upf. Universitat Pompeu Fabra Barcelona

TESI DOCTORAL UPF 2014

Mapping eQTL networks with mixed graphical models

Inma Tur Mongé

Department of Experimental and Health Sciences

Supervisor: Dr. Robert Castelo

Barcelona, 2014

A n'Adrià.

Acknowledgments

La feina de tots aquests anys no hauria estat possible sense l'ajuda i el suport de la gent que ha estat al meu costat, així que moltes gràcies a tots!

En primer lloc, vull donar les gràcies al meu supervisor Robert Castelo. Gràcies per totes les coses que m'has ensenyat, en especial, el valor de la feina ben feta; per la paciència que has tingut amb mi durant aquests anys; per transmetre'm la teva passió per la recerca; per veure el got mig ple quan jo l'he vist buit del tot; per haver-me donat l'oportunitat de viatjar i conéixer llocs i gent tan interessants; per facilitar-me les coses sempre i per recordar-me que tot costa (molt) i tot és lent, però al final acaba sortint.

I want to express my gratitude to Prof. Alberto Roverato for his valuable comments beyond technical aspects and his constant collaboration throughout these years. I am also very grateful to Prof. Steffen Lauritzen for giving me the opportunity to visit him at the Department of Statistics of the University of Oxford. It was a pleasure working in such an enriching environment, having insightful discussions with Sofia Massa and Kayvan Sadeghi, among others, and specially, having the privilege to meet Sir David Cox.

Voldria donar les gràcies també a la Marta Casanellas per haver-me introduït en el món de la genètica des de les aules de la FME i pel seu suport a l'inici del doctorat.

Vull donar les gràcies a totes les persones del PRBB amb qui he compartit despatx, passadís, dinars, retreats i berenars. El bon ambient que hi ha hagut sempre ha estat essencial per combatre els moments més difícils de la tesi. Thank you, Sonja, for being my unique FG-mate. Gràcies a la Mireya, Macarena, Steve, Nico, Eneritz, Jaume i Amadís perquè algun dia demostrarem que menjar muntanyes de pop ajuda a ser millors investigadors. Gracias Alice por ser la única que valora tener una pizarra en casa! En especial, vull donar les gràcies a la Núria per ser un pilar més que fonamental durant tots aquests anys i a la Laia i l'Anna, per tots els dinars, totes les converses i totes les cerveses de Plaça Joanic.

Evidentment, no em puc oblidar dels meus amics matemàtics que tan importants han estat per mi al llarg d'aquests anys i amb els qui he compartit una infinitat de vivències i, fins i tot, un doctorat! En especial, gràcies Arnau, Víctor i Guillem per tots els mojitos a la BdM. Elena, gràcies per estar sempre tan pendent de mi; Cris, per ser la persona més positiva que he conegut mai; Laura, Tere, gràcies pels sopars de nenes i per Peñíscola. Gràcies Dani per compartir els primers anys de tesi; Marga, per ser la millor amfitriona a Paris; Gina, per transmetre'm aquesta contínua energia i Eva, pels km's que hem compartit. També vull donar les gràcies a l'Ignasi, Xous i Míguel per tots els divendres. Sense vosaltres, aquests darrers 3 anys haurien estat molt avorrits! I a la Neus, per arreglar el món des de qualsevol bar de Gràcia.

També em vull enrecordar de sa gent d'Eivissa que fan que no m'oblidi de les meves arrels tot i la distància. Gràcies als boixos: Joan Torres, Toni (gràcies per acollir-me als EEUU!), Joan Roig (gràcies pels dinars a la Pompeu) i David. I gràcies a ses boixes: Alba (te debo una media!), Mariajo, Vanessa i Lara, per les visites a Barcelona, per ses Pujades a sa catedral i perquè sou les millors. I també vull donar les gràcies a na Mònica Domínguez, na Mònica Yern, en Santi Bofill i, en especial, en Santi de la Osa, per despertar-me la curiositat per les matemàtiques. Gràcies Maria Balle, durant tots aquests anys sempre has estat un referent per jo!

Vull agrair als meus pares el suport i els ànims que he rebut constantment des d'Eivissa, perquè si he arribat fins aquí, és gràcies als valors de la constància, l'esforç i la responsabilitat que m'heu transmès des de ben petita. Gràcies Carlos, Miquel i Ana per intentar entendre què he estat fent durant aquests darrers anys i per no deixar de sorprendre'm. Gràcies també per tot a Mari, Paco, Lourdes, Nacho i Sílvia.

Finalment, aquesta tesi sí que no hauria estat el mateix sense el suport incondicional de n'Adrià. Gràcies per la teva paciència infinita d'aquests anys, per ajudar-me sempre en tot i, tot i així, deixar-me seguir compartint aquesta aventura al teu costat.

Abstract

Expression quantitative trait loci (eQTL) mapping constitutes a challenging problem due to the high-dimensional multivariate nature of continuous gene expression traits and discrete genotypes from genetical genomics experiments.

Next to the expression heterogeneity produced by confounding factors and other sources of unwanted variation, indirect effects spread throughout genes as a result of genetic, molecular and environmental perturbations. Disentangling direct from indirect effects while adjusting for unwanted variability should help us moving from current parts list of molecular components to understanding how these components work together in networks of eQTL and gene to gene associations.

There is a large body of statistical methodology to tackle this challenge within the context of linear models for continuous data. However, little has been investigated in using graphical Markov models (GMMs) and conditional independence on mixed continuous and discrete data from genetical genomics data sets, which are powerful tools for the analysis of complex associations.

In this thesis we investigate the use of mixed GMMs to estimate eQTL networks from data. We develop procedures to simulate these models and data from them to gather insight into the propagation of additive effects throughout the network. We derive the parameters for a likelihood ratio exact test that enables use of higher-order conditional independence with mixed GMMs. We exploit this test in the context of limited-order correlations and marginal distributions to obtain estimates of the underlying eQTL network. We show in the context of a yeast genetical genomics data set, that this estimate leads to a sparser network with more direct associations that provide valuable insight into the genetic control of gene expression in yeast. We develop an algorithm for accurate estimation of the genetic effects of eQTLs in the presence of missing data. All algorithms described in this thesis are implemented in the R/Bioconductor package `qpgraph`.

Resum

La cartografia genètica dels trets quantitatius d'expressió (eQTL) esdevé un gran repte degut a la naturalesa multivariant d'alta dimensionalitat dels trets continus d'expressió gènica i els genotips discrets dels experiments de *genetical genomics*.

A més de l'heterogeneïtat de l'expressió produïda pels factors de confusió i altres fonts de variabilitat no desitjada, els efectes indirectes s'estenen per tots els gens com a resultat de perturbacions genètiques, moleculars i ambientals. L'identificació d'efectes directes tot ajustant pels efectes de variabilitat no desitjada, ens hauria de permetre entendre com els diferents components moleculars interaccionen en xarxes d'associacions entre eQTLs i gens.

Per abordar aquest problema, existeixen nombrosos mètodes estadístics en el context dels models lineals per a dades contínues. En canvi, els models gràfics de Markov (GMMs) i la independència condicional, tot i que són eines adients per a l'estudi d'associacions complexes, han estat poc investigades en el context de dades mixtes contínues i discretes de *genetical genomics*.

En aquesta tesi, investiguem l'ús dels GMMs mixtes per a estimar xarxes d'eQTLs. Desenvolupem procediments per a simular GMMs mixtes i simular dades a partir d'aquests models per tal d'investigar la propagació dels efectes additius a través de la xarxa. Derivem els paràmetres d'un test de versemblança exacte que ens permet utilitzar independències condicionals d'ordre gran amb els GMMs mixtes. Utilitzem aquest test en el context de correlacions d'ordre limitat i distribucions marginals per a obtenir estimacions de la xarxa d'eQTLs subjacent. També mostrem que, en el context d'un conjunt de dades *genetical genomics* de llevat, aquesta estimació dóna lloc a una xarxa esparsa amb associacions més directes que ens proporcionen informació rellevant sobre el control genètic de l'expressió dels gens en llevat. Desenvolupem un algoritme per estimar de manera acurada els efectes genètics dels eQTLs a partir de dades *missing*. Tots els algoritmes descrits en aquesta tesi estan implementats en el paquet de R/Bioconductor `qpgraph`.

Preface

The field of molecular biology has experienced an unprecedented progress during the last two decades with the arrival of the genomics era. Technological advances are providing an exponentially increasing amount of high-dimensional biological data sets. This explosion of data has enabled researchers to elaborate the list of molecular components from many biological systems, including human. In the next few years, we will probably witness the completion of those lists for many model organisms and diseases.

One of the next challenges in biology will be to understand how all these molecular components work together to implement the regulatory mechanisms underlying the biological function of organisms. Current assays to genotype and profile the expression of genes already produce high-dimensional data sets that convey many of the problems ahead to address in this challenge.

One such problems is disentangling direct from indirect relationships between genes and genotypes, which requires assessing the association between a pair of these variables while conditioning on the rest of genes and genotypes. The idea of conditioning was already introduced by Francis Galton in the 19th century and has become one of the most fundamental concepts in statistics and data analysis. However, the high-dimensional nature of current biological data sets preclude conditioning simply on the rest of the variables since classical statistical procedures for this purpose do not hold when the number of variables is much larger than the sample size.

This thesis investigates the problem of analyzing high-dimensional genotype and gene expression data by means of conditional independence and mixed graphical models.

Barcelona, February 2014

Contents

Abstract vii
Resum ix
Preface xi
Contents xiv

1 Introduction 1
1.1 Research objectives 8
1.2 Outline of this thesis 8

2 Quantitative genetics 11
2.1 Introduction 11
2.2 Variance components 12
2.3 Heritability 16
2.4 Genetic linkage 20
2.5 QTL mapping 23
2.5.1 Detection of QTL 25
2.5.2 Location of QTL 34
2.5.3 Effects of QTL 35
2.6 Discussion 36

3 Genetics of gene expression 39
3.1 Introduction 39
3.2 Genetical genomics data 42
3.3 eQTL mapping methods 47
3.4 eQTL regulatory mechanisms 51
3.5 Discussion 56

4 Mixed graphical Markov models of eQTL networks 59
4.1 Introduction 59
4.2 Preliminaries 60
4.3 Undirected Gaussian GMMs 62

4.4 Mixed GMMs 64
4.5 Simulation of eQTL networks with mixed GMMs 69
4.6 Flow of genetic additive effects through gene expression . 78
4.7 Estimation of additive effects under selection bias 78
4.8 Concluding remarks 81

5 An exact test to probe higher-order eQTL associations 83
5.1 Introduction 83
5.2 An exact test of conditional independence 84
5.3 Testing higher-order eQTL associations 93
5.4 Concluding remarks 99

6 Learning eQTL networks with mixed GMMs 101
6.1 *q*-Order correlation graphs 101
6.2 Estimation of eQTL networks as qp-graphs 103
6.3 qp-Graphs estimates of eQTL networks are enriched for *cis*-eQTL associations 106
6.4 Higher-order conditioning leads to sparser eQTL networks with more direct *trans*-acting associations 110
6.5 An estimate of the yeast eQTL network 117
6.6 Concluding remarks 123

7 Learning eQTL networks from data with missing observations 125
7.1 Introduction 125
7.2 Complete-case analysis for the non-rejection rate 126
7.3 An EM algorithm to estimate the likelihood ratio statistic 127
7.4 Simulation of genetical genomics data with missing values 131
7.5 Accuracy of likelihood ratio estimates 131
7.6 Accuracy of phenotypic variance estimates 132
7.7 Analysis of the non-rejection rate 133
7.8 Concluding remarks 135

Conclusions and discussion 137

Bibliography 156

1. Introduction

Understanding biological function has been a subject of study and debate throughout human history. First descriptions of biological phenomena can be traced back to the 4th century b.c. by Aristotle. Between centuries 17th and early 20th it become a matter of heated debate whether biological function could be only exerted by living organisms, a postulate known as *vitalism.* One of the most prominent vitalists was Louis Pasteur (1822-1895) who conducted in 1858 a series of experiments that showed that fermentation could not take place without the presence of living cells. Edward Buchner (1860-1917), however, was years later able to produce fermentation without the presence of living yeast cells and, thus, disproved vitalism. His work earned him the Nobel Prize in Chemistry in 1907. In the mid-20th century the field of biology experienced a revolution with the interaction of two seemingly unrelated research areas at that time: biochemistry and genetics (see Figure 1.1).

Biochemistry studies biological function through the analysis of the chemical processes related to molecular components of organisms. In particular, biochemists investigate the structure, function and interactions of biological macromolecules, such as proteins, by examining their chemical reactions. On the other hand, genetics investigates biological function by taking whole organisms with different characteristics, crossing them and studying how the different traits are inherited to offspring.

The foundations of modern genetics were established in the mid 19th century by Gregor Mendel (1822-1884) although the importance of his work was not fully understood and accepted until the beginning of the 20th century. In 1865 and 1866, Mendel published the results of his breeding experiments with pea plants. In these publications, he showed that the inheritance patterns of some traits whose values can be classified in different categories, the so-called *qualitative traits*, obeyed simple statistical rules. In particular, he established that each individual has two copies of the same unit of inheritance (gene), and that one of them is randomly transmitted to the offspring so that the offspring receive a copy from each parent. Moreover, he stated that the gene responsible for the variation of a trait is segregated independently of the gene responsible for any other trait. These rules were called the Law of Segregation and the Law of Independent Assortment, respectively, and are known as Mendel's Laws of Inheritance. However, at that time, Mendel's work was ignored. It was in 1900, when Hugo de Vries (1848-1935), Carl Correns (1864-1935), and Erik Tschermak (1871-1962) independently realized the importance of Mendel's Laws of Inheritance after carrying out experiments that supported his theory. Indeed, Mendel's work was also a proof that application of quantitative methods to genetics could be highly useful.

Shortly after, William Bateson (1861-1926), who introduced the term *genetics* in 1906, Thomas Morgan (1866-1945) and Alfred Sturtevant (1891-1970) continued developing the theory of Mendel and made important contributions in the field of genetics. In particular, Bateson and Morgan described the phenomenon of *genetic linkage*, which occurs when the segregation of a gene during the meiosis is not independent of the segregation of another gene that is located proximal to it on a chromosome (Bateson et al., 1905; Morgan, 1911). Alfred Sturtevant studied the linear disposition of genes on chromosomes and built the first *genetic map* (Sturtevant, 1913). This allowed him to introduce a technique based on the comparison of the inheritance pattern of a trait with the inheritance pattern of chromosomal regions, known as *genetic mapping*.

In the same period, Francis Galton (1822-1911) started to apply statistical methods to study human differences. Galton was also the first to introduce the concepts of correlation and conditioning already in 1880s in order to establish a degree of resemblance between relatives (Galton, 1889). He compared physical attributes of parents and children and described the fact that offspring values always tend to regress to the mean value for the population. His successor, Karl Pearson (1857-1936) introduced many statistical methods which are commonly used today in many fields, including biology, epidemiology and medicine. For example, Pearson developed the correlation coefficient from the idea already introduced by Galton, defined the p-value in the context of the χ^2 test and developed Principal Component Analysis (PCA), which are among the most powerful and used statistical methods in many areas of research, including biology.

The statistical techniques that Galton and Pearson invented constituted the basis of the biometric school, that was distinguished by the use of statistical models to study heredity and evolution. However, in contrast to Mendelians, biometricians were interested in the traits that presented a continuous variation within populations, called *quantitative traits*. They pointed out that these were the important characters in evolution. Indeed, Pearson led a movement against Mendelians arguing that Mendel's laws could not be applied to quantitative traits since such traits did not show a simple pattern of inheritance.

The disagreement between Mendelians and biometricians was finally resolved by Ronald A. Fisher (1890-1962) in 1918 (Fisher, 1918). In this paper, he showed how variation of continuous traits could be the result of the action of many discrete genetic loci each of which is inherited according to Mendel's laws. This paper is considered the funding work of the field of *quantitative genetics*, that is, the study of the inheritance of quantitative traits.

Sewall Wright (1889-1988) and John B.S. Haldane (1892-1964) were also major contributors to this field. Wright worked on quantitative genetics of animal and plant breeding (Wright, 1920) but his major contribution was the development of the statistical method of path analysis (Wright, 1934), one of the first methods using a statistical model with a graphical interpretation, which is used to describe the directed dependencies among a set of variables (e.g., to determine the interrelations among the factors which determine the values of traits such as the weight of guinea pigs). On the other hand, Haldane was focused on human genetics and, particularly, on the creation of human genetic linkage maps (Haldane and Bell, 1937).

Indeed, right after Bateson, Morgan and Sturtevant's contributions, geneticists started to identify and locate on the chromosomes the specific genes responsible for phenotypes whose variation was affected by single genes with Mendelian segregation (e.g., Sax, 1923). They did not succeed, however, when they tried to identify the genes for more complex (non-Mendelian) phenotypes or diseases. The major advances in genetics of the following decades showed how hard and challenging this problem was.

For the first half of the 20th century, scientists were studying biochemistry and genetics as two separated but complementary research areas to study biological function. However, in the middle of the 20th century it was discovered that there was an intimate link between genes and proteins and the way in which biology was understood changed abruptly.

In 1953, James D. Watson (1928-) and Francis Crick (1916-2004) discovered the structure of DNA (Watson and Crick, 1953). Thanks to this discovery it was finally found the connection between biochemistry and genetics, which constituted the starting point of a new research field, called *molecular biology* (see Figure 1.1). Concretely, it was determined that the DNA had a double-helix structure, with two strands of nucleotides that contain the genetic information. The process by which the structure of DNA was used as a template to copy the genetic information from parents to offspring began to be understood. In 1958, Francis Crick proposed an explanation of the flow of genetic information contained in a gene within a cell (Crick, 1958). Briefly, the DNA sequence of a gene is transcribed into an RNA molecule and this one is translated into a protein. It was called the *central dogma of molecular biology*. Yet, the view of the central dogma at that time was essentially descriptive and theoretical.

During the following two decades, genetics and biochemistry were combined to develop model systems of living organisms using bacteria, fruit flies, yeast and mice. Using these model systems, basic mechanisms of gene expression regulation (Jacob and Monod, 1961) started to be understood. One of the most important discoveries was the *recombinant DNA* technology (Jackson et al., 1972). Recombinant DNA are molecules composed of DNA sequences from different organisms (chimeric DNA) used to identify single genes, reproduce or edit them in order to investigate their function (Figure 1.1).

