, Wiley Online Library, v. 49, n. 4, p. 685–693, 1995. Cited 2 times on pages 18 and 42.

HALDANE, J. B. S. **The causes of evolution**. [S.l.]: Princeton University Press, 1932. Cited on page 17.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982. Cited on page 35.

HASAN, A.; COTOBAL, C.; DUNCAN, C. D.; MATA, J. Systematic analysis of the role of rna-binding proteins in the regulation of rna stability. 2014. Cited on page 16.

HAYES, B. J.; VISSCHER, P. M.; GODDARD, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. **Genetics research**, Cambridge University Press, v. 91, n. 1, p. 47–60, 2009. Cited on page 19.

HOULE, D. Genetic covariance of fitness correlates: what genetic correlations are made of and why it matters. **Evolution**, Wiley Online Library, v. 45, n. 3, p. 630–648, 1991. Cited 2 times on pages 18 and 42.

HUGHES, T. R.; MARTON, M. J.; JONES, A. R.; ROBERTS, C. J.; STOUGHTON, R.; ARMOUR, C. D.; BENNETT, H. A.; COFFEY, E.; DAI, H.; HE, Y. D. *et al.* Functional discovery via a compendium of expression profiles. **Cell**, Elsevier, v. 102, n. 1, p. 109–126, 2000. Cited on page 20.

JIN, W.; RILEY, R. M.; WOLFINGER, R. D.; WHITE, K. P.; PASSADOR-GURGEL, G.; GIBSON, G. The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. **Nature genetics**, Nature Publishing Group, v. 29, n. 4, p. 389, 2001. Cited on page 21.

KEENE, J. D. Rna regulons: coordination of post-transcriptional events. **Nature reviews. Genetics**, Nature Publishing Group, v. 8, n. 7, p. 533, 2007. Cited 2 times on pages 16 and 20.

KEMPTHORNE, O. A biometrics invited paper: Logical, epistemological and statistical aspects of nature-nurture data interpretation. **Biometrics**, JSTOR, p. 1–23, 1978. Cited on page 18.

____. Heritability: Uses and abuses. **Genetica**, Springer, v. 99, n. 2, p. 109–112, 1997. Cited on page 18.

KENT, W. J. Blat—the blast-like alignment tool. **Genome research**, Cold Spring Harbor Lab, v. 12, n. 4, p. 656–664, 2002. Cited on page 27.

LAWRENSON, K.; LI, Q.; KAR, S.; SEO, J.-H.; TYRER, J.; SPINDLER, T. J.; LEE, J.; CHEN, Y.; KARST, A.; DRAPKIN, R. *et al.* Cis-eqtl analysis and functional validation of candidate susceptibility genes for high-grade serous ovarian cancer. **Nature communications**, Nature Research, v. 6, 2015. Cited on page 22.

LEE, S. H.; YANG, J.; GODDARD, M. E.; VISSCHER, P. M.; WRAY, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. **Bioinformatics**, Oxford Univ Press, v. 28, n. 19, p. 2540–2542, 2012. Cited on page 42.

LI, C. Population subdivision with respect to multiple alleles. **Ann Hum Genet**, v. 33, n. 1, p. 23–29, 1969. Cited on page 29.

LUKONG, K. E.; CHANG, K.-w.; KHANDJIAN, E. W.; RICHARD, S. Rna-binding proteins in human genetic disease. **Trends in Genetics**, Elsevier, v. 24, n. 8, p. 416–425, 2008. Cited on page 17.

MANIATIS, T.; REED, R. An extensive network of coupling among gene expression machines. **Nature**, Nature Publishing Group, v. 416, n. 6880, p. 499–506, 2002. Cited on page 16.

MITCHELL-OLDS, T. Pleiotropy causes long-term genetic constraints on life-history evolution in brassica rapa. **Evolution**, Wiley Online Library, v. 50, n. 5, p. 1849–1858, 1996. Cited on page 20.

MONTAVON, T.; SOSHNIKOVA, N.; MASCREZ, B.; JOYE, E.; THEVENET, L.; SPLINTER, E.; LAAT, W. de; SPITZ, F.; DUBOULE, D. A regulatory archipelago controls hox genes transcription in digits. **Cell**, Elsevier, v. 147, n. 5, p. 1132–1145, 2011. Cited on page 21.