Moreover, some years later, scientists were able for the first time to read the specific nucleotide sequence of a DNA fragment. This was possible due to a very important technological advance led by Frederick Sanger (1918-2013). Sanger and his colleagues introduced in 1977 a chain-termination method which constitutes the basis of the techniques for reading the sequence of DNA (Sanger et al., 1977). The Sanger method was the most widely used technology for DNA sequencing during the rest of the century. Further, another crucial contribution was the Polymerase Chain Reaction (PCR) technology developed by Kary Banks Mullis (1944-) in 1983. PCR is a biochemical technology that provides a quick way to isolate a specific fragment of DNA and copy (amplify) it many times, generating thousands to millions of copies of a particular DNA sequence. This is often used to detect the presence of specific DNA sequences for many applications, including DNA cloning for sequencing, functional analysis of genes and diagnosis of genetic diseases, among others.

At that time, researchers were studying genes one at a time and they were able to sequence single genes responsible for certain human diseases. However, in order to find the genes responsible for any human disease, scientists started to think about the idea to sequence the complete human genome. By the 1980s, the improvement of sequencing technologies increased the scope and speed of DNA sequencing and allowed researchers the sequencing of the complete genome of the human mitochondrion and chloroplast and also the complete genome of certain bacterias and model organisms such as yeast.

Finally, in the 1990s, sequencing the entire DNA of the human genome was started by an internationally coordinated project, the Human Genome Project (HGP). A first draft of the human genome was completed in 2001 (Venter et al., 2001; International Human Genome Sequencing Consortium, 2001).

The determination of the entire DNA sequence of organisms offered a systematic look at the entire genome and enabled the analysis of the function and structure of all the genes simultaneously. The analysis of the entire genome at once was determinant in the study of the genes responsible for complex traits and diseases. Thus, studies conducted by molecular geneticists focused on single genes gave way to genome-wide analysis, thereby giving a completely new perspective of genetics. This marked the beginning of a new field in genetics that was called *genomics* (see Figure 1.1), a term that was already coined in 1986 by Thomas Roderick.

In the same decade, thanks to the development of the array-based technology (DNA microarrays), millions of polymorphic DNA regions, called single nucleotide polymorphisms (SNPs), were discovered and genotyped in parallel throughout samples. This gave rise to the development of statistical methods to systematically identify the regions of the genome affecting the variation of quantitative traits, a problem called *quantitative trait locus (QTL) mapping* (Falconer and Mackay, 1996). The use of DNA microarrays led also to an explosion of genome-wide association studies (GWAS) of complex traits which look for associations between the genetic markers and human phenotypes and diseases in case-control studies (The Wellcome Trust Case Control Consortium, 2007).

Microarray technology was not only used at the level of DNA but also at the level of RNA. Millions of short oligonucleotides complementary to cDNA sequences of known genes enabled simultaneously measuring the abundance of RNA molecules, that is, the expression level (Schena et al., 1995) for thousands of genes. These experiments allow the visualization of the pattern of up and down regulation of genes which, in turn, provides a signature of the assayed cellular state. It enabled, for instance, the discovery of new breast cancer subtypes (Perou et al., 2000).

High-throughput microarray technology produced the first genetics and molecular high-dimensional data sets, that is, data sets in which the number of recorded variables is much greater than the number of available samples. The successes of these high-throughput analyses of multiple genes in addition to the power of genotyping led in the early 21st century to merge both strategies and introduce genetical genomics studies. Ritsert C. Jansen and Jan-Peter Nap proposed that the expression level of a gene could be considered as a quantitative trait (Jansen and Nap, 2001) whose variability could be partly due to genetics. Indeed, Brem et al. (2002) showed that gene expression was an heritable trait in yeast. Genetic loci affecting gene expression are called eQTLs and finding them was called the *eQTL mapping problem*. In fact, integrating and analyzing different sources of molcular data (such as SNPs, gene expression profiles or even the level of proteins or metabolite products), the so-called *integrative genomics*, has become a fruitful strategy to get a deeper understanding of the flow of information from DNA to organismal phenotypes.

Throughout the last years, the improvement of DNA sequencing technologies is dramatically lowering the cost of sequencing and increasing the speed and accuracy of molecular profiling assays. These methods, the so-called high-throughput sequencing technologies (HTS) or next-generation sequencing technologies (NGS), are intended to parallelize the sequencing process, producing millions of sequencing reactions simultaneously. These sequences can be of DNA or RNA, and can cover all the genome (whole-genome sequencing) or some parts (exome and target sequencing), as well as the complete transcriptome (RNA sequencing).

As a consequence, these technologies are producing high-dimensional data sets that represent an underlying complex network of relationships between the recorded molecular variables. These high-dimensional technologies present an opportunity to address the problem of identifying the genetic basis of complex traits and to investigate their regulatory mechanisms. Therefore, the current challenge for researchers is to develop appropriate quantitative methods intended to analyze these high-dimensional biological data sets and to properly infer information that help biologists to narrow the gap between DNA and organismal phenotypes and human diseases.

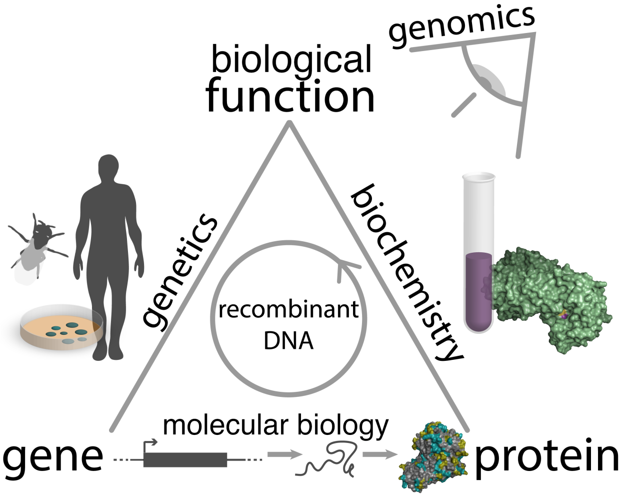

Figure 1.1: Biology triangle. This diagram shows the connection between biochemistry, genetics, molecular biology and genomics. Biological function can be understood through the study of the inheritance of genes (genetics) or through the study of molecular components such as proteins (biochemistry). Genes and proteins are connected through the central dogma of molecular biology. The discovery of recombinant DNA made operational this diagram: if we have a gene, we can read out the protein by DNA sequencing and if we have an antibody against the protein, we might find the gene that encodes such protein. We are also able to knock out a gene or isolate a protein from the rest of the organism and see what function it subserves. Finally, genomics gives us a global view of this diagram by looking at the entire genome at once. Adapted from *https://courses.edx.org/static/content-mit-7012x~2013_Spring/images/ResourceBox/sera/biology_triangle.png*.

1.1 Research objectives

In this thesis we tackle the challenging problem of eQTL mapping. The main purpose of this research is to develop a statistical methodology that enables the dissection of non-spurious associations between genetic variants and gene expression profiles by exploiting the high-dimensional nature of genetical genomics data. To this end, we focus on the theory of mixed graphical Markov models (Lauritzen and Wermuth, 1989). For that, this work aims to:

1. Implement a procedure to simulate classical graphical Markov models for mixed discrete (genotypes) and continuous (gene expression profiles) variables and data from them such that they represent features of a network of eQTL associations, an eQTL network, from experimental crosses of model organisms.

2. Develop a measure to test the association between a genetic variant (discrete variable) and a Gaussian distributed phenotype or between two Gaussian distributed phenotypes such that it enables the adjustment for the expression of all the other genes, other phenotypic variables and possible confounding effects.

3. Adapt this measure to test the association between variables from data sets with missing genotype values.

4. Provide an strategy to learn a mixed graphical model as an eQTL network estimate from data where the number of variables (p) is much larger than the number of observations (n).

5. Study the estimated eQTL network of yeast in order to get a deeper understanding of the regulatory architecture of gene expression profiling.

1.2 Outline of this thesis

The work in this thesis is presented in the following way:

- We start by introducing in detail the field of quantitative genetics in **Chapter 2**. Further, we provide an overview of the QTL mapping problem whose objective is to locate the genomic regions that affect quantitative phenotypes. We pay special attention to the developed statistical methodologies used to conduct these studies.

- In **Chapter 3** we provide an exhaustive explanation of the case in which the quantitative traits used in QTL mapping are gene expression profiles, the so-called eQTL mapping studies. We discuss their main challenges and drawbacks of current methodologies to address this problem.

- The theoretical framework employed to develop our eQTL mapping approach, the theory of mixed graphical Markov models (Lauritzen and Wermuth, 1989), is introduced in **Chapter 4**. We also propose and implement an algorithm to simulate eQTL network models from experimental crosses and data sets from them.

- In the next chapter (**Chapter 5**), we present an exact likelihood ratio test to test the conditional independence between a genetic variant and a gene expression profile or between gene expression profiles. We assess its performance against the χ^2 asymptotic approach used in QTL mapping through some experiments with synthetic and real data from yeast.

- Next, in **Chapter 6** we provide a multivariate approach to estimate eQTL networks by using the previously presented exact conditional independence test. By applying this methodology to a real genetical genomics data set from yeast (Brem and Kruglyak, 2005), we obtain a sparser eQTL network with more direct and more functionally related *trans*-associations in comparison to the eQTL network obtained with a classical univariate approach.

- Finally, in **Chapter 7** we provide an approach to address the problem of eQTL mapping from data with missing genotypes and verify its suitability with synthetic data. We also provide an expectation-maximization algorithm to obtain accurate estimates of the exact likelihood ratio statistic.

2. Quantitative genetics

2.1 Introduction

The experiments conducted by Gregor Johan Mendel with pea plants, published in 1865 and 1866, laid the foundations to understand how certain traits that are caused by single units of inheritance, called genes, are passed from parents to offspring. Since the rediscovery of Mendel's laws at the beginning of the 20th century, genes responsible for this type of traits, known as Mendelian traits, started to be mapped to chromosomal locations in experimental organisms (Sturtevant, 1913). In humans, many diseases, among other phenotypes, have been catalogued as Mendelian traits and their inheritance pattern (recessive, dominant, X-linked, etc) has been extensively studied since the 1980's. By that time, geneticists started to use naturally occurring DNA variation as genetic markers to trace inheritance in families and find the gene causing the variation in the trait. The first studies were performed in yeast (Petes and Botstein, 1977) and then in human families (Botstein et al., 1980). The Hungtington's disease (Gusella et al., 1983; MacDonald et al., 1993) or cystic fibrosis (Kerem et al., 1989) are just two examples of Mendelian diseases for which the causal gene has been found.

In contrast, most phenotypic characters in animals and plants including common genetic disorders (Plomin et al., 2009) are complex in the sense that they are affected by a combination of many gene loci and other non-genetic factors such as environment (Falconer and Mackay, 1996). Some of these phenotypes show a continuous variation in the population (for instance, the height, the body weight or the blood pressure) and they are known as quantitative traits. Ronald A. Fisher showed that quantitative traits could be explained by Mendelian inheritance if multiple genes affect the trait (Fisher, 1918). That paper can be seen as an extension of the simple Mendelian inheritance since it studies the combination of simultaneous segregation of genes at many loci and the environment in which a population of individuals are expressed. It constitutes the foundation of the field of quantitative genetics, the area of genetics that studies the inheritance of quantitative traits.

Understanding the variation of DNA and the environmental factors that contribute to the variation of phenotype, as well as the inheritance mode of this variation, is of great importance in evolution, in medicine (to predict disease risk and to develop personalized treatments) and in animal and plant breeding (to improve the selection of economically important traits). In particular, a region of the genome that affects the variation of quantitative phenotypes is called a quantitative trait locus (QTL). The identification of QTLs has been possible through the development of specific statistical techniques; such analyses are called QTL mapping studies.

In this chapter we review the very basic statistics underlying the genetics of quantitative traits (Section 2.2). We also explain how to calculate the degree of inheritance of these traits (Section 2.3) and how genes associated to quantitative traits segregate from parents to offspring (Section 2.4). Finally, the analysis of QTL mapping is addressed in the last section.

2.2 Variance components

The first law of Mendel, called the Law of Segregation, states that each individual (of a diploid organism) has two copies of the same gene for any trait and that each parent transmits a randomly selected copy of the gene to its offspring. The alternative versions of a gene are called alleles and the genotype refers to the two alleles present in an individual for a certain gene.

In the simplest scenario where the variation of a trait is caused by only one gene with two possible alleles of equal frequency, the combination of alleles results in three distinct genotypes (see Figure 2.1). Depending on the relationship between the alleles, the genotypes produce either three different phenotype values (if the effect of the alleles is additive) or two (when the alleles show complete dominance).

Then, how can a continuous distribution of phenotypic values arise? As explained above, a quantitative trait is affected by the combination of many genes so that the number of possible genotypes is greater and, consequently, the number of trait values increases. Furthermore, most of the quantitative traits are affected by the environment or other non-

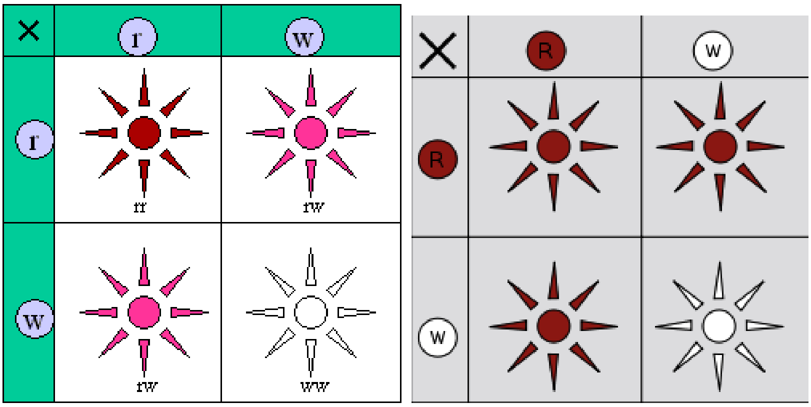

Figure 2.1: Phenotype classes of a Mendelian trait. Left: Incomplete dominance is shown when the three possible combinations of the alleles results in three distinct phenotype values, i.e., the alleles contribute additively to the trait. This is the behaviour of additive or codominant models. Right: a dominant phenotype is the one that is shown easily, the presence of R allele is enough to express the red colour; the white color is a recessive trait. In this case, the model is said to be dominant. *Adapted from http://en.wikipedia.org/wiki/Mendelian_inheritance.*

genetic circumstances which may blur the genetical differences between the phenotypic classes giving rise to continuous, characterized as normal distributions (Figure 2.2). Hence, the phenotypic value P observed in an individual can be interpreted as the sum of the genetic effects G and environmental effects E due to all the other non-genetic factors (Falconer and Mackay, 1996, pg. 108):

$$P = G + E\,.$$

The genetic value G can be further divided into three other components (Falconer and Mackay, 1996, pg. 119):

$$G = A + D + I\,. \tag{2.1}$$

The first component, A, corresponds to the additive effects and it arises from the sum of the effects attributable to separate loci. The second component, D, corresponds to the dominance effect among the alleles at the same locus, and therefore, D represents the interaction between alleles. Finally, the third component corresponds to the interaction between different loci, called epistasis. In the case of quantitative traits, if there is no epistasis, we say that the loci act additively on the phenotype (Figure 2.3).

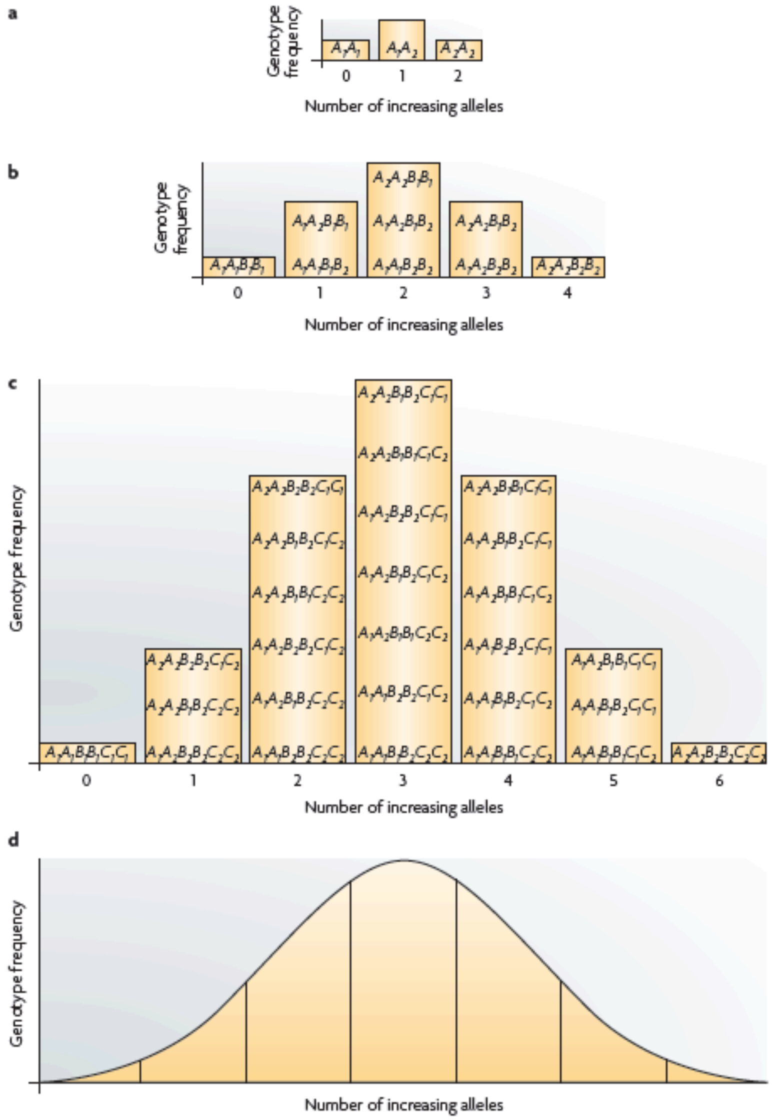

Figure 2.2: Continuous distribution of a quantitative trait assuming equal and additive effects. Histograms of genotype frequencies showing the number of phenotypic values as a function of alleles from one (a), two (b) and three (c) loci. As the number of loci affecting a trait increases, in addition to the effect of environment, the phenotype variation gets closer to a normal distribution (d). *Taken from Plomin et al. (2009).*

In contrast to qualitative trait values, which can be divided into classes, phenotypic values of quantitative traits have to be measured. To this end, two statistical parameters are used: the mean, which calculates the average value of a quantitative trait in a population, and the variance.

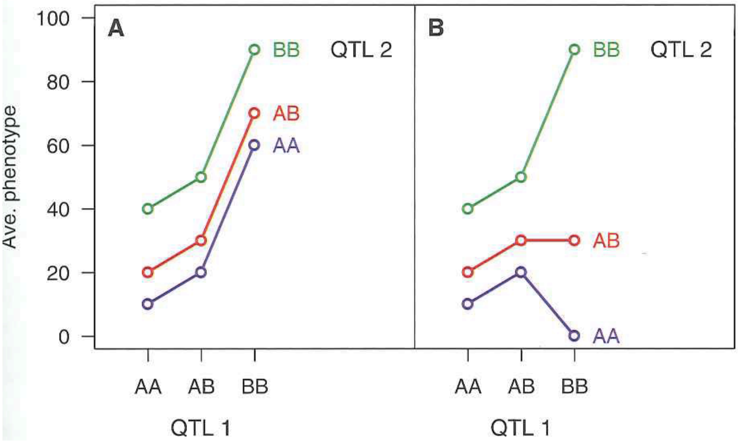

Figure 2.3: Additive and epistatic effects of two loci, called QTLs, affecting a quantitative trait. A) The average phenotype values G according to the genotype of two QTLs are plotted. The pattern of the effect at one locus is the same regardless of the effect of the other locus; these QTLs show additive effects. B) Here, the effect of one locus depends on the genotype of the other locus, the two QTLs interact and they show epistatic effects. *Taken from Broman and Sen (2009).*

One of the challenging points in the study of the genetics of quantitative traits is to distinguish among the different components of the variability of a trait in a population. To do this, we use the variance. In particular, the partition of the variance into components attributable to different causes, allows us to estimate the amount of phenotypic variability due to genetic effects, which is crucial when measuring the degree of inheritance of the trait (Section 2.3).

The total phenotypic variance V_P is partitioned into the variance that is associated to the genetic effects V_G, and the variance that is associated with environmental effects V_E. Genetics and environment may interact, which involves an additional component of variance, V_{GE}:

$$V_P = V_G + V_E + V_{GE} .$$

Following Equation (2.1), genetic variance can also be divided into the additive component V_A, the dominance component V_D and variance due to epistasis, V_I:

$$V_G = V_A + V_D + V_I \,.$$

2.3 Heritability

The first attempt to measure the proportion of the variability of a quantitative trait that is inherited from parents to offspring was proposed by Sir Francis Galton. At the end of the 19th century, Galton showed that the height of 928 individuals could be explained as a function of their parents' mean height (Galton, 1889) as it is shown in Figure 2.4. In particular, if offspring values are regressed against the average value of the parents, the slope of the regression line determines the degree of how heritable a phenotype is with reference to the resemblance of offspring and parents (Falconer and Mackay, 1996). This degree is known as the heritability of a trait. However, the concept of heritability was formally introduced by Sewall Wright and Ronald Fisher (Fisher, 1918; Wright, 1920). It is a key concept in the study of the response to selection in evolutionary biology and agriculture, and to the prediction of disease risk in medicine (Visscher et al., 2008; Tenesa and Haley, 2013). Heritability is a dimensionless parameter that is estimated with reference to a specific population and environment and, in some cases, even to the individual's lifetime. For this reason, it can be compared across populations and within them in order to improve, for instance, plant and animal breeding programmes.

For plant and animals breeders, or to investigate the genetical factors of a disease, it is interesting to know the fraction of the phenotypic variability of a trait that is due to genetic variance. In this case, the broad-sense heritability,

$$H^2 = \frac{V_G}{V_P} \,,$$

gives us this fraction, also known as the degree of genetic determination.

However, in order to predict the phenotype of an offspring from a cross of two parents, one is interested in the estimation of the relative contribution that the parents will make to their progeny. Because parents only transmit one copy of each gene to their offspring, these (other than twins) only share one or none copies that are identical by descent, that is, identical copies of the same ancestral allele. Hence, effects that are based on sharing two copies, such as dominance or

FIG.10.

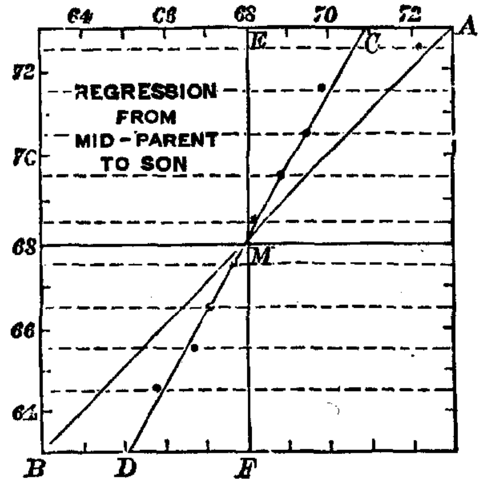

Figure 2.4: Regression towards the mean. Francis Galton represented in this graph the relationship between offspring height (y-axis) as a function of their parents' mean height (x-axis). The line connecting the points C and D is the regression line of these two data sets. *Taken from Galton (1889).*

other non-additive genetic effects, do not contribute to the phenotypic resemblance between relatives. Therefore, the proportion of the total phenotypic variation that is due to heritable effects only involves the additive component of the variance in the numerator:

$$h^2 = \frac{V_A}{V_P}\,.$$

This is called the narrow-sense heritability (or heritability) and it is also known as the degree of resemblance between relatives.

Heritability and the estimation of offspring phenotype

In balanced designs, the narrow-sense heritability has been calculated as the regression of offspring on parental phenotypes (Galton, 1889). Conversely, if we have determined the narrow-sense heritability of a trait and we know several population values we can estimate the phenotypic value of an offspring T_0 as:

$$T_0 = T + h^2(T^* - T)\,,$$

where T is the population mean and T^* is the midparent value. Estimation of h^2 can also rely on the phenotypic correlations of full or half sibs and on the difference in the correlation between monozygotic (MZ) and dizygotic (DZ) twin pairs (Falconer and Mackay, 1996). In some cases, these estimations may be biased due to the presence of epistasis and shared environment (Zaitlen et al., 2013).

Linear mixed-effects models are, in general, more appropriate to estimate the additive genetic variance when the design is unbalanced or the individuals have a mixture of relationships. These models are based on the estimation of identity-by-descent (IBD) covariance matrices (Powell et al., 2010) which specify the genetic relatedness of individuals in the data set. Traditionally, IBD matrices were calculated with respect to a base population from a known pedigree. However, these matrices are now estimated from the whole-genome SNP data (Visscher et al., 2006) so that pedigree information is not necessary. In principle, these matrices (also known as realized relationship matrices) make it possible to obtain more accurate and reliable estimates for the narrow-sense heritability. In Visscher et al. (2006), the heritability of height was estimated from a data set of full sibs and, in this case, the result ($h^2 = 0.8$) was consistent with previous estimations obtained from the difference in the correlation between MZ and DZ in twin pairs. Recently, a population consisting of closely and distantly related individuals was used to reduce the bias due to non-additive variance components to estimate the heritability of 11 human quantitative and 12 dichotomous traits (Zaitlen et al., 2013).

Heritability and the response to selection

Narrow-sense heritability contributes to predict how a population will respond when artificial or natural selection is applied. The relationship between h^2 and the response to selection is given by

$$R = h^2 S \,,$$

where R is the selection response and S is the selection coefficient. The selection response R is given by the difference between the population mean before selection and the mean of the offspring of selected parents after one generation of selection. The selection coefficient S corresponds to the difference between the unselected population mean and the mean of the selected parents. When artificial selection is applied, h^2 can be of crucial importance in the design of breeding schemes in agriculture.

Missing heritability

It has been shown (Manolio et al., 2009) that the identified genetic variants affecting complex traits do not explain all the heritability of the trait. This means that the proportion of the phenotypic variance attributable to the additive effects of the identified variants, denoted by η^2, is smaller than the heritability due to all the loci affecting the trait, h^2. The fraction of heritability that is not explained by the identified loci, $1 - \eta^2/h^2$, is commonly known as the missing heritability (Manolio et al., 2009).

Many explanations have been proposed to disentangle the causes of the missing heritability of quantitative traits. One of the possible causes is the underestimation of the genetic additive variation V_A of η^2. For instance, it may be that many loci of small effects have not yet been identified (Yang et al., 2010) or that rare variants contributing to the genetic variation of the trait are not captured by current genotyping platforms since they are not sufficiently frequent in the population (Pritchard, 2001). On the other hand, Zuk et al. (2012) state that a substantial portion of the missing heritability could arise from the overestimation of the denominator, h^2. This occurs since the estimation of h^2 implicitly assumes that the trait is explained by a model that does not involve genetic interactions among loci (epistasis) while, according to these authors, there are no reasons that justify this assumption.

Recently, Bloom et al. (2013) conducted a study in which they dissect the genetic architecture of 46 quantitative traits in a population of 1,008 yeast segregants. This population results from an experimental cross between a wild and a laboratory strain so that the dominance and the environmental effects can be controlled. In this study, the authors find that many loci of small effects affect the variation of the quantitative traits. More importantly, the identified loci explain a very high proportion of heritability and authors verify that the main reason for the remaining missing heritability is its insufficient sample size. Therefore, under these experimental conditions, the authors conclude that the heritability for the traits under study is not missing. Although the complexity of human traits is greater and, consequently, the causes for the missing heritability may be very diverse, the work of Bloom et al. (2013) constitutes an important reference for further design studies.

2.4 Genetic linkage

The second law of Mendel states that genes responsible for different traits are passed independently of one another from parents to offspring. This is known as the Law of Independent Assortment. It means that, for a pair of alleles responsible for a certain trait, the allele selected to be passed to the offspring is independent of the selected allele that is responsible for any other trait. While this is true for genes that are located on different chromosomes, this is not always the case when both genes are on the same chromosome.

During meiosis, the process by which the reproductive cells (gametes) are formed, recombination events take place. A recombination event (Morgan, 1911) implies an odd number of crossovers, that is, the exchange of genetic material between parental chromosomes (see Figure 2.5). Recombination events create new combinations of alleles in the gametes. Thus, if a recombination event takes place between two genes on the same chromosome, they may be separated onto different chromatids (a chromatid is one copy of a duplicated chromosome) and they are not segregated together.

However, if a recombination event does not take place between two genes, they are inherited together. If this is the case, these genes are said to be genetically linked. Bateson et al. (1905) provided evidence of the fact that genes located close to each other on a chromosome tend to be inherited together.

Genetic distance

Alfred Sturtevant, a student of Thomas H. Morgan, observed in a series of experiments with fruit flies that the amount of crossovers between genes differs and that the probability of ocurring a recombinant event increases with the physical distance of the genome. Hence, this led to the idea that the distance between two genes could be measured in terms of the crossover frequency between them. Concretely, the genetic distance between two genes on a chromosome is defined by their recombination fraction which is the proportion of offspring that have genetic material of differing parental origins at the two loci (recombinant offspring). Because the number of recombinant offspring cannot exceed the number of parental ones, the proportion of recombinant offspring never exceeds

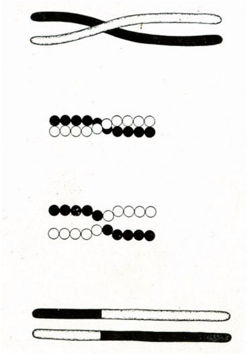

Figure 2.5: Representation of a crossover. A crossover is the exchange of genetic material between parental chromosomes. *Taken from Morgan (1916).*

0.5. The genetic distance between two loci in a chromosome is measured in terms of centimorgans (cM) where 1 cM stands for a recombination fraction of 0.01, that is, one gamete out of 100 is recombinant for those two loci.

Genetic markers

A genetic marker can be described as a variation in the DNA sequence whose location on a chromosome is known. At the time when the structure of DNA was yet unknown, finding genes, known as gene mapping, was performed in most organisms by using traditional genetic markers, for instance, genes encoding easily observable phenotypes such as blood type (Landsteiner, 1901). The insufficient number of this type of traits in several organisms limited the gene mapping investigations and motivated, throughout the 20th century, the development of other genetic markers.

Ideally, marker loci should be highly polymorphic and abundant. A genetic marker may be a long DNA sequence, such as a simple sequence repeat (SSR or microsatellites) or a restriction fragment length polymorphism (RFLP), or a short one, like a single nucleotide polymorphism (SNP). SNPs are DNA sequence variations occurring when a single nucleotide differs in the genome between the members of the population. They are the most widely used genetic markers for understanding the genetics of quantitative phenotypes due to the rapid development and cost reduction of high-throughput technologies used to type them.

Genetic maps

A representation of the sequence and position of genes or genetic markers in terms of their genetic distance is called a linkage map or a genetic map (Figure 2.6). Alfred Sturtevant (Sturtevant, 1913) was the first to build a genetic map of a chromosome of *Drosophila Melanogaster*.

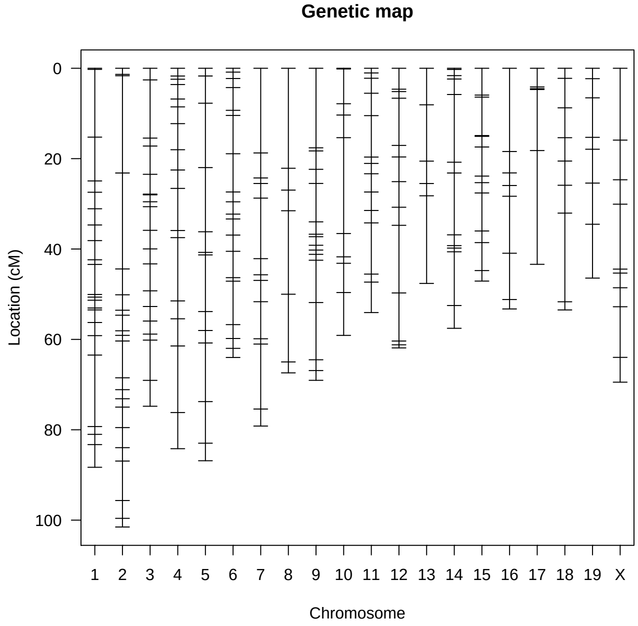

Figure 2.6: Genetic map of mouse. This genetic map displays the position of the genetic markers typed in the data from Leduc et al. (2012).

To build a genetic map we have to estimate the recombination fraction between each pair of loci according to a model for the recombination process. One option is to assume that crossover locations occur at random (no crossover interference model). This implies that recombination events in disjoint intervals are independent and that genotypes along a chromosome form a Markov chain. However, in most organisms, crossovers are not randomly occurring. In this case, we consider a positive crossover interference model.

Because the recombination fraction between two loci varies between 0 and 0.5, it is not an additive measure across a genetic map, unlike the genetic distance. This means that if loci A, B and C are located at 5 cM intervals on a map, locus C is 10 cM from locus A, but the recombination fraction between A and C is not 10%. The mathematical relationship between recombination fractions and genetic distances is described by the map functions (Speed, 2005). There are different map functions according to the recombination process. For instance, the Haldane map function assumes no interference (Haldane and Bell, 1937), Kosambi map function assumes the level of interference in humans, and the Carter-Falconer map function assumes the level of interference in mice.

2.5 QTL mapping

One of the main interests of quantitative geneticists is to study which genes are affecting quantitative traits, where these genes are located and how large is the magnitude of their effects. This is accomplished by looking for associations between the variation of a polymorphic region of the genome and the variation of a quantitative trait of interest through the linkage of this region to marker loci for which their genotype is known. This region is called quantitative trait locus (QTL) and the process of identifying QTLs is called quantitative trait locus mapping (QTL mapping).

These studies have been carried out since the beginning of the 20th century (Sax, 1923) although the revolution of QTL mapping studies started around the 80's. The advances in the technology enabled the implementation of cheaper and more efficient genotyping methods, such as the development of microarrays, favoring the discovery of abundant polymorphic markers. These new genotyping methods in addition to the development of statistical methodologies have led to a proliferation of QTL mapping studies.

Linkage mapping and association studies

There are two main types of QTL mapping analysis according to the presence (linkage analysis) or absence (association studies) of relationships between the analyzed individuals (Lander and Schork, 1994; Altshuler et al., 2008; Mackay et al., 2009).

In linkage analysis studies, the relationships between the individuals are known. As a consequence, genetic markers that are linked to the causal variants segregate together unless a recombination event occurs. By contrast, association mapping studies are conducted in a set of unrelated individuals. In natural populations, many recombination events can occur throughout many generations, and therefore, only those markers that are strongly linked to the causal loci will be associated to the quantitative trait under study.

The blocks of loci that are inherited together (haplotype blocks) in populations of related individuals tend to be larger than haplotypes in populations of unrelated individuals. Hence, association mapping studies can localize QTLs more precisely than linkage analysis studies. In fact, the availability of milions of polymorphic markers has led to a dramatic growth of genome-wide association studies (GWAS) and large-scale genome-wide maps of QTLs are reported for quantitative traits of human biology and diseases (Altshuler et al., 2008; Kruglyak, 2008; The International HapMap Consortium, 2007; The 1000 Genomes Project Consortium, 2010).

Experimental crosses

QTL mapping studies have been conducted in humans as well as in nonhuman species, for instance, model organisms such as yeast or mice. In particular, the population resulting from an experimental cross from two inbred lines (or inbred strains) of a model organism is the simplest population in which QTL mapping can be performed (Broman and Sen, 2009). In these populations, the non-genetic factors such as the environment, the life history as well as the genetic composition can be under control providing more accurate phenotypic measurements and reducing the variance due to non-genetic effects. As long as the model and the study design are appropriate, the conclusions extracted from a QTL study in a model organism can be relevant for human genetics (Lehner, 2013).

In order to perform a QTL mapping analysis in a model organism we need a population that is genetically variable for the quantitative trait of interest. This population is the result of a cross between two homozygous inbred strains (both chromosomes have the same alleles at all positions) that ideally differ genetically with regard to the trait of interest. Then, for each individual in the population, we measure the phenotypic value and the genotypic values corresponding to a set of polymorphic genetic markers. Additionally, we need a genetic map specifying the locations of these marker loci on the chromosome.

Inbred strains are achieved by repeated sibling mating. If two inbred lines, raised under the same conditions, differ in a quantitative trait of interest, one can inspect the genetic basis of this trait by crossing the lines. This cross results in a population, F_1, whose individuals are heterozygous at all markers and QTLs because they receive one copy of each chromosome from one of the strains and another copy from the other strain. If F_1 individuals are crossed again, for example, with one of the parental strains (BC or backcross) or with themselves (F_2 or intercross), a new set of genetically variant individuals arise (see Figure 2.7). These are the result of a combination of the parental strains due to the occurrence of recombinant events during meiosis. This new population is expected to carry the genetic variability which should enable mapping QTLs associated to the trait of interest.