MOORE, M. J. From birth to death: the complex lives of eukaryotic mrnas. **Science**, American Association for the Advancement of Science, v. 309, n. 5740, p. 1514–1518, 2005. Cited 2 times on pages 16 and 17.

MOORE, M. J.; SCHWARTZFARB, E. M.; SILVER, P. A.; MICHAEL, C. Y. Differential recruitment of the splicing machinery during transcription predicts genome-wide patterns of mrna splicing. **Molecular cell**, Elsevier, v. 24, n. 6, p. 903–915, 2006. Cited on page 16.

MORLEY, M.; MOLONY, C. M.; WEBER, T. M.; DEVLIN, J. L.; EWENS, K. G.; SPIELMAN, R. S.; CHEUNG, V. G. Genetic analysis of genome-wide variation in human gene expression. **Nature**, NIH Public Access, v. 430, n. 7001, p. 743, 2004. Cited 3 times on pages 20, 22, and 53.

ORR, H. A. Adaptation and the cost of complexity. **Evolution**, BioOne, v. 54, n. 1, p. 13–20, 2000. Cited 2 times on pages 21 and 53.

PAABY, A. B.; ROCKMAN, M. V. The many faces of pleiotropy. **Trends in Genetics**, Elsevier, v. 29, n. 2, p. 66–73, 2013. Cited on page 19.

PARÉ-BRUNET, L.; GLUBB, D.; EVANS, P.; BERENGUER-LLERGO, A.; ETHERIDGE, A. S.; SKOL, A. D.; RIENZO, A.; DUAN, S.; GAMAZON, E. R.; INNOCENTI, F. Discovery and functional assessment of gene variants in the vascular endothelial growth factor pathway. **Human mutation**, Wiley Online Library, v. 35, n. 2, p. 227–235, 2014. Cited on page 22.

POWELL, J. E.; HENDERS, A. K.; MCRAE, A. F.; WRIGHT, M. J.; MARTIN, N. G.; DERMITZAKIS, E. T.; MONTGOMERY, G. W.; VISSCHER, P. M. Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. **Genome research**, Cold Spring Harbor Lab, v. 22, n. 3, p. 456–466, 2012. Cited on page 57.

PRICE, A. L.; ZAITLEN, N. A.; REICH, D.; PATTERSON, N. New approaches to population stratification in genome-wide association studies. **Nature Reviews Genetics**, Nature Publishing Group, v. 11, n. 7, p. 459–463, 2010. Cited 2 times on pages 29 and 30.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A.; BENDER, D.; MALLER, J.; SKLAR, P.; BAKKER, P. I. D.; DALY, M. J. *et al.* Plink: a tool set for whole-genome association and population-based linkage analyses. **The American Journal of Human Genetics**, Elsevier, v. 81, n. 3, p. 559–575, 2007. Cited 2 times on pages 33 and 42.

RITCHIE, M. E.; PHIPSON, B.; WU, D.; HU, Y.; LAW, C. W.; SHI, W.; SMYTH, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. **Nucleic Acids Research**, v. 43, p. doi: 10.1093/nar/gkv007, 2015. Cited 2 times on pages 28 and 29.

ROBIN, X.; TURCK, N.; HAINARD, A.; TIBERTI, N.; LISACEK, F.; SANCHEZ, J.-C.; MÜLLER, M. proc: an open-source package for r and s+ to analyze and compare roc curves. **BMC bioinformatics**, BioMed Central, v. 12, n. 1, p. 77, 2011. Cited on page 35.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987. Cited on page 30.

SAX, K. The association of size differences with seed-coat pattern and pigmentation in phaseolus vulgaris. **Genetics**, Genetics Society of America, v. 8, n. 6, p. 552, 1923. Cited on page 18.

SCHADT, E. E.; MOLONY, C.; CHUDIN, E.; HAO, K.; YANG, X.; LUM, P. Y.; KASARSKIS, A.; ZHANG, B.; WANG, S.; SUVER, C. *et al.* Mapping the genetic architecture of gene expression in human liver. **PLoS biology**, Public Library of Science, v. 6, n. 5, p. e107, 2008. Cited on page 22.

SCHADT, E. E.; MONKS, S. A.; DRAKE, T. A.; LUSIS, A. J. *et al.* Genetics of gene expression surveyed in maize, mouse and man. **Nature**, Nature Publishing Group, v. 422, n. 6929, p. 297, 2003. Cited 3 times on pages 22, 53, and 57.