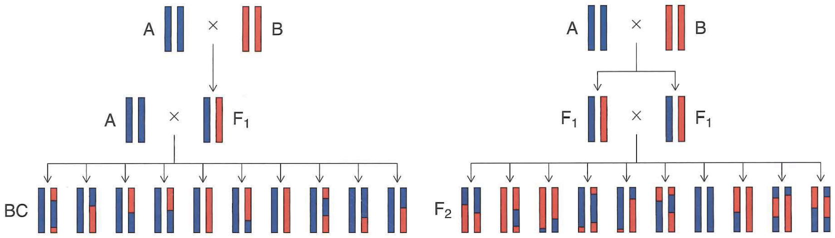

Figure 2.7: Representation of the two main types of cross between two inbred lines: a backcross (left) and intercross (right). If the parental strains are denoted as A and B, the F_1 individuals will have genotype AB at each locus. In a backcross, the F_1 population is crossed with one of the parental strains, for instance, with the A parental strain. Thus, BC individuals will have genotype AA or AB. In an intercross, the F_1 population is self-crossed so that F_2 individuals will have genotype AA, AB or BB at each locus. *Taken from Broman (2001).*

2.5.1 Detection of QTL

In this section we review the most important aspects that are related to the detection of QTLs as well as the basic statistical methodologies developed in this field.

Single marker regression

The simplest method for QTL mapping is called single marker regression (Soller et al., 1976) which is, basically, an analysis of variance (ANOVA). The basic assumption underlying this method is that given the QTL genotype g at a certain marker, the phenotype y follows a normal distribution

$$y_i|g_i \sim \mathcal{N}\left(\mu_{g_i}, \sigma^2\right) , \tag{2.2}$$

where the subscript g_i stands for the genotype of individual i at the marker. Thus, the phenotype has different mean values according to the genotype, $E(y|g) = \mu_g$, but we assume common variance, $\operatorname{var}(y|g) = \sigma_g^2 = \sigma^2$.

The scheme of this approach is summarized as follows: first, consider each marker individually and split the individuals into groups according to their genotype at that marker (2 groups in the case of a backcross, 3 if it is an intercross). Then, compare the average of phenotype values. We assume that, if the marker is linked to the QTL, we would see consistent differences between the averages of the different groups; otherwise, the phenotype distributions are approximately the same. The evidence for linkage to a QTL is assessed by computing a statistic called the LOD score which corresponds to the $\log_{10}$ likelihood ratio comparing the null hypothesis that there is no QTL at the interrogated marker and the alternative hypothesis that there is a QTL at that marker. Large LOD scores entail a greater evidence of the presence of a QTL at that position.

Let's see how to calculate the LOD score in a backcross. Consider the null hypothesis of no QTL. In this case, we assume that the phenotype of a population of n individuals follows a normal distribution with a single mean μ, independent of the genotypes, $y_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$, so that its likelihood function is:

$$\begin{aligned}\mathcal{L}_0 &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y_i-\mu\right)^2}{2\sigma^2}\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\sum_{i=1}^{n}\frac{\left(y_i-\mu\right)^2}{2\sigma^2}\right) = \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\right) .\end{aligned} \tag{2.3}$$

The maximum likelihood estimates (MLEs) of the parameters μ and σ^2 are $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$, respectively. If we plug these expressions in Equation (2.3), we obtain

$$\begin{aligned}\mathcal{L}_0 &= \left(\frac{2\pi}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\frac{2}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}\right) \\ &= \left(\frac{2\pi}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right) . \qquad (2.4)\end{aligned}$$

Analogously, the likelihood function under the alternative hypothesis, where we assume that $y_i|g_i \sim \mathcal{N}(\mu_{g_i}, \sigma^2)$, reduces to

$$\mathcal{L}_1 = \left(\frac{2\pi}{n}\sum_{i=1}^{n}(y_i-\bar{y}_{g_i})^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right) , \qquad (2.5)$$

where in this case, the MLE of μ is $\hat{\mu} = \bar{y}_{g_i} = \sum_{i=1}^{n} 1_{g_i}(i) y_i / \sum_{i=1}^{n} 1_{g_i}(i)$ and $1_{g_i}(i)$ denotes an indicator function that is 1 if and only if individual i has genotype g_i and 0, otherwise. The MLE of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu}_{g_i})^2$. Finally, we take the $\log_{10}$ of the ratio of $\mathcal{L}_0$ and $\mathcal{L}_1$ to obtain the LOD score:

$$\begin{aligned}\text{LOD} &= \log_{10}\left(\frac{\mathcal{L}_1}{\mathcal{L}_0}\right) = \log_{10}\left(\frac{\left(\frac{2\pi}{n}\right)^{-\frac{n}{2}}\left(\sum_{i=1}^{n}(y_i-\bar{y}_{g_i})^2\right)^{-\frac{n}{2}}\exp(-\frac{n}{2})}{\left(\frac{2\pi}{n}\right)^{-\frac{n}{2}}\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)^{-\frac{n}{2}}\exp(-\frac{n}{2})}\right) \\ &= \log_{10}\left(\frac{\left(\sum_{i=1}^{n}(y_i-\bar{y}_{g_i})^2\right)^{-\frac{n}{2}}}{\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)^{-\frac{n}{2}}}\right) = \frac{n}{2}\log_{10}\left(\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y}_{g_i})^2}\right) \\ &= \frac{n}{2}\log_{10}\left(\frac{\text{RSS}_0}{\text{RSS}_1}\right) , \qquad (2.6)\end{aligned}$$

where RSS_0 and RSS_1 stand for the residual sum of squares of the null and the alternative hypothesis, respectively.

The model for a backcross can be equivalently expressed as a linear model. If, for individual i, we take $z_i = 1$ if $g_i = \text{AA}$ and $z_i = -1$ if $g_i = \text{AB}$, we have that

$$y_i = \mu + \beta z_i + \varepsilon_i , \qquad (2.7)$$

where we assume that the ε_i are independent and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Therefore, the test for an association between a marker and a phenotype may be done by using a t test with $n-2$ degrees of freedom.

In order to calculate the LOD score for an intercross, we proceed as before. In this case, the only difference is that the genotype of a marker can take three different values: AA, AB or BB so that under the alternative hypothesis, the phenotype follows three normal distributions with different means according to the genotype at the marker. Furthermore, an intercross can be modeled as follows:

$$y_i = \mu + \alpha z_{i1} + \delta z_{i2} + \varepsilon_i\,, \tag{2.8}$$

where we take $z_{i1} = -1, 0, +1$ if g_i is AA, AB or BB, respectively, and α represents the coefficient for the additive effect of the marker on the phenotype. Additionally, the dominance effect is encoded in the second term where we take $z_{i2} = +1$ if $g_i =$ AB and $z_{i2} = 0$, otherwise. In this case, the linkage between a QTL and the phenotype may be tested by calculating the F_{p_1,p_2} statistic from ANOVA with parameters $p_1 = 2$ and $p_2 = n - 3$. In particular, there is a direct correspondence between the F statistic and the LOD score:

$$\begin{aligned} F &= \left(\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1}\right)\left(\frac{n - \mathrm{df} - 1}{\mathrm{df}}\right) \\ &= \left(\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1} - 1\right)\left(\frac{n - \mathrm{df} - 1}{\mathrm{df}}\right) \\ &= \left(10^{\frac{2}{n}\mathrm{LOD}} - 1\right)\left(\frac{n - \mathrm{df} - 1}{\mathrm{df}}\right), \end{aligned}$$

where df $= 2$ is the number of degrees of freedom of the model.

The main advantage of single marker regression is its simplicity which allows one, for instance, to easily incorporate covariates (Broman and Sen, 2009).

Interval mapping

Since, in general, we do not observe genotypes at the QTL but only at marker loci, the resolution for detecting and locating the QTL is subjected to the markers that show the greatest differences between the means of the different genotype groups. Therefore, it may happen that the QTL is far from all the markers so that the power to detect and locate it is low. In addition to this, if one applies single marker regression, individuals with missing genotypes and/or phenotypes must be discarded.

These shortcomings were first addressed by Lander and Botstein in 1989 (Lander and Botstein, 1989). In this paper, a new method named interval mapping was proposed and it is still today one of the most popular approaches for QTL mapping in experimental crosses. Here, not only the markers are tested to search for a QTL, but also each position on the genome is interrogated. Since markers and QTLs are associated due to linkage, a genetic map is used to calculate the probabilities of genotypes of each individual i at a certain position k in the genome based on the genotypes at its nearest flanking markers ($g_i^{(k-1)}$ and $g_i^{(k+1)}$ at positions $k-1$ and $k+1$), that is, $p_{ij}^{(k)} = \Pr\left(g_i^{(k)} = j | g_i^{(k-1)}, g_i^{(k+1)}\right)$ (Table 2.1). Given the genotypes at these flanking markers, the conditional phenotype values follow a mixture of normal distributions. In particular, the density function for the phenotype of individual i regarding the position k is

$$\phi(y_i;\mu,\sigma^2) = \sum_j p_{ij}^{(k)}\ \phi\left(y_i;\mu_j,\sigma^2\right)\ , \tag{2.9}$$

where ϕ denotes the density function of the normal distribution and the sum goes over all possible genotypes j (2 in the case of a backcross and 3 if we have an intercross). Several methods have been developed to calculate the LOD score based on different strategies to deal with missing genotype data: Lander and Botstein (1989) estimate the parameters with maximum likelihood under the mixture model using the expectation-maximization (EM) algorithm (Dempster et al., 1977); Haley and Knott (1992) and Martínez and Curnow (1992) proposed a faster method called Haley-Knott regression based on approximations of the mixture model; finally, a method based on the imputation of missing data was introduced by Sen and Churchill (2001). The main advantage of interval mapping is that the inference of QTLs at any position in the genome increases the power to locate and to estimate the effects of QTLs. Additionally, individuals with missing genotype data are not discarded anymore. On the contrary, the main drawback of this approach is that the calculations are computationally more expensive.

Statistical significance

Once LOD scores have been computed we need to determine the level of statistical significance for the existence of a QTL at a particular locus. If a single hypothesis test is performed at one specific marker, it is enough

Marker genotype		Probability of QTL genotype at position k	
$g^{(k-1)}$	$g^{(k+1)}$	$g^{(k)} = \text{AA}$	$g^{(k)} = \text{AB}$
AA	AA	$\frac{(1-r_{lQ})(1-r_{rQ})}{1-r_{lr}}$	$r_{lQ}r_{rQ}/(1-r_{lr})$
AA	AB	$(1-r_{lQ})r_{rQ}/r_{lr}$	$r_{lQ}(1-r_{rQ})/r_{lr}$
AB	AA	$r_{lQ}(1-r_{rQ})/r_{lr}$	$(1-r_{lQ})r_{rQ}/r_{lr}$
AB	AB	$r_{lQ}r_{rQ}/(1-r_{lr})$	$\frac{(1-r_{lQ})(1-r_{rQ})}{1-r_{lr}}$

Table 2.1: Conditional probabilities of QTL genotypes in a backcross. Here we show the probabilities of QTL genotypes given the gentoypes at two flanking markers, $g^{(k-1)}$ and $g^{(k+1)}$, in a backcross where r_{lQ} is the recombination fraction between $g^{(k-1)}$ and the QTL, r_{rQ} is the recombination fraction between $g^{(k+1)}$ and the QTL and r_{lr} is the recombination fraction between both markers.

to consider the null hypothesis that there is not a QTL at that specific locus and calculate the null distribution under this hypothesis. The 95th percentile of this distribution is usually used as the LOD threshold. Then, LOD scores above the threshold are considered significant so that the null hypothesis is rejected and we can accept that there is a QTL at that locus. In the case of single marker regression, it is known that the following transformation of the LOD score

$$D = 2\ln(10)\text{LOD} = n\ln(10)\log_{10}\left(\frac{\text{RSS}_0}{\text{RSS}_1}\right) = n\ln\left(\frac{\text{RSS}_0}{\text{RSS}_1}\right) , \qquad (2.10)$$

follows asimptotically a χ^2 distribution with 1 and 2 degrees of freedom for a backcross and an intercross, respectively.

However, as we are performing a large number of tests across the genome we need to correct for the test multiplicity in order to avoid large false positive rates. To this end, we formulate the global null hypothesis that there is no QTL anywhere in the genome which, as we shall see below, does not follow a χ^2 distribution. In this case, LOD scores larger than the 95th percentile of the null distribution mean that there is a QTL somewhere in the genome. Assume that we perform m tests across the genome and that the probability that some LOD score is larger than the LOD threshold, denoted by γ, is α. Then, we can write

$$\alpha = \Pr\left(\bigcup_{i=1}^{m}\text{LOD}_i > \gamma\right) = 1 - \Pr\left(\bigcup_{i=1}^{m}\text{LOD}_i \leq \gamma\right) \qquad (2.11)$$

and, in fact, this is equivalent to

$$\alpha = 1 - \Pr\left(\max_{i=1,\ldots,m} \mathrm{LOD}_i \leq \gamma\right) = \Pr\left(\max_{i=1,\ldots,m} \mathrm{LOD}_i > \gamma\right) . \qquad (2.12)$$

Therefore, the distribution of the global null hypothesis is equivalent to the distribution of the genome-wide maximum LOD scores. However, this global null hypothesis mainly depends on the number of inspected genome positions so that deriving the null distribution is not straightforward. At first, Lander and Botstein (1989) provided statistical thresholds through computer simulations as well as analytic computations of the null distribution. Nowadays, the most common way to calculate the null distribution is by performing permutations (Churchill and Doerge, 1994; Manichaikul et al., 2007): first, we permute the phenotype data relative to the genotypes which remain as in the original data set; then, we apply the QTL method to the permuted data and finally, we take the maximum LOD score. This operation is repeated a certain number of times (as reported by Broman and Sen (2009) or Doerge (2002), 1,000 times is enough) and a curve of maximum LOD scores is obtained. The precision of the null distribution increases with the number of permutations. However, this entails an enormous increase in the computational cost of the method. There are other alternatives to obtain thresholds such as the bootstrap resampling (Good, 2005) which is based on the randomization with replacement of phenotype values.

In order to gain consistency across the different methods and the different stringency thresholds, some authors (Leduc et al., 2012) not only report significant QTLs but also use different terms, proposed by Lander and Kruglyak (1995), regarding the strength of the linkage. This terms are based on the number of times that one would expect to see a result at random in a dense, complete genome scan and include suggestive, significant, highly significant, and confirmed linkages.

Covariates

As we already explained in Section 2.2, not only the genetic effects (QTLs) affect the variability of the phenotypes but also there are variables, such as environment, sex, diet or even a second QTL, that may contribute to the variation of the traits. Usually, if the effect of these other variables is large, it can be useful to include them in the analysis as covariates in order to reduce the residual variation, and therefore, increase the power to detect the causal QTL. Otherwise, some QTL effects may be masked or the detection of QTLs under some circumstances may not replicate under others.

The covariates may have an additive effect on the phenotype. In this case, the QTL and the covariate are independent so we would test the model

$$y_i = \mu + \beta g_i + \gamma x_i + \varepsilon \,, \tag{2.13}$$

where y_i is the phenotype of individual i, g_i refers to the genotype of the QTL, x_i corresponds to the covariate and β and γ are their corresponding coefficients, against the null model

$$y_i = \mu + \gamma x_i + \varepsilon \,. \tag{2.14}$$

Furthermore, in addition to the additive effect, the covariates may also interact with the QTL. In this case, the effect of the QTL on the phenotype depends on the value of the covariate. In particular, if the covariate is another QTL, we would test for the epistatic effect between both QTLs. This can be modeled by adding an interacting effect in the alternative model:

$$y_i = \mu + \beta g_i + \gamma_1 x_i + \gamma_2 g_i x_i + \varepsilon \,. \tag{2.15}$$

In general, methods that have been described previously can be extended to include the effects of covariates. However, when adding QTLs as covariates, it is more appropriate to use the composite interval mapping method (Zeng, 1994; Jansen and Stam, 1994).

Epistasis

Although the methods that search for individual QTLs work well, they ignore the more realistic scenario in which a trait is affected by multiple linked and/or interacting QTLs, that is, when QTLs show epistasis. In this case, the phenotype can be modeled as

$$y = \mu + \sum_i \beta_i g_i + \sum_i \sum_j \gamma_{ij} g_i g_j \,, \tag{2.16}$$

where y is the phenotype, the first summatory corresponds to the additive effects of the genotypes and the second one refers to the interaction effects between them.

The problem of identifying the set of QTLs and their possible interactions that fit better the data is challenging. From a statistical point of view, it is considered a model selection problem (Hastie et al., 2009; Sillanpää and Coriander, 2002) whose ultimate goal is finding pairs of interacting QTLs. In the QTL mapping framework, it consists of four steps: first, we need to select the QTL model (that is, if we restrict the QTLs to have additive effects or we allow them to show epistatic effects); second, we choose a method to fit the data to model; next, we search through the space of models (the main approaches include forward selection, backward selection, stepwise selection, randomized searches), and finally, we compare these models (for instance, by applying a penalized likelihood (Manichaikul et al., 2009) or by using the BIC or AIC criteria).

The main advantage of methods that search for multiple QTLs is that the resolution and power to find interacting QTLs increases. Moreover, by controlling for the presence of a QTL, the residual variation may be reduced and, consequently, the power to detect additional QTLs increases. However, the implementation of these methods is, in general, difficult due to the large number of potential QTLs and combinations of interacting QTLs, and therefore, they are computationally very expensive. Among the methods developed for this purpose, we can find multiple QTL extensions of the Haley-Knott regression, the multiple imputation approach (reviewed in Broman and Sen, 2009) or the multiple interval mapping (Kao et al., 1999). We can also highlight the composite interval mapping approach proposed by Zeng (1994) and Jansen and Stam (1994), the two step strategy developed by Sen and Churchill (2001) or the Bayesian approaches reviewed in Yi and Shriner (2007). Most of these methods are implemented in different statistical software (Basten and Zeng, 2002; Yandell et al., 2007; Arends et al., 2010).

Pleiotropy

Often, a mutation in a single gene can have an effect on multiple traits. This is called pleiotropy. If the QTL affects all the traits in the same manner, it is probable that these traits are implicated in the same genetic pathway or are part of the same biological process. By contrast, when the QTL affects the traits in opposite direction, usually these traits are implicated in fitness components (Falconer and Mackay, 1996, pg. 344).

In some cases, spurious pleiotropic associations may arise (Solovieff et al., 2013). This is often the case when a SNP is found to be associated to different traits that are not functionally related. Among other causes, it may happen that this SNP is genotyped in a region of high linkage disequilibrium (LD). LD stands for the non-random association of alleles at different loci. Hence, the SNP may be associated to many QTLs located in different genes with distinct functions. Challenges and strategies to identify and understand pleiotropic effects in human complex traits have been recently reviewed in Solovieff et al. (2013). Pleiotropic effects are widespread in cellular traits, such as gene expression.

2.5.2 Location of QTL

In addition to find evidence for the existence of QTLs associated to a quantitative trait, we also want to locate them precisely along the genome and to identify the gene responsible for the functional variation of the trait. This task can result tedious since the QTL regions can be very large and may contain many genes. However, it can be easier if we take into account some aspects such as the recombination fraction between loci or the degree of LD. If there are large blocks of markers in LD we don't need to use all of them in order to detect an association between the markers and the phenotype but, on the contrary, it will be more difficult to locate the QTL with precision. In general, as the density of markers used in the study is higher, the localization of QTLs is more precise.

The location of the QTLs in the genome can give us some hints about its functional role and, eventually, it can give some clues about the gene that affects the trait (Mackay et al., 2009). For example, QTLs that are non-synonymous polymorphisms and that are located in a coding region lead to different amino acid sequences so that the variation of the quantiative trait may be easy to explain. On the other hand, QTLs that are synonymous polymorphisms in coding regions are usually associated with mRNA stability (Nackley et al., 2006). QTLs located in promoters and introns may affect the transcription factor binding and mRNA splicing, and therefore, can affect the pattern of expression of genes and their isoforms.

Genes that show a functional relationship with the QTL, can be investigated by using functional complementation (Mackay, 2001). Further, techniques such as positional cloning or targeted gene replacement can be applied to identify the gene carrying the genetic variation (Falconer and Mackay, 1996; Mackay, 2001). Despite the detection of many significant QTLs, the number of identified genes that explain the functional role of the QTL remains small compared to the large amount of reported QTL mapping studies (Roff, 2007).

2.5.3 Effects of QTL

The statistical power to map QTLs depends on many factors including the type of cross, the sample size, the density of markers, the stringency of the LOD threshold or the size of the effect of the QTLs. The effect of a QTL is defined as the difference in the phenotype averages among the QTL genotype groups. In a backcross, it refers to the additive effect and it is defined as the difference between the average phenotypes for the heterozygotes and the homozygotes $a = \mu_{\mathrm{AA}} - \mu_{\mathrm{AB}}$. In an intercross, where we have three different genotypes, AA, AB, and BB, the additive effect of a QTL is $a = (\mu_{\mathrm{AA}} - \mu_{\mathrm{BB}})/2$ whereas the dominance effect is defined as $d = \mu_{\mathrm{AB}} - (\mu_{\mathrm{AA}} + \mu_{\mathrm{BB}})/2$.

Selection bias

The estimated effect of a QTL is, in general, subject to certain bias. Let's assume that we test the presence of a QTL in different data sets and we calculate its estimated effect in each of the data sets. Some of these values will be larger than the true effect of the QTL and some will be smaller. In these latter cases, the power to detect the QTL may be insufficient and the QTL may not be detected. Therefore, the average QTL effect calculated from the data sets for which the QTL has been detected is larger than its true effect. This is called selection bias.

In particular, selection bias is large when the true effect is moderate or low, and therefore, the power to detect the QTL is low. On the contrary, a QTL with a very large effect, has also sufficient power to be detected and, in this situation, the difference between the estimated and the true effect of the QTL is negligible.

Estimation of the phenotypic variance explained by a QTL

Over the last decades, the advances of the technology have enabled the analysis of data sets consisting of a high marker density as well as large sample sizes so that the mapping resolution has increased notably. As a consequence, many QTLs of small effects that were not detected previously due to the lack of power have now been identified (Mackay et al., 2009). Moreover, the ability to detect and estimate QTL effects in experimental crosses can be increased through an accurated and careful experimental design (Broman and Sen, 2009). However, we have seen in Section 2.3 that, in general, QTLs do not account for all the genetic variation of the trait and the causes of this missing heritability are not always understood (Manolio et al., 2009).

In particular, we can measure the strength of the results by calculating the variance attributable to a given locus. The proportion of phenotypic variance explained by one QTL (the narrow-sense heritability due to the QTL) is defined as

$$\eta^2 = \frac{\mathrm{var}\{E(y|g)\}}{\mathrm{var}(y)}\,, \tag{2.17}$$

where y is the phenotype and g represents the genotype of the QTL. This equation can be written in terms of the additive a and dominance d effect of the QTL on the phenotype so that, for a backcross, we have

$$\eta^2 = \frac{a^2}{a^2 + 4\sigma^2}\,, \tag{2.18}$$

and, for an intercross, we have

$$\eta^2 = \frac{2a^2 + d^2}{2a^2 + d^2 + 4\sigma^2}\,,$$

where σ^2 is the residual variance of the phenotype. Moreover, the phenotypic variance explained by one QTL can be easily calculated from the LOD score:

$$\eta^2 = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_0} = 1 - 10^{-\frac{2}{n}\mathrm{LOD}}\,. \tag{2.19}$$

In this case, the difference between the variance attributable to the QTL and the narrow-sense heritability of the trait is the fraction of missing heritability.

2.6 Discussion

Understanding the genetic basis of variation for quantitative traits is one of the most challenging problems of current biology. QTL mapping studies have emerged as a particularly helpful technique to bridge the gap between DNA and quantitative phenotypes and diseases. The main purpose of QTL mapping studies is to associate the variation of DNA sequence to the variation of multiple organismal phenotypes while ignoring all other intermediate steps along the path from genotype to phenotype. However, it has been shown that some cellular traits are heritable. Therefore, the strategy followed by classical QTL mapping studies can be applied to any complex trait, for instance, to RNA transcript abundance, an strategy called eQTL mapping (Brem et al., 2002), or to the amount of protein produced from genes, called pQTL mapping (Wu et al., 2013; Albert et al., 2014).

The availability of these different sources of data adds new dimensions to enable the characterization of gene regulation. By jointly analyzing different cellular traits we achieve a more complete and comprehensive view of the flow of information between the genotype and the organismal phenotype.

Therefore, the future of QTL mapping lies in the integration of different types of data, often called integrative genomics analyses, which enables the study of complex traits from a network perspective. To deepen in this question, in the next chapter we study one of the primary types of integrative genomics strategies, called QTL mapping of gene expression profiles.

3. Genetics of gene expression

3.1 Introduction

Most cellular functions are executed by proteins and proteins are encoded by genes (Figure 3.1). The process by which the information from a gene is used to synthesize a protein in a cell is called gene expression. Each step in the information flow from DNA to protein through RNA may be regulated. Gene regulation allows the cell to control and adjust the amount and type of proteins it manufactures with the ultimate goal of determining its structure and function. The throughput at which a cell makes proteins is greatly influenced by the synthesis of its RNA transcripts, and therefore, by its gene expression levels. Thus, the ability to quantify the level at which a particular gene is expressed can give a huge amount of information about its transcriptional regulation and ultimate function. Gene expression levels are measured by quantifying the steady-state of messenger RNA (mRNA) abundance within cells or tissues.

Microarray technology has been used for some decades to analyze changes in gene expression (for example, under different conditions such as case-control studies or time-course experiments). Microarrays have been also used to measure the variation of DNA (for example, with respect to a phenotype) as we have seen in the previous chapter. However, it is not until the beginning of the 21st century that the simultaneous assay of both types of data, variation of DNA and variation of gene expression, on the same biological samples is performed with high-throughput technologies. This first type of integrative genomics data has enabled researchers tackling new questions in genetics by coupling the power of traditional genetics and the technological successes of genomics (Figure 3.2). This strategy is called genetical genomics (Jansen and Nap, 2001; Li and Burmeister, 2005; Rockman and Kruglyak, 2006).

The basic principle of genetical genomics is that the expression level of a gene in the population, that is, the amount of mRNA produced by the

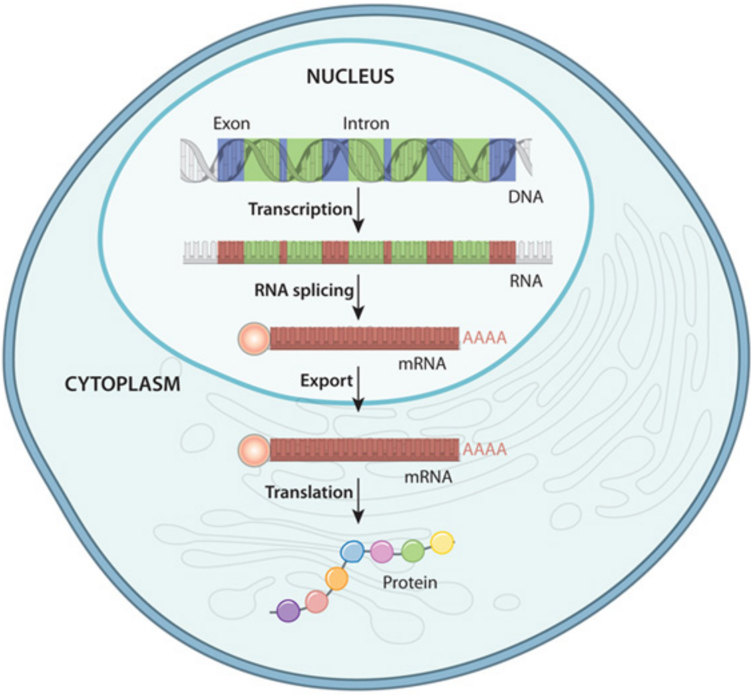

Figure 3.1: Central dogma of molecular biology. DNA sequence is first replicated and then transcribed into mRNA. Green regions (introns) do not contribute to the synthesis of the protein and they are removed. The exons (red) are spliced together forming a mature mRNA molecule that is exported out of the nucleus of the cell. Once in the cytoplasm, the mRNA is translated into a protein. *Taken from http://www.nature.com/scitable/topicpage/gene-expression-14121669.*

gene, is treated as a quantitative trait. The first analyses of genetical genomics data were performed on model organisms (Brem et al., 2002; Schadt et al., 2003) and immediately revealed that the impact of genetic variation on the expression level of genes is common, widespread and highly heritable (Brem et al., 2002). Since then, the study of gene expression genetics, that is, the localization of regions in the genome, called expression quantitative trait loci (eQTLs), that influence the cellular phenotype of gene expression has been widely tackled. In fact, because eQTL studies connect the variation at the DNA sequence level to the one at the RNA level, gene expression levels may be viewed as intermediate molecular phenotypes which can help narrowing the gap between the DNA polymorphisms and higher-level phenotypes (Schadt

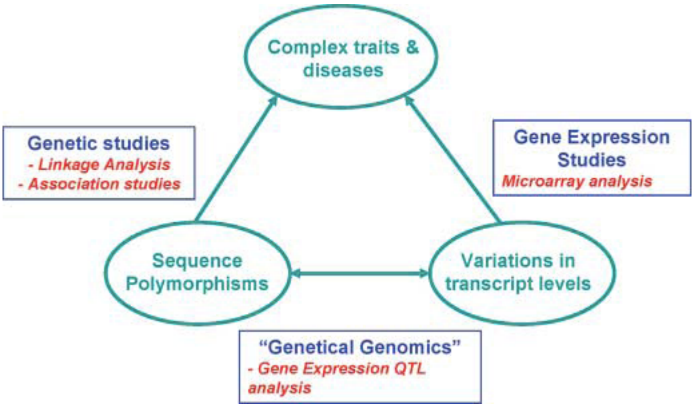

Figure 3.2: Genetical genomics. The combination of traditional genetic analysis which study the genetics of complex traits and microarray studies that analyze the variation of transcript levels leads to the study of the genetics of gene expression. *Taken from Li and Burmeister (2005).*

et al., 2003, 2005; Cookson et al., 2009).

The strategy followed by eQTL mapping studies is analogous to that of QTL analysis. First, individuals are genotyped at a set of molecular markers. Second, gene expression is profiled in the tissue or the experimental condition of interest. Then, a statistical procedure is used to find eQTLs that show significant associations with the mRNA levels. These eQTLs can be broadly categorized according to their location relative to the gene whose expression levels are being mapped onto: a *cis*-eQTL is located at the same genomic region of its target gene; if the eQTL is located elsewhere on the genome, we say that it is a *trans*-eQTL. One of the strengths of eQTL studies is that a prior knowledge of the regulatory mechanisms is not required to find the location of the genome regions that underly the transcriptional variation. Once the eQTLs are detected, bioinformatics methods can be used to identify functional relationships among the transcripts affected by the same loci and to investigate the structure of the underlying regulatory networks (Brem et al., 2002; Bing and Hoeschele, 2005; Schadt et al., 2005; Liu et al., 2008; Flassig et al., 2013).

However, gene expression is a high-dimensional multivariate trait involving measurements from thousands of genes that act coordinately under complex molecular regulatory programs. This makes eQTL mapping a very challenging problem for two reasons. One is that gene expression profiles can be highly correlated as a product of gene regulation. When marginal eQTL associations that only involve one gene at a time are inspected, the correlation structure of genes complicates the distinction between direct and indirect effects. Another is that high-throughput expression profiling can be very sensitive to non-biological factors of variation such as batch effects (Leek and Storey, 2007; Leek et al., 2010), introducing heterogeneity and spurious correlations between gene expression measurements. These artifacts may compromise the statistical power to map truly biological eQTLs (Stegle et al., 2010) or show up as interesting genetic switches with broad pleiotropic effects affecting a large number of genes, commonly known as eQTL hotspots (Leek and Storey, 2007; Breitling et al., 2008).

One approach to address these shortcomings is to consider the eQTL mapping problem from a multivariate perspective. Thus, by using natural genetic variation as a source of perturbations, it may be possible to infer a network of causal relationships between eQTLs and genes (Chen et al., 2007). Further, an immediate advantage over the classical univariate approach is that the presence of simultaneous variation at multiple loci allows the exploration of a larger space of variation than that performed by carrying out single perturbations (Rockman, 2008).

In the rest of the chapter we provide details of every step in an eQTL mapping study. First, a detailed description of the genetical genomics data sets is provided. In Section 3.3, we review the principal methods that have been developed for eQTL mapping. The classification of eQTLs according to their location is explained in Section 3.4. In the last section, we discuss the main challenges of eQTL mapping studies.

3.2 Genetical genomics data

One of the benefits of eQTL mapping studies is that they can be conducted in many types of populations, for instance, in a progeny from a cross between two parent strains of a model system such as yeast (Brem et al., 2002) or in a set of unrelated individuals (Fairfax et al., 2012). Individuals composing the population are genotyped for a set of polymorphic markers covering the genome. The transcript abundance is also measured for each gene by using high-throughput gene expression profiling technology, such as microarrays or RNA-seq. Both classes of information compose the genetical genomics data sets (Figure 3.3).

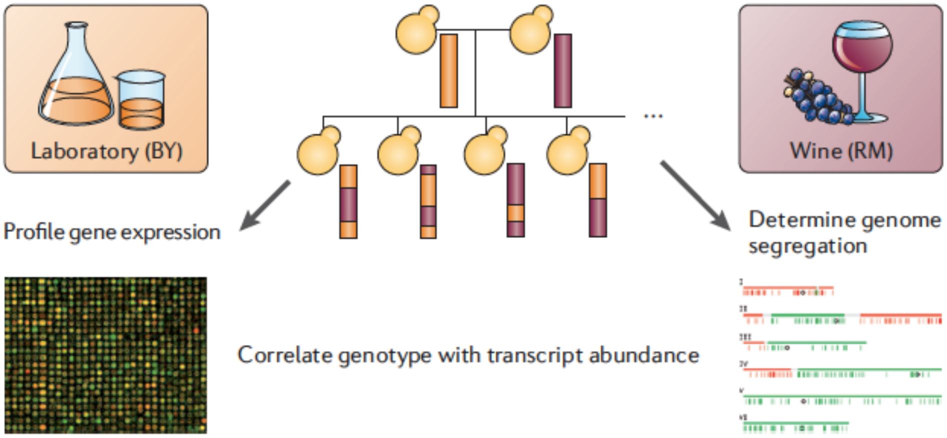
Laboratory (BY)
Wine (RM)
Profile gene expression
Determine genome segregation
Correlate genotype with transcript abundance

Figure 3.3: eQTL mapping experimental design. This figure illustrates the experimental design for a cross between two yeast strains (Brem et al., 2002) that resulted in a population of 112 samples. The gene expression profile (left) and the genotypes for a set of genetic markers (right) were obtained by using the microarray technology and constitute the genetical genomics data set used to find the genetic variants that correlate with the gene expression values. *Taken from Rockman and Kruglyak (2006).*

Array-based gene expression profiling

A microarray chip is a slide composed by a collection of probes which are used to measure in parallel the expression levels of a large number of genes or to genotype multiple regions of a genome. The concept of microarrays emerged in 1991 (Fodor et al., 1991). Since then, different types of arrays have been developed to identify gene expression signatures, detect important genetic variants or aid in cataloging the diverse molecular patterns underlying biological and physiological processes.

Schematically, a microarray experiment consists of the following steps: first, a chip is built containing millions of oligonucleotides (probes) which are short sequences of DNA corresponding to parts of the genome we are interested in. Second, RNA is extracted from the samples and reverse-transcribed into complementary DNA (cDNA). These cDNA samples are fragmented and labeled with fluorochrome, a fluorescent chemical compound whose fluorescence's intensity can be measured by a scanner. Then, cDNA fragments bind to their respective (complementary) probes onto the array, in a process called hybridization. Fragments that did not bind to any sequence are washed off so that only strongly paired strands remain hybridized. Finally, a scanner measures the intensity of the fluorochrome of each probe in terms of a fluorescence scale that typically ranges from 1 to 2^{16} - 1 and, in order to ease its interpretation, the $\log_2$ value of intensity is usually taken. This intensity is proportional to the amount of mRNA that binds to the probe and determines the gene expression level.

The first gene expression arrays used a two channel technology and hybridized two samples together (e.g., case and control, wild type and mutant or a tissue sample and a reference). Each sample is labeled with a different color, green (fluorochrome Cy3) and red (fluorochrome Cy5). The two cDNA samples are mixed and hybridized to a single microarray which is then scanned to visualize the relative differences in abundance of the RNA present in each sample. In order to avoid the bias caused by dye-specific effects, the experiment design includes a dye swap which consists in duplicating hybridizations with the same samples using the opposite labeling scheme.

After dual channel, single channel microarrays were introduced. Single channel microarrays use only one color to label the samples and each sample is applied on a separate chip. These arrays provide an estimation of mRNA abundance level for each gene. Therefore, the comparison of two sets of conditions requires two separate single-dye hybridizations. One of the most popular microarray platforms are high-density oligonucleotide arrays produced by Affymetrix (Lockhart et al., 1996).