SCHMID, R.; BAUM, P.; ITTRICH, C.; FUNDEL-CLEMENS, K.; HUBER, W.; BRORS, B.; EILS, R.; WEITH, A.; MENNERICH, D.; QUAST, K. Comparison of normalization methods for illumina beadchip humanht-12 v3. **BMC genomics**, BioMed Central Ltd, v. 11, n. 1, p. 349, 2010. Cited on page 29.

SCHMITZ, S.; CHERNY, S. S.; FULKER, D. W. Increase in power through multivariate analyses. **Behavior Genetics**, Springer, v. 28, n. 5, p. 357–363, 1998. Cited 2 times on pages 21 and 53.

SERVIN, B.; STEPHENS, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. **PLoS genetics**, Public Library of Science, v. 3, n. 7, p. e114, 2007. Cited 3 times on pages 39, 42, and 43.

SEXTON, T.; CAVALLI, G. The role of chromosome domains in shaping the functional genome. **Cell**, Elsevier, v. 160, n. 6, p. 1049–1059, 2015. Cited on page 21.

SIVAKUMARAN, S.; AGAKOV, F.; THEODORATOU, E.; PRENDERGAST, J. G.; ZGAGA, L.; MANOLIO, T.; RUDAN, I.; MCKEIGUE, P.; WILSON, J. F.; CAMPBELL, H. Abundant pleiotropy in human complex diseases and traits. **The American Journal of Human Genetics**, Elsevier, v. 89, n. 5, p. 607–618, 2011. Cited on page 20.

SPEED, D.; HEMANI, G.; JOHNSON, M. R.; BALDING, D. J. Improved heritability estimation from genome-wide snps. **The American Journal of Human Genetics**, Elsevier, v. 91, n. 6, p. 1011–1021, 2012. Cited 4 times on pages 30, 31, 34, and 41.

SPITZ, F.; GONZALEZ, F.; DUBOULE, D. A global control region defines a chromosomal regulatory landscape containing the hoxd cluster. **Cell**, Elsevier, v. 113, n. 3, p. 405–417, 2003. Cited on page 21.

STEARNS, F. W. One hundred years of pleiotropy: a retrospective. **Genetics**, Genetics Soc America, v. 186, n. 3, p. 767–773, 2010. Cited on page 19.

STEPHENS, M. A unified framework for association analysis with multiple related phenotypes. **PloS one**, Public Library of Science, v. 8, n. 7, p. e65245, 2013. Cited 5 times on pages 20, 21, 39, 42, and 43.

STRANGER, B. E.; MONTGOMERY, S. B.; DIMAS, A. S.; PARTS, L.; STEGLE, O.; INGLE, C. E.; SEKOWSKA, M.; SMITH, G. D.; EVANS, D.; GUTIERREZ-ARCELUS, M.; PRICE, A.; RAJ, T.; NISBETT, J.; NICA, A. C.; BEAZLEY, C.; DURBIN, R.; DELOUKAS, P.; DERMITZAKIS, E. T. Patterns of <italic>cis</italic> regulatory variation in diverse human populations. **PLoS Genet**, Public Library of Science, v. 8, n. 4, p. e1002639, 04 2012. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pgen.1002639>. Cited on page 27.

TARCHINI, B.; DUBOULE, D. Control of hoxd genes' collinearity during early limb development. **Developmental cell**, Elsevier, v. 10, n. 1, p. 93–103, 2006. Cited on page 21.

TESLOVICH, T. M.; MUSUNURU, K.; SMITH, A. V.; EDMONDSON, A. C.; STYLIANOU, I. M.; KOSEKI, M.; PIRRUCCELLO, J. P.; RIPATTI, S.; CHASMAN, D. I.; WILLER, C. J. *et al.* Biological, clinical, and population relevance of 95 loci for blood lipids. **Nature**, NIH Public Access, v. 466, n. 7307, p. 707, 2010. Cited on page 22.

ULE, J.; JENSEN, K. B.; RUGGIU, M.; MELE, A.; ULE, A.; DARNELL, R. B. Clip identifies nova-regulated rna networks in the brain. **Science**, American Association for the Advancement of Science, v. 302, n. 5648, p. 1212–1215, 2003. Cited on page 16.