Microarrays need to be pre-processed to remove the variance of the gene expression measurements due to sample-specific effects so that individual samples can be more appropriately compared. A popular method to pre-process one channel microarrays is the Robust Multichip Average (RMA) algorithm (Irizarry et al., 2003) and, more recently, the frozen Robust Multiarray Analysis (fRMA) (McCall et al., 2010). Analogously, the `limma` software (Smyth, 2005) provides functions to pre-process two channel microarray experiments: for instance, the background correction adjusts for cross hybridization resulting from nonspecific binding of the fluorochrome on the array whereas normalization gives the same distribution of values to each chip.

Throughout this thesis, gene expression data is assumed to come from microarrays. In Chapter 6, we use a real data set whose gene expression levels come from a two channel microarray experiment (Brem and Kruglyak, 2005).

Sequence-based gene expression profiling

The development of high-throughput sequencing (HTS) technologies, also known as next generation sequencing (NGS), has enabled profiling gene expression by sequencing based methods, called RNA-seq (Wang et al., 2009). RNA-seq provides both the sequence and relative quantification of RNA molecules that are present at any particular time in a specific cell type or tissue (Nagalakshmi et al., 2008; Mortazavi et al., 2008; Battle et al., 2013). RNA-seq not only provides accurate RNA expression levels but it also enables a deeper understanding of the transcriptome of different species and cell types. In particular, it has become a powerful technology to detect allele-specific expression (Rozowsky et al., 2011) or to quantify isoform abundances, also known as isoform quantification (Li and Dewey, 2011). RNA-seq has also been used to analyze alternative splicing (Morin et al., 2008).

A typical RNA-seq experiment consists of the following steps. First, RNA is isolated and purified and then it is randomly fragmented into small pieces. These fragments of RNA are reverse-transcribed to obtain a collection of cDNA fragments, referred to as a library, with specific sequences (adapters) attached to one or both ends. These adapters are needed to bind to the complementary adapters on the sequencing device. cDNA fragments are amplified so that clusters of the same fragment are built. Each of these fragments are then sequenced in parallel and millions of short sequence reads are produced. Several sequencing technologies are available for RNA-seq, for instance, the systems developed by 454 Life Sciences (Roche) (Margulies et al., 2005) or Illumina (Bennett et al., 2005).

Before proceeding with a downstream analysis, sequenced reads need to be processed. In the case of gene expression estimation, the reads are mapped to a reference genome. The read mapping process presents both biological and technical challenges and several strategies have been developed to overcome them (Langmead et al., 2009; Li and Durbin, 2009; Marco-Sola et al., 2012). Then, gene expression level is measured by counting the number of reads mapping to a gene, exon or transcript, and normalized to avoid possible biases. As a result, a genome-wide transcription map is built and the transcriptional structure and the level of expression is obtained for each gene.

RNA-seq data have some advantages over the microarray data. These include, among others, greater sensitivity of gene expression measurements and reduced technical variation. In particular, the added value of RNA-seq lies in a larger dynamic range than with fluorescent-based microarray measurements, and more power to interrogate lowly expressed or unknown genes. However, methods to analyze NGS data are still evolving and the path that researchers need to follow from sample processing to analysis results is still longer than with gene expression microarrays.

Batch effects

In order to obtain robust and reproducible results with microarray and RNA-seq data, it is necessary to analyze many biological replicates (Hansen et al., 2011). If we consider different samples we have more statistical power to detect changes in the expression level of genes. However, these biological samples cannot always be processed at once and they are often divided into different groups.

A set of samples that are processed together is called a batch. In a biological experiment, this sample grouping may introduce a systematic bias called batch effect. Batch effects can occur because measurements are affected by different laboratory conditions (Leek et al., 2010). For example, a batch effect may arise if groups of microarrays are processed in different laboratories, by different technicians or in several days. Different reagent lots can also affect the final measurements. These sources lead to increased variability and decreased power to detect the real biological signal. If an appropriate block design is employed, batch effects can be identified, controlled and eventually corrected. However, in a bad experimental design, batch effects may correlate with the variable of interest and lead to wrong conclusions.

This source of non-biological variation is not removed by pre-processing techniques such as normalization (Leek et al., 2010; Hansen et al., 2011) and specific methods need to be applied to properly address this bias.

Genotyping

Analogously to gene expression arrays, genetic markers can be genotyped by using high-density oligonucleotide SNP arrays. In a SNP array, hundreds of thousands of probes compose a small chip so that many SNPs can be interrogated simultaneously. Since SNP alleles only differ in one nucleotide it is difficult to achieve an optimal hybridization. To overcome this issue, SNP arrays are designed to have several redundant probes that have the SNP site in different locations. These probes also contain mismatches to the SNP allele. Then, we can determine the specific allele of each SNP by comparing the differential amount of hybridization of the target DNA to each of these redundant probes.

SNPs may also be genotyped by using NGS technology (Liti et al., 2009). NGS data may also suffer from error rates that are due to multiple factors, including base-calling and alignment errors. For a review of the methods developed to overcome these problems, the reader may be referred to Nielsen et al. (2011).

Finally, SNP data usually have a certain rate of missing data, which may affect the results of their downstream analysis. The percentage of missing genotypes in a SNP data set may vary from 5% to 20% (Huentelman et al., 2005) and specific techniques, such as imputation, are required to avoid the practical consequences of this problem.

3.3 eQTL mapping methods

In this section we review some of the approaches that have been developed to map genetic variants that affect gene expression.

Many methods for eQTL mapping are based on the insight that expression levels can be analyzed with statistical approaches in the same manner as any other quantitative trait phenotype. For this reason, a straightforward way to map eQTLs onto the genome is by treating each gene as an independent continuous trait and applying classical QTL mapping techniques (Kendziorski and Wang, 2006; Tesson and Jansen, 2009) as we have discussed in Chapter 2. These methods are mostly based on univariate linear models. The expression level of an individual gene is modeled as a linear combination of the interrogated eQTL (which can be a genotyped marker or any other putative location on the genome) and other explanatory variables representing, for instance, the effects of environment, sex or other loci.

Methods that search heuristically for a combination of many loci affecting the variation of a phenotype (Zeng, 1994; Broman and Sen, 2009) and those that test for epistasis (Lander and Botstein, 1989; Haley and Knott, 1992) can also be applied to eQTL mapping. The evidence for an association of an eQTL is usually measured by the computation of the $\log_{10}$ ratio of the model assuming the eQTL and the null model of independence between the eQTL and the gene. Statistical issues related to the significance of the LOD score are equally treated as in QTL mapping studies.

Identifying non-spurious eQTL associations

When applying a classical QTL mapping technique, we are not taking into account the high-dimensionality of the gene expression trait and their correlated structure. Therefore, by treating each gene expression individually and looking for its marginal association to an eQTL, we may find many significant eQTLs which, in fact, are indirectly affecting many genes. Moreover, a wide range of confounding effects lie hidden in the data which, if not properly addressed, can lead to both spurious and missed associations. The importance of identifying non-spurious *trans*-acting eQTLs has been largely reported in the literature.

The nature of the confounding effects, also known as expression heterogeneity, is diverse. We have seen in the previous section that gene expression data may be affected by technical batch effects (Leek et al., 2010). Confounding effects can also arise due to population structure such as race or family-relatedness, which exist if subgroups of individuals are more closely related, that is, genetically more similar among each other than with the rest of the population. In general, if the confounding effect is identified, such as sex or an environmental variable, it can be included in the model as a covariate. Otherwise, several strategies have been developed to estimate, adjust and eventually remove the effects of the confounders.

One approach to adjust batch effects is the one proposed by Leek and Storey (2007) called surrogate variable analysis (SVA). First, this method applies a single value decomposition to a residual expression matrix that is obtained after removing the signal of the primary variable of interest from the gene expression measurements. From this decomposition, the method identifies the singular vectors that explain the variation of the residual matrix in order to determine the confounding signatures. Then, the set of genes that is associated to each confounding signature is identified and from each of these sets a surrogate variable is built. Finally, these surrogate variables can be used as main effects in a linear model to adjust for unwanted variability. Another method that includes the confounding factors in the model as main effects is the one proposed by Stegle et al. (2010).

Another technique to address batch effects is called ComBat (Johnson et al., 2007). This technique directly removes known batch effects. The main disadvantage of this and other techniques that remove, instead of adjusting for, unwanted variability is that it may also remove a fraction of the interesting biological variability.

Some years ago, linear mixed models started to be exploited with great success: Kang et al. (2008a) proposed a method called intersample correlation emended (ICE) which corrects for technical confounding factors whereas in Kang et al. (2008b), they propose a method (EMMA) to correct for population structure. Two years later, Listgarten et al. (2010) developed a strategy to jointly correct for both types of confounding effects. In linear mixed models, known confounding effects and SNP data are modeled as fixed effects whereas unknown confounding factors are taken as random effects.

Beyond the univariate models that explicitly control for confounding effects, there are other approaches that have been developed with the objective of identifying non-spurious associations between genetic markers and genes, in particular, non-spurious *trans*-associations. These approaches use multivariate techniques to integrate a gene regulatory network model into the QTL mapping framework to provide a systems view of the underlying genotype-phenotype map, called eQTL network. To achieve this purpose, the primary strategy of these methods is to consider the correlation structure among transcripts. Among them, we can include methods based on sparse factor analysis (Parts et al., 2011), structural equation models (Liu et al., 2008), fused lasso regression methods (Kim and Xing, 2009), a combination of existing machine learning algorithms (Ackermann et al., 2012), sparse partial least squares (Chun and Keleş, 2009), random forests methods (Michaelson et al., 2010) or regularization techniques (Yin and Li, 2011).

In the Bayesian context, we can highlight the Bayesian networks that use conditional mutual information with constraint-based algorithms (Zhu et al., 2004) or the ones that use the BIC criterion with score-based algorithms (Neto et al., 2010). Some years ago, Zhang et al. (2010) proposed a Bayesian approach for detecting modules of co-expressed genes and their linked markers so that pleiotropic and epistatic effects can be identified.

Mixed graphical model theory (Lauritzen and Wermuth, 1989) is another multivariate statistical framework used to infer eQTL networks. Instead of assessing the marginal dependence between a genetic marker and a gene, mixed graphical models rely on the concept of conditional dependence. This is a more strict notion of association than marginal correlation and constitutes a powerful tool for distinguishing direct from indirect effects. Most of the approaches using conditional independence (Bing and Hoeschele, 2005; Chen et al., 2007; Neto et al., 2008; Kang et al., 2010) are based on conditional independence tests of order one, that is, they are restricted to conditioning on one other gene to disentangle direct and indirect relationships. An alternative is provided by Edwards et al. (2010) whose method restricts the search of mixed graphical models to tree network topologies that provide useful insights. However, these topologies restrict again the order of correlations to one.

Dealing with missing genotypes

As introduced in the previous section, an important problem regarding genotyping is that a fraction of markers may have missing values in a sample. Different strategies are proposed to overcome this problem.

A straightforward solution, known as complete-case analysis, is to omit observations with missing values. In contrast, methods that replace missing data with substituted values, have become a common alternative. This technique, called imputation, enables the use of existing learning algorithms for complete data but at the risk of introducing potential biases and errors. Diverse approaches have been applied to the problem of imputing missing genotypes such as the nearest-neighbor search of Roberts et al. (2007) or the multiple imputation proposed by Sen and Churchill (2001) implemented in Broman et al. (2003) for genotype data from experimental crosses.

3.4 eQTL regulatory mechanisms

A useful way to characterize eQTLs is to distinguish them between local and distant according to the genomic position of the eQTL with respect to the gene whose expression level is affected by the eQTL. eQTLs are local if they are located in the same chromosome and within some physical or genetic distance of a chosen point in the gene sequence, typically the transcription start site (TSS). They are distant if their position is elsewhere in the genome. This classification is strictly positional and it does not say anything about the molecular nature nor the molecular mechanism of the eQTL.

However, instead of local and distant, the most commonly used terms are *cis* and *trans*, respectively, although the original definition of these latter terms has an implicit mechanistic meaning (see Figure 3.4). Genomic regions acting in *cis*, also known as *cis* regulatory elements, have an allele-specific effect on the expression level of the gene they regulate. *Cis* regulatory elements act at the level of DNA and often regulate the transcription initiation or RNA stability. These include, for instance, promoter regions, enhancers, splice sites or poly-A signals. By contrast, *trans* regulatory elements are those that affect the regulation of both alleles of the target gene (in diploid organisms). Genes that encode proteins or functional RNA such as transcription factors or insulator proteins are examples of *trans*-acting regulatory elements.

Therefore, the classification of *cis* and *trans*-eQTLs based on the distance to the target gene sometimes may lead to confusion. For this reason, some authors such as Rockman and Kruglyak (2006) prefer the terms local and distant. For instance, it may happen that an eQTL is located in a gene encoding a transcription factor that affects an adjacent target gene (Ronald et al., 2005). Since the eQTL is close to the target gene, we will classify this eQTL as a *cis*-eQTL although it is located in a *trans* regulatory element. On the other hand, an eQTL located in a *cis* regulatory element may be classified as *trans* due to a very conservative definition of the region we enforce to contain *cis*-eQTLs, called *cis*-window.

Yet, since we usually find *cis* regulatory elements near the gene they regulate and *trans* regulatory elements far from the target gene, it is common to classify (Figure 3.5) local eQTLs as *cis*-eQTLs and distant eQTLs as *trans*-eQTLs (Brem et al., 2002; Bing and Hoeschele,

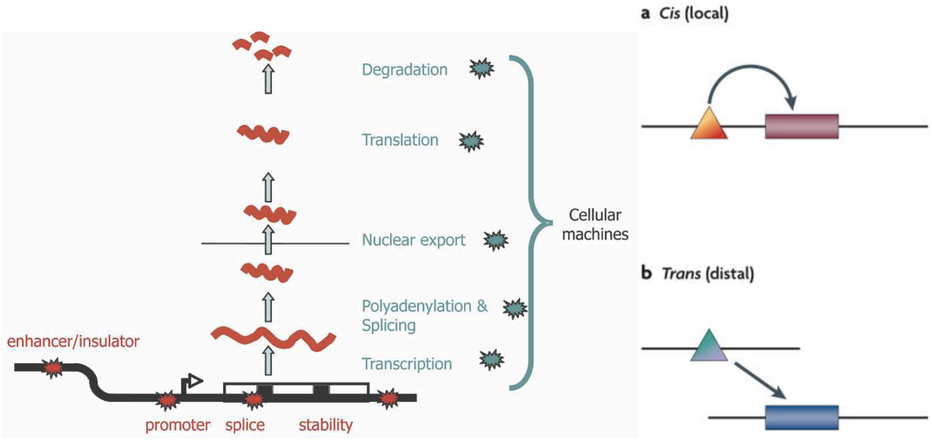

Figure 3.4: Left: Genetic variation that influences gene expression may act in *cis* if it is located in regulatory sequences (red stars). Genetic variants in the molecular machinery that interact with *cis*-regulatory sequences are considered *trans*-eQTLs (blue stars). *Taken from Williams et al. (2007).* Right: representation of a *cis* (a) and a *trans* (b) eQTL (triangle) affecting a gene (rectangle). *Taken from Cheung and Spielman (2009).*

2005; Leek and Storey, 2007; Listgarten et al., 2010; Fu et al., 2012). Throughout this thesis we use the terms *cis* and *trans* to define a local and a distant eQTL, respectively.

Cis-eQTLs

On average, *cis*-eQTLs (see the dots on the diagonal of Figure 3.5) explain a higher percentage of the gene expression variability than *trans*-eQTLs (Schadt et al., 2003). A plausible explanation is that *cis*-eQTLs have, in general, more direct mechanisms of regulation than those located in a different region. As a consequence, we usually find that *cis*-eQTLs have stronger associations than *trans*-eQTLs (Schadt et al., 2003; Cheung and Spielman, 2009; Listgarten et al., 2010) so that the power to detect local linkages is higher and they are more likely to be detected in data sets with smaller sample sizes. Further, *cis*-eQTLs provide immediate hints to what the causal gene is (Ronald et al., 2005). Nevertheless, the number of *cis*-associations found in eQTL studies depends on the size of the *cis*-window which can vary from 10,000 base pairs (10kb) in yeast (Brem et al., 2002) to 500kb in human (Kang

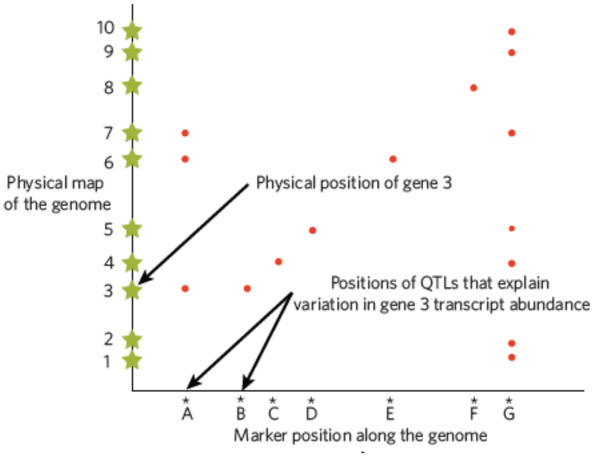

Figure 3.5: eQTL dot plot. In this figure, the y-axis represents the physical position in the genome of the gene corresponding to each transcript and the x-axis represents the physical position of the genetic markers. Red dots lying on the diagonal correspond to *cis*-associations whereas those lying outside the diagonal correspond to *trans*-associations. An eQTL hotspot is represented by the vertical band on the right. *Taken from Rockman (2008).*

et al., 2008a; Listgarten et al., 2010). Still, the maximum number of *cis*-acting eQTL grows linearly in the number of gene expression profiles whereas there is a combinatorial number of *trans*-eQTLs. Overall, it can be stated that *cis*-eQTLs are more easily detectable than *trans*-eQTLs (Schadt et al., 2003).

Trans-eQTLs

On the other hand, *trans*-acting eQTLs (see the dots outside the diagonal of Figure 3.5) have proven to be crucial to our understanding of complex regulatory mechanisms. For instance, locus control regions (Li et al., 2002) enhance the expression of distal genes under tissue specific conditions such as the one affecting the human β-globin locus (Grosveld et al., 1987). *Trans*-acting mechanisms have become more relevant in a recent contribution (Westra et al., 2013) in which they identified 233 SNPs *trans*-associated to different complex traits. In particular, some of these *trans*-eQTLs affect genes known to be altered in individuals with diseases such as systemic lupus erythematosus or type 1 diabetes.

However, *trans*-acting variants are, in general, difficult to identify. A first reason has to do with the fact that there are not prior clues about the location of *trans*-eQTLs, unless additional information is used. Therefore, markers covering the entire genome have to be tested for linkage to the target gene. A second reason is that *trans*-acting eQTLs normally have small effects on the target genes (Cheung and Spielman, 2009). This is probably because many *trans*-eQTLs affect the same gene and the effect of each genetic variant is small so that large sample sizes are required to increase the power and, consequently, be able to detect these small effects (Cheung and Spielman, 2009). On the other hand, it is difficult to find true direct *trans*-associations since small additive effects may also result from the propagation of large effects through the underlying correlated structure of genes, the effect of confounding factors and selection bias. Hence, although the number of *trans*-eQTLs identified in a typical eQTL study is usually higher than that of *cis*-eQTLs (Yvert et al., 2003), a large number of these *trans* associations may be spurious if not properly addressed (Section 3.3).

Not only the mapping of non-spurious *trans*-eQTLs onto the genome is problematic, but also the identification of the causal sequence is, in general, complicated. Mechanisms of *trans*-regulation are complex and not well understood. For example, one might think that known regulators of gene expression such as transcription factors or signalling molecules could be more involved in *trans*-eQTLs due to the direct mechanism of regulation between the regulator and the target genes. However, in a study conducted in yeast, Yvert et al. (2003) found that *trans*-eQTLs were not enriched for transcription factors showing, once again, the challenges of mapping *trans*-eQTLs.

eQTL hotspots

eQTL hotsposts are regions of the genome that are associated to a high number of correlated transcripts (Breitling et al., 2008). They appear as vertical bands such as the one in Figure 3.5.

eQTL hotspots may arise as a result of the presence of some confounding effects in the correlation structure of gene expression which lead to spurious associations (Li and Burmeister, 2005) between the genes if these artifacts are not controlled. Strategies such as the ones proposed by Kang et al. (2008b), Leek and Storey (2007) or Listgarten et al.

(2010) show that, after correcting for the confounding effects, many eQTL hotspots disappear.

On the contrary, some studies (Wessel et al., 2007; Kang et al., 2008a; Curtis et al., 2013) prove that, even after the adjustment for the confounding effects, a number of significant eQTL hotspots are still detected, suggesting their biological origin. These eQTL hotspots are often called master regulators. For example, when a non-synonymous change is found in the coding region of a transcription factor, the amino acid change may lead to a different regulation of their target genes if this change affects, for instance, the DNA binding domain. Further, those transcripts that show a *cis*-linkage with the hotspot might lead directly to the causative gene underlying the regulation of all the other transcripts. Nevertheless, Breitling et al. (2008) suggest that this regulatory nature of eQTL hotspots may be explained by the fact that if one gene, co-expressed with many other genes, shows a spurious eQTL association, the rest of co-expressed genes may also show these eQTL associations. Therefore, the interpretation of eQTL hotspots should be done with caution.

There are two good measures of the statistical power to identify eQTL associations due to true genetic effects and, in particular, those involving an eQTL hotspot. One is the enrichment in *cis*-eQTLs after correcting for the confounding effects and the other is the concordance of association between biologically replicated data sets (Kang et al., 2008b).

Multiple genetic variants regulate gene expression

The regulation of the mRNA level of a gene may be dependent on many factors and, thus, potentially influenced by more than one genetic variant. However, a systematic search for multiple loci is difficult due to the statistical complexity implicit in the combinatorial calculations required for this task. Even so, some studies (Schadt et al., 2003; Brem and Kruglyak, 2005) have reported the existence of multiple loci mapping to the same gene although the methods employed were not designed for this purpose. In fact, Brem and Kruglyak (2005) show that only in 3% of the genes their expression is influenced by a single eQTL whereas more than a 50% are affected by more than 5 eQTLs.

3.5 Discussion

eQTL mapping constitutes a very challenging problem due to the multivariate nature of gene expression profiles and its corresponding correlated structure. Additionally, spurious associations may arise as a consequence of confounding effects (e.g., batch effects) underlying gene expression data; thereby complicating even more the identification of direct eQTL associations and, in particular, of *trans*-eQTLs. These spurious eQTL associations lead to dense eQTL networks and the underlying regulatory mechanisms of these associations are difficult to infer.

Therefore, univariate methods used in classical QTL mapping in which we consider one genetic marker at a time and test whether it is linked to the phenotype, are not able to distinguish between direct and indirect eQTL associations.

During the last decade, many approaches based on different statistical frameworks (linear model theory or Bayesian statistics) have been developed to address the challenges of eQTL mapping (Kang et al., 2008a; Kim and Xing, 2009; Listgarten et al., 2010; Liu et al., 2008; Parts et al., 2011; Zhu et al., 2004; Neto et al., 2010). While these models tackle in some way the challenges of eQTL mapping, their interpretation becomes harder with the increasing complexity of the underlying statistical principles in which many of them are based.

On the other hand, methods based on conditional dependences allow the assessment of the association while controlling for other factors of interest. These may include other transcripts or genetic markers in addition to known confounding effects. These approaches provide an easier statistical interpretation of the resulting eQTL associations and constitute a suitable strategy to identify non-spurious eQTL associations. However, currently available methods (Edwards et al., 2010; Bing and Hoeschele, 2005; Chen et al., 2007; Neto et al., 2008; Kang et al., 2010) are restricted to conditional independence tests of order one; thereby limiting the ability to detect spurious eQTL associations that are not controlled by conditioning on just one other factor.

For this reason, further development of methods based on conditional dependences of higher-order is required in order to identify non-spurious eQTL associations with an easier statistical interpretation. Such strategies would be able to infer sparser eQTL networks that allow a better dissection of the underlying regulatory mechanisms of gene expression profiling.

4. Mixed graphical Markov models of eQTL networks

4.1 Introduction

Gene expression is a high-dimensional multivariate vector resulting from the simultaneous measurement of thousands of genes. The expression of these genes is regulated in a coordinated way by a wide range of complex molecular mechanisms so that gene expression profiles can be highly correlated. To infer an eQTL network that captures this correlated structure, it makes sense to study the joint distribution of all gene expression profiles rather than the marginal distribution of each gene.

However, the high-dimension of gene expression profiles increases the complexity of its joint analysis unless some type of structure simplifies the joint distribution of this multivariate vector. To this purpose, we use graphical Markov models (Whittaker, 1990; Cox and Wermuth, 1996; Lauritzen, 1996; Edwards, 2000) which are probabilistic models that are based on the statistical concept of conditional independence (see below). Concretely, a GMM is a family of probability distributions sharing conditional independence restrictions that are represented by means of a graph. Different types of graphs, such as undirected graphs or acyclic digraphs (also known as DAGs or Bayesian networks) determine distinct classes of GMMs.

Learning the structure of a graph from data where the sample size n is much larger than the number of random variables p, is a well-studied problem (Lauritzen, 1996; Chickering, 2002; Castelo and Kočka, 2003). On the other hand, during the last decade, technological advances in the instrumentation employed in the biomedical field have enabled the recording of large data sets that form multivariate samples of n observations through a typically much larger number p of random variables, i.e., $p \gg n$. In this setting, traditional assumptions underlying learning procedures do not hold and a substantial amount of work in GMM research has been devoted to the problem of learning the structure of the graph from data with $p \gg n$. Most of these contributions have been developed for Gaussian GMMs learned from pure continuous (multivariate normal) data. These approaches can be broadly categorized in regularization techniques (Friedman et al., 2008), dimension-reduction procedures (Segal et al., 2006) and limited-order correlations (Castelo and Roverato, 2006).

However, when dealing with data composed of mixed discrete and continuous random variables, such as genetical genomics data, the theory of mixed GMMs (Lauritzen, 1996; Edwards, 2000) has not been yet fully exploited. A first evidence is that existing learning procedures based on conditional independence tests are of order one (e.g., Bing and Hoeschele, 2005; Kang et al., 2010; Edwards et al., 2010).

Moreover, the underpinnings of eQTL networks, such as the propagation of additive effects through the gene network, are not well understood by the genetics community. Yet, a useful way to understand how additive effects from genetic variants propagate to gene expression and continuous phenotypes, is to simulate mixed GMMs and simulate data from these models. The simulation of data from mixed GMMs is also useful since the assessment of GMM learning procedures on simulated data is essential to verify the correctness or the asymptotic behavior of learning procedures.

In the first sections of this chapter we review the theory of GMMs that will be used in the rest of the thesis. Next, we describe how to simulate the parameters of classical mixed GMMs and how to simulate data sets from them to recreate biological features from eQTL networks, such as pleiotropic effects (Section 4.5). In Section 4.6, we simulate data from mixed GMMs to show how additive effects from genetic variants propagate to gene expression and continuous phenotypes. Finally, in Section 4.7, we use these simulations to show which is the estimated additive effect of the eQTLs under the effect of selection bias (Chapter 2, pg. 35).

4.2 Preliminaries

Marginal and conditional independence

In this section, we define the concepts of marginal independence and conditional independence.

Definition 1. Let X, Y be two random variables (r.v.'s) that admit a joint density function $f_{XY}(x, y)$. We assume that all functions on a discrete space are considered continuous functions. X and Y are said to be *marginally independent*, noted as $X \perp\!\!\!\perp Y$, if their joint density function factorizes into the product of their marginal density functions for all x and y:

$$f_{XY}(x, y) = f_X(x) f_Y(y)\,.$$

Equivalently, X and Y are marginally independent if the conditional density function of X given $Y = y$ does not depend on y:

$$X \perp\!\!\!\perp Y \iff f_{Y|X}(y|x) = f_Y(y)\,,$$

and, further,

$$X \perp\!\!\!\perp Y \iff f_{X|Y}(x|y) = f_X(x)\,.$$

Definition 2. Let X, Y, Z be three r.v.'s with a joint density function $f_{XYZ}(x, y, z)$. We say that X and Y are *conditionally independent* given Z, noted as

$$X \perp\!\!\!\perp Y | Z\,,$$

if, for each value z such that $f_Z(z) > 0$, $f_{XY|Z}(x, y|z)$ can be factorized as:

$$f_{XY|Z}(x, y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z)\,,$$

for all values x, y. Equivalently,

$$\begin{aligned}
X \perp\!\!\!\perp Y | Z &\iff f_{XYZ}(x, y, z) = f_{XZ}(x, z) f_{YZ}(y, z) / f_Z(z)\,,\\
X \perp\!\!\!\perp Y | Z &\iff f_{X|YZ}(x|y, z) = f_{X|Z}(x|z)\,,\\
X \perp\!\!\!\perp Y | Z &\iff f_{XYZ}(x, y, z) = h_{XZ}(x, z) k_{YZ}(y, z) \text{ for some functions } h,\ k\,,\\
X \perp\!\!\!\perp Y | Z &\iff f_{XYZ}(x, y, z) = f_{X|Z}(x|z) f_{Y|Z}(y|z) f_Z(z)\,.
\end{aligned}$$

Graph theory

Here, we introduce some basic concepts of graph theory.

Definition 3. An *undirected graph* G is defined by a pair of sets $G = (V, E)$ where V denotes the set of p vertices and E denotes the edge set $E \subseteq V \times V$. The edges of an undirected graph are represented by lines.

Definition 4. Two vertices α, $\beta \in V$ are *adjacent* if there exists an edge between them in G.

Definition 5. A graph $G = (V, E)$ is *complete* if every pair of vertices α, $\beta \in V$ are adjacent in G, that is, $(\alpha, \beta) \in E$.

Definition 6. A *path* of length m from the vertex α to the vertex β is a sequence $\alpha = \alpha_1 = \alpha_2 = \ldots = \alpha_m = \beta$ of distinct vertices such that $(\alpha_i, \alpha_{i+1}) \in E$ for all $i = 1, \ldots, m - 1$. A *graph cycle* is a path such that $\alpha = \beta$.

Definition 7. A *chordless cycle* of a graph $G = (V, E)$ is a graph cycle of length at least four that has no *cycle chords*. A *chord* is an edge between two non-consecutive vertices of a cycle.

We have illustrated a chordless cycle in the graph on the left of Figure 4.1.

Definition 8. Let α and β be two vertices of an undirected graph $G = (V, E)$. A subset $C \subseteq V$ is said to be an $(\alpha, \beta)-$*separator* if all paths from α to β intersect C.

Definition 9. Let be A, B, C three disjoint subsets of V. C is said to *separate* A from B if it is an $(\alpha, \beta)-$separator for every $\alpha \in A$, $\beta \in B$.

Graphs in Figure 4.1 illustrate the concept of separation.

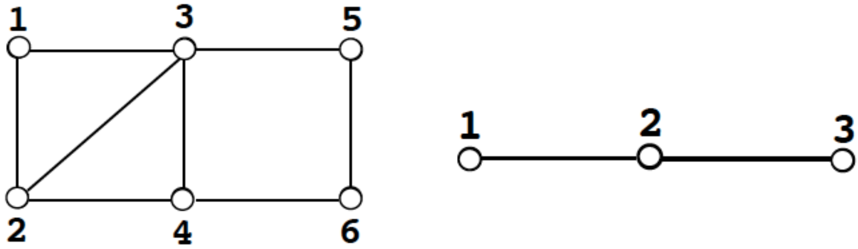

Figure 4.1: Chordless cycle and separation. On the left, the cycle formed by vertices 3, 5, 6 and 4 is a chordless cycle. By contrast, the edge between vertices 2 and 3 is a chord of the cycle formed by the vertices 1, 3, 4 and 2. On the other hand, the set $C = \{3, 4\}$ separates $A = \{1, 2\}$ from $B = \{5, 6\}$. On the right, vertex 2 is a $(1, 3)-$separator.

4.3 Undirected Gaussian GMMs

Undirected GMMs (Lauritzen, 1996) are statistical models representing probability distributions P_V involving a vector $X_V = \{X_1, \ldots, X_p\}$ of p

r.v.'s indexed by V. Undirected GMMs are determined by undirected graphs $G = (V, E)$ such that there is one to one correspondence between the vertices of G and the variables in X_V.

The connection between the conditional independence restrictions of the r.v.'s in X_V and their graphical representation by means of an undirected graph $G = (V, E)$ is given by the following property.

A probability distribution P_V associated to a GMM with an undirected graph G is said to be *Markov* with respect to G if it holds that $\forall$ $\alpha, \beta \in V$, such that α and β are not adjacent in G (i.e., $(\alpha, \beta) \notin E$) the corresponding r.v.'s X_α and X_β are conditionally independent given the rest of the variables,

$$X_\alpha \perp\!\!\!\perp X_\beta | X_{V \setminus \{\alpha,\beta\}} \, .$$

This is the *pairwise Markov* property. If P_V is positive and continuous, the pairwise Markov property is equivalent to the *global Markov* property (Lauritzen, 1996). The global Markov property states that for disjoint subsets $A, B, C \subseteq V$ we have that $X_A \perp\!\!\!\perp X_B | X_C$ whenever C separates A from B in G.

Moreover, a probability distribution P_V is *faithful* to G if all the conditional independence relationships in P_V can be read off the undirected graph G through the Markov property.

An undirected *Gaussian* GMM is the family of multivariate normal distributions that are Markov with respect to a given undirected graph $G = (V, E)$. Let X_V be a multivariate vector of p continuous r.v.'s that follow a p-variate Gaussian distribution $\mathcal{N}_p(\mu, \Sigma)$ with mean vector μ and positive definite covariance matrix Σ. Assume that the inverse of the covariance matrix $K = \Sigma^{-1}$, called the *concentration* matrix, is well defined. In this case, the connection between conditional independence and the multivariate normal distribution is reflected in the concentration matrix K.

In particular, we have that given two r.v.'s $X_\alpha, X_\beta \in X_V$, they are conditionally independent given the rest of r.v.'s, $X_\alpha \perp\!\!\!\perp X_\beta | X_{V \setminus \{\alpha,\beta\}}$, if and only if the corresponding partial correlation coefficient is zero:

$$\rho_{\alpha\beta . V \setminus \{\alpha,\beta\}} = \frac{-k_{\alpha\beta}}{\sqrt{k_{\alpha\alpha} k_{\beta\beta}}} = 0 \, ,$$

where $K = \{k_{ij}\}_{i,j \in V} = \Sigma^{-1}$. Concretely, Lauritzen (1996, Prop. 5.2) shows that

$$\rho_{\alpha\beta.V\setminus\{\alpha,\beta\}} = 0 \iff k_{\alpha\beta} = 0 \iff (\alpha, \beta) \notin E\,. \qquad (4.1)$$

Thus, two r.v.'s are independent given the remaining ones if and only if the corresponding element of the inverse covariance matrix is zero, and therefore, the structure of $G = (V, E)$ can be derived from the zero pattern of K.

An important subclass of Gaussian GMMs are *decomposable* Gaussian GMMs. These models are defined by decomposable undirected graphs:

Definition 10. An undirected graph $G = (V, E)$ is *decomposable* if it is complete, or there exist three disjoint sets (A, B, C) of V such that

- $V = A \cup B \cup C$,
- A and B are non-empty,
- C separates A from B,
- C is complete in G,
- $A \cup C$ and $B \cup C$ are also decomposable.

In such case, (A, B, C) form a *decomposition* of the undirected graph $G = (V, E)$. Equivalently, G is decomposable if it does not contain chordless cycles of length greater than 3, that is, G is *chordal* or *triangulated.*

4.4 Mixed GMMs

Mixed GMMs (Lauritzen and Wermuth, 1989) are statistical models representing probability distributions involving discrete r.v.'s, denoted by I_δ with $\delta \in \Delta$, and continuous r.v.'s, denoted by Y_γ with $\gamma \in \Gamma$. This class of GMMs are determined by *undirected marked graphs* $G = (V, E)$ with p marked vertices grouped into two subsets $V = \Delta \cup \Gamma$, and with edge set $E \subseteq V \times V$. Vertices $\delta \in \Delta$ are depicted by solid circles, $\gamma \in \Gamma$ by open ones and the entire set of them, V, index a vector of r.v.'s $X = (I, Y)$. In our context, continuous r.v.'s Y correspond to genes and discrete r.v.'s I to markers or eQTLs; see Figure 4.2 for a graphical representation of one such mixed GMM. We denote the joint sample space of X by:

$$x = (i, y) = \{(i_\delta)_{\delta \in \Delta}, (y_\gamma)_{\gamma \in \Gamma}\} ,$$

where i_δ are discrete values corresponding to genotype alleles from the marker or eQTL I_δ and y_γ are continuous values of the expression from gene Y_γ. The set of all possible joint discrete levels i is denoted as $\mathcal{I}$. Following Lauritzen and Wermuth (1989), we assume that the joint distribution of the variables X is *conditional Gaussian* (also known as CG-distribution) with density function:

$$f(x) = f(i, y) = p(i)|2\pi\Sigma(i)|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(y - \mu(i))^T \Sigma(i)^{-1}(y - \mu(i))\right\} .$$

This distribution has the property that continuous variables follow a multivariate normal distribution $\mathcal{N}_{|\Gamma|}(\mu(i), \Sigma(i))$ conditioned on the discrete variables. The parameters $(p(i), \mu(i), \Sigma(i))$ are called *moment characteristics* where $p(i)$ is the probability that $I = i$, and $\mu(i)$ and $\Sigma(i)$ are the conditional mean and the covariance matrix of Y which may depend on i. If the covariance matrix is constant across the levels of $\mathcal{I}$, that is, $\Sigma(i) \equiv \Sigma$, the model is *homogeneous*. Otherwise, the model is said to be *heterogeneous*.