VALENTE, E. M.; SILHAVY, J. L.; BRANCATI, F.; BARRANO, G.; KRISHNASWAMI, S. R.; CASTORI, M.; LANCASTER, M. A.; BOLTSHAUSER, E.; BOCCONE, L.; AL-GAZALI, L. *et al.* Mutations in cep290, which encodes a centrosomal protein, cause pleiotropic forms of joubert syndrome. **Nature genetics**, Nature Publishing Group, v. 38, n. 6, p. 623, 2006. Cited on page 20.

WAGNER, G. P. The influence of variation and of developmental constraints on the rate of multivariate phenotypic evolution. **Journal of Evolutionary Biology**, Wiley Online Library, v. 1, n. 1, p. 45–66, 1988. Cited on page 53.

WAGNER, G. P.; ALTENBERG, L. Perspective: complex adaptations and the evolution of evolvability. **Evolution**, Wiley Online Library, v. 50, n. 3, p. 967–976, 1996. Cited on page 21.

WAGNER, G. P.; KENNEY-HUNT, J. P.; PAVLICEV, M.; PECK, J. R.; WAXMAN, D.; CHEVERUD, J. M. Pleiotropic scaling of gene effects and the'cost of complexity'. **Nature**, Nature Publishing Group, v. 452, n. 7186, p. 470, 2008. Cited 2 times on pages 53 and 54.

WAGNER, G. P.; PAVLICEV, M.; CHEVERUD, J. M. The road to modularity. **Nature Reviews Genetics**, Nature Publishing Group, v. 8, n. 12, p. 921–931, 2007. Cited 2 times on pages 21 and 55.

WAGNER, G. P.; ZHANG, J. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. **Nature reviews. Genetics**, Nature Publishing Group, v. 12, n. 3, p. 204, 2011. Cited on page 19.

WANG, K.; LI, M.; HAKONARSON, H. Annovar: functional annotation of genetic variants from high-throughput sequencing data. **Nucleic acids research**, Oxford University Press, v. 38, n. 16, p. e164–e164, 2010. Cited on page 48.

WEN, Y.; GAMAZON, E. R.; BLEIBEL, W. K.; WING, C.; MI, S.; MCILWEE, B. E.; DELANEY, S. M.; DUAN, S.; IM, H. K.; DOLAN, M. E. An eqtl-based method identifies cttn and zmat3 as pemetrexed susceptibility markers. **Human molecular genetics**, Oxford University Press, v. 21, n. 7, p. 1470–1480, 2011. Cited on page 22.

WEST, M. A.; KIM, K.; KLIEBENSTEIN, D. J.; LEEUWEN, H. V.; MICHELMORE, R. W.; DOERGE, R.; CLAIR, D. A. S. Global eqtl mapping reveals the complex genetic architecture of transcript-level variation in arabidopsis. **Genetics**, Genetics Soc America, v. 175, n. 3, p. 1441–1450, 2007. Cited 2 times on pages 22 and 53.

WILLIAMS, G. C. Pleiotropy, natural selection, and the evolution of senescence. **Evolution**, AAAS, v. 11, n. 11, p. 398–411, 1957. Cited on page 20.

WINGREEN, N. S.; MILLER, J.; COX, E. C. Scaling of mutational effects in models for pleiotropy. **Genetics**, Genetics Soc America, v. 164, n. 3, p. 1221–1228, 2003. Cited on page 53.

WRIGHT, S. Systems of mating. i. the biometric relations between parent and offspring. **Genetics**, Genetics Society of America, v. 6, n. 2, p. 111, 1921. Cited on page 17.

YANG, J.; BENYAMIN, B.; MCEVOY, B. P.; GORDON, S.; HENDERS, A. K.; NYHOLT, D. R.; MADDEN, P. A.; HEATH, A. C.; MARTIN, N. G.; MONTGOMERY, G. W. *et al.* Common snps explain a large proportion of the heritability for human height. **Nature genetics**, Nature Research, v. 42, n. 7, p. 565–569, 2010. Cited on page 19.

YANG, J.; LEE, S. H.; GODDARD, M. E.; VISSCHER, P. M. Gcta: a tool for genome-wide complex trait analysis. **The American Journal of Human Genetics**, Elsevier, v. 88, n. 1, p. 76–82, 2011. Cited 7 times on pages 8, 31, 34, 38, 39, 41, and 42.

ZHOU, X.; CARBONETTO, P.; STEPHENS, M. Polygenic modeling with bayesian sparse linear mixed models. **PLoS Genet**, Public Library of Science, v. 9, n. 2, p. e1003264, 02 2013. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pgen.1003264>. Cited on page 31.