Canonical parameters and conditional Gaussian interactions

We can write the logarithm of the density in terms of the *canonical parameters* $(g(i), h(i), K(i))$:

$$\log f(i, y) = g(i) + h(i)^T y - \frac{1}{2} y^T K(i) y \,, \tag{4.2}$$

where

$$\begin{aligned} g(i) &= \log(p(i)) - \frac{1}{2}\log|\Sigma(i)| - \frac{1}{2}\mu(i)^T \Sigma(i)^{-1}\mu(i) - \frac{|\Gamma|}{2}\log(2\pi) \,, \\ h(i) &= \Sigma(i)^{-1}\mu(i) \,, \qquad (4.3) \\ K(i) &= \Sigma(i)^{-1} \,. \end{aligned}$$

CG-distributions satisfy the Markov property if and only if their canonical parameters are expanded into interaction terms such that only those interactions among adjacent vertices are present (Lauritzen, 1996). Therefore, these terms are expanded as follows:

$$g(i) = \sum_{d \subseteq \Delta} \lambda_d(i), \quad h_\gamma(i) = \sum_{d \subseteq \Delta} \eta_d(i)_\gamma, \quad k_{\gamma\eta}(i) = \sum_{d \subseteq \Delta} \psi_d(i)_{\gamma\eta} \, , \tag{4.4}$$

where the interaction terms are described as:

- $\lambda_d(i)$, with $d \subseteq \Delta$ complete in G, are the discrete interactions among the variables indexed by d. If $|d| = 1$, the term is called main effect of the variable in d. If $d = \emptyset$, the term $\lambda_\emptyset$ is constant.
- $\eta_d(i)_\gamma$, with $d \cup \{\gamma\}$ complete in G, represent mixed linear interactions between X_γ and the variables indexed by d in Equation (4.2). If $d = \emptyset$, the term $\eta_{\emptyset\gamma}$ is called linear main effect of the variable X_γ.
- $\psi_d(i)_{\gamma\eta}$, with $d \cup \{\gamma, \eta\}$ complete in G, represent quadratic interactions between X_γ, X_η and the variables indexed by d in Equation (4.2). If $\gamma = \eta$ and $d = \emptyset$, we refer to quadratic main effects. If the model is homogeneous, there are not mixed quadratic interactions, i.e., $\psi_d(i)_{\gamma\eta} = 0$ for $d \neq \emptyset$.

Plugging these expansions in Equation (4.2), we obtain

$$\log f(i, y) = \sum_{d \subseteq \Delta} \lambda_d(i) + \sum_{d \subseteq \Delta} \sum_{\gamma \in \Gamma} \eta_d(i)_\gamma y_\gamma - \frac{1}{2} \sum_{d \subseteq \Delta} \sum_{\gamma, \eta \in \Gamma} \psi_d(i)_{\gamma\eta} y_\gamma y_\eta \, , \tag{4.5}$$

where $\lambda_d(i) = 0$ unless d is complete in G, $\eta_d(i)_\gamma = 0$ unless $d \cup \{\gamma\}$ is complete in G, and $\psi_d(i)_{\gamma\eta} = 0$ unless $d \cup \{\gamma, \eta\}$ is complete in G.

Decomposable mixed GMMs

An important subclass of mixed GMMs is defined by decomposable marked graphs (Figure 4.2).

Definition 11. A triple (A, B, C) of disjoint subsets of V forms a *decomposition* of an undirected marked graph G if $V = A \cup B \cup C$ and

- C is a complete subset of V,
- C separates A from B and,
- $C \subseteq \Delta$ or $B \subseteq \Gamma$.

Thus, an undirected marked graph G is *decomposable* if there exists a proper decomposition (A, B, C) such that the subgraphs $G_{A\cup C}$ and $G_{B\cup C}$ are decomposable. When this holds, $C_1 = \{A \cup C\}$ and $C_2 = \{B \cup C\}$ induce complete subgraphs, called *cliques*, and $S = \{C\}$, also inducing a complete subgraph, is called a *separator* of G.

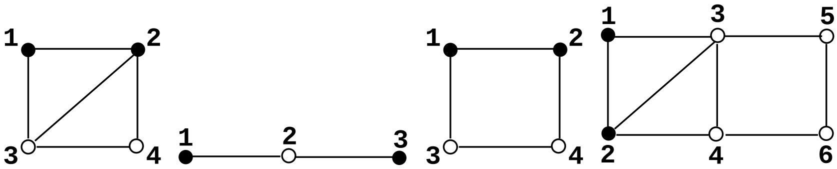

Figure 4.2: Examples of graphs representing mixed GMMs. The graph depicted on the left is decomposable whereas the other three are non-decomposable.

Analogously, when G is undirected, decomposability of G holds if and only if G does not contain chordless cycles of length larger than 3 and does not contain any path between two non-adjacent discrete vertices passing through continuous vertices only.

Maximum Likelihood Estimates of mixed GMMs

Let $\mathcal{X} = \{x^{(\nu)}\} = \{(i^{(\nu)}, y^{(\nu)})\}$ be a sample of $\nu = 1, ..., n$ independent and identically distributed observations from a CG-distribution. For an arbitrary subset $A \subseteq V$, we abbreviate to $i_A = i_{A\cap\Delta}$, $\mathcal{I}_A = \mathcal{I}_{\Delta\cap A}$ and

$y_A = y_{A\cap\Gamma}$ and the following sampling statistics are defined:

$$
\begin{aligned}
n(i) &= \#\left\{\nu : i^{(\nu)} = i\right\} && (4.6)\\
s(i) &= \sum_{\nu : i^{(\nu)} = i} y^{(\nu)} && (4.7)\\
\bar{y}(i) &= s(i)/n(i) && (4.8)\\
ss(i) &= \sum_{\nu : i^{(\nu)} = i} y^{(\nu)} (y^{(\nu)})^T && (4.9)\\
ssd(i) &= ss(i) - s(i)s(i)^T/n(i) && (4.10)\\
ssd_A(A) &= \sum_{i_A \in \mathcal{I}_A} ssd_{A\cap\Gamma}(i_A) && (4.11)\\
ssd_A &= ssd_A(V) && (4.12)\\
ssd &= ssd(V) && (4.13)
\end{aligned}
$$

Lauritzen (1996, Prop. 6.9) shows that the likelihood function for the heterogeneous, saturated model attains its maximum if and only if $ssd(i)$ is positive definite for all $i \in \mathcal{I}$, which is almost surely equal to the event that $n(i) > |\Gamma|$ for all $i \in \mathcal{I}$. If the maximum likelihood estimate (MLE) exists, it is determined as having moment characteristics equal to the empirical moments, i.e.,

$$\hat{p}(i) = n(i)/n, \quad \hat{\mu}(i) = \bar{y}(i), \quad \hat{\Sigma}(i) = ssd(i)/n(i)\,. \tag{4.14}$$

This sample size constraint is milder with the homogeneous, saturated model whose likelihood function attains its maximum if and only if $n(i) > 0$ for all $i \in \mathcal{I}$ and ssd is positive definite (Lauritzen, 1996, Prop. 6.10), which cannot occur whenever $n < |\Gamma| + |\mathcal{I}|$. If $n \geq |\Gamma| + |\mathcal{I}|$, this is almost surely equal to the event that $n(i) > 0$ for all $i \in \mathcal{I}$. In such a case, the MLEs of the moment characteristics are

$$\hat{p}(i) = n(i)/n, \quad \hat{\mu}(i) = \bar{y}(i), \quad \hat{\Sigma} = ssd/n\,. \tag{4.15}$$

However, in this case, it follows that saturated mixed GMMs cannot be directly estimated from data with $p \gg n$, using only the formulae described above.

For the unsaturated case, decomposable mixed GMMs also admit explicit MLEs. In the homogeneous case, Lauritzen (1996, Prop. 6.21) shows that the MLE exists almost surely if and only if $n(i_C) \geq |C \cap \Gamma| +$

$|\mathcal{I}_C|$ for all cliques C of G and $i_C \in \mathcal{I}_C$. In this case, it is given with the following canonical parameters (Lauritzen, 1996, pg. 189):

$$\hat{p}(i) = \prod_{j=1}^{k} \frac{n(i_{C_j})}{n(i_{S_j})}, \tag{4.16}$$

$$\hat{h}(i) = n\left\{ \sum_{j=1}^{k} \left[ssd_{C_j}(C_j)^{-1}\bar{y}_{C_j}(i_{C_j})\right]^{|\Gamma|} - \left[ssd_{S_j}(S_j)^{-1}\bar{y}_{S_j}(i_{S_j})\right]^{|\Gamma|} \right\}, \tag{4.17}$$

$$\hat{K} = n\left\{ \sum_{j=1}^{k} \left[ssd_{C_j}(C_j)^{-1}\right]^{|\Gamma|} - \left[ssd_{S_j}(S_j)^{-1}\right]^{|\Gamma|} \right\}, \tag{4.18}$$

where $S_1 = \emptyset$.

The matrices $[M]^{|\Gamma|}$ of Equation (4.18) are defined as follows: given a matrix $M = \{m_{\gamma\eta}\}_{|A|\times|A|}$ of dimension $|A| \times |A|$ with $A \subseteq \Gamma$, $[M]^{|\Gamma|}$ is a $|\Gamma| \times |\Gamma|$ matrix such that

$$[M]^{|\Gamma|}_{\gamma\eta} = \begin{cases} m_{\gamma\eta}, & \text{if } \{\gamma, \eta\} \in A. \\ 0, & \text{otherwise.} \end{cases}$$

Analogously, in Equation (4.17), $[M]^{|\Gamma|}$ is a $|\Gamma|$-length vector obtained from a $|A|$-length vector M.

4.5 Simulation of eQTL networks with mixed GMMs

Disentangling direct and indirect associations of genes and genetic variants through gene expression and natural variation requires some underlying model of the eQTL network. Here, we assume that gene expression forms a p-multivariate sample following a CG-distribution given the joint probability of all genetic variants so that a sensible model of an eQTL network is a mixed GMM. A mixed GMM enables the integration of discrete genotypes with continuous expression measurements as a multivariate statistical model satisfying a set of restrictions of conditional independence encoded by means of a graph. To gather insight into how this model represents the underlying associations between genetic variants and continuous genes and phenotypes, in this section we describe how to simulate eQTL networks with homogeneous mixed GMMs and data from them. In particular, we restrict ourselves to the case of a backcross, in which each marker can have two different genotypic values, but these simulations could be extended to other cross models allowing for other than linear additive effects (codominant model) on the mixed associations, such as dominance effects. As shown in the next two sections, such an exercise enables gathering a deeper understanding into the flow of genetic additive effects arising from eQTLs and propagating through the gene network under this type of statistical models.

The procedures described in this section are all implemented as part of the functionalities of the R/Bioconductor package called `qpgraph`.

Simulation of undirected graphs

The first step to simulate a GMM consists of simulating its associated graph $G = (V, E)$ which, in this case, defines the structure of the eQTL network. Discrete r.v.'s I, associated to discrete vertices Δ correspond to eQTLs and continuous r.v.'s Y associated to continuous vertices Γ to expression profiles, such that $V = \Delta \cup \Gamma$ and $|V| = p$.

In the context of genetical genomics data we make the assumption that discrete genotypes affect continuous gene expression measurements and not the other way around. Thus, we consider the underlying graph G as a marked graph with mixed edges, where some are directed and represented by arrows and some are undirected. More concretely, G will have arrows pointing from discrete vertices to continuous ones and undirected edges between continuous vertices. From this restriction, it follows immediately that there are no semi-directed cycles and allows one to interpret these GMMs as *chain graphs*. Chain graphs are graphs formed by undirected subgraphs connected by directed edges (Lauritzen, 1996, pg. 7).

One of the basic assumptions made by procedures that estimate the structure of GMMs when $p \gg n$ is that the underlying graph structure is sparse. In the present context, this means that the number of present eQTL and gene-gene associations is much smaller than the total possible number of them. Therefore, to explore the performance of estimation procedures, we are interested in sampling graphs with a fine-tune control of the level of sparseness. In our case, we sample the undirected subgraph that defines the gene network from the subclass of undirected d-regular graphs (Harary, 1969). These are graphs with a constant vertex degree d (see Figure 4.3) which makes the graph density D a linear function of d:

$$D = \frac{d}{p_\Gamma - 1},$$

where $p_\Gamma = |\Gamma|$ is the number of genes and $p_\Gamma \cdot d$ is even. The constant degree d also bounds the size of any minimal subset separating every pair of vertices (Castelo and Roverato, 2006, pg. 2646).

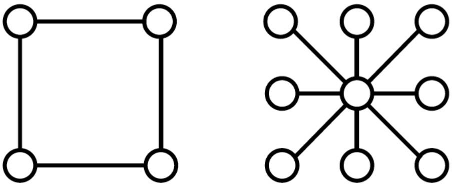

Figure 4.3: A d-regular graph with vertex degree $d = 2$ is shown on the left. On the right, the star graph is not d-regular since the vertex in the middle has degree $d = 8$ while the other vertices have degree $d = 1$.

Finally, we restrict eQTL relationships, which correspond to mixed edges, to at most one per gene.

Simulation of parameters of a homogeneous mixed GMM

After simulating the underlying graph structure G, we need to simulate the parameters of the CG-distribution represented by G (a conditional covariance matrix Σ, a conditional mean vector $\mu(i)$ and a vector of probabilities $p(i)$ corresponding to discrete levels) with given marginal linear correlations on the pure continuous (gene-gene) associations and given additive effects on the mixed (marker/eQTL-gene) ones. We simulate *homogeneous* mixed GMMs. In the context of genetical genomics data, this assumption implies that genotype affects only the mean expression level of genes and not the correlations between them.

Conditional covariance matrix:

Once the structure of the graph G is obtained, a random homogeneous conditional covariance matrix Σ is generated as follows. Let $G_\Gamma \subseteq G$

denote the subgraph of $p_\Gamma = |\Gamma|$ pure continuous vertices and let ρ denote the desired mean marginal correlation between each pair of continuous r.v.'s (X_γ, X_η) such that $(\gamma, \eta) \in G_\Gamma$. Let $\mathcal{S}^+(G_\Gamma) \subset \mathcal{S}^+$ denote the set of all $p_\Gamma \times p_\Gamma$ positive definite matrices in $\mathcal{S}^+$ such that every matrix $S \in \mathcal{S}^+(G_\Gamma)$ satisfies that $\{S^{-1}\}_{ij} = 0$ whenever $i \neq j$ and $(i,j) \notin G_\Gamma$.

We simulate Σ such that $\Sigma \in \mathcal{S}^+(G_\Gamma)$ in two steps. First, we build an initial *incomplete matrix* Σ_0 with elements $\{\sigma^0_{ij}\}$ if either $i = j$ or $(i,j) \in G_\Gamma$, and the remaining elements unspecified. Second, we search for a *positive completion* of Σ_0, which consists of filling up Σ_0 in such a way that the resulting $\Sigma \in \mathcal{S}^+(G_\Gamma)$.

It can be shown (Grone et al., 1984) that if Σ_0 admits a positive completion, then it is unique and it belongs to $\mathcal{S}^+(G_\Gamma)$. This means that, given a suitable Σ_0, we can use algorithms for maximum likelihood estimation or Bayesian conjugate inference (Roverato, 2002) as matrix completion algorithms. To this end, we first draw Σ_0 from a Wishart distribution $W_{p_\Gamma}(\Lambda, p_\Gamma)$ with $\Lambda = DRD$, $D = \text{diag}(\{\sqrt{1/p_\Gamma}\}_{p_\Gamma})$ and $R = \{R_{ij}\}_{p_\Gamma \times p_\Gamma}$ where $R_{ij} = 1$ for $i = j$ and $R_{ij} = \rho$ for $i \neq j$. It is required that $\Lambda \in \mathcal{S}^+$ and this happens if and only if $-1/(p_\Gamma - 1) < \rho < 1$ (Seber, 2007, pg. 317). Finally, we apply an iterative regression procedure for maximum likelihood estimation of Gaussian graphical models with known structure (Hastie et al., 2009, pg. 634) as matrix completion algorithm to obtain Σ from the initially sampled Σ_0.

Probability of discrete levels:

Since we are simulating a backcross, each discrete r.v. takes two possible values $i_\delta = \{1, 2\}$ with equal probability:

$$p(i_\delta = 1) = \frac{1}{2} \quad \text{and} \quad p(i_\delta = 2) = \frac{1}{2} \; .$$

We assume that discrete r.v.'s representing eQTLs are marginally independent between them. Therefore, joint levels $i \in \mathcal{I}$ follow a uniform distribution, that is, $p(i) = 1/|\mathcal{I}| \; \forall i \in \mathcal{I}$.

Conditional mean vector:

The values of the mean vector $\mu(i)$ are determined from the strength of the mixed interactions between discrete and continuous r.v.'s. For this reason, we force each discrete variable I_δ to have an additive effect $a_{\delta\gamma}$ on the continuous variable Y_γ. For the case of a backcross, this implies that

$$a_{\delta\gamma} = \mu_\gamma(i_\delta = 1) - \mu_\gamma(i_\delta = 2) = \frac{\sum\limits_{i':i_\delta=1} p(i')\mu(i')}{\sum\limits_{i':i_\delta=1} p(i')} - \frac{\sum\limits_{i':i_\delta=2} p(i')\mu(i')}{\sum\limits_{i':i_\delta=2} p(i')}\,, \quad (4.19)$$

where the random vector $\mu(i)$, conditioned on the discrete levels $\mathcal{I}$, is generated from (4.3). Since $p(i) = 1/|\mathcal{I}|$, Equation (4.19) reduces to

$$a_{\delta\gamma} = \frac{1}{|\mathcal{I}|/2} \sum_{i':i_\delta=1} \mu(i') - \frac{1}{|\mathcal{I}|/2} \sum_{i':i_\delta=2} \mu(i')\,. \quad (4.20)$$

The values of the canonical parameter $h(i) = \{h_\gamma(i)\}_{\gamma\in\Gamma}$, determine the strength of the mixed interactions between discrete and continuous r.v.'s and they are generated as in (4.4). In particular,

$$h_\gamma(i) = \begin{cases} \eta_{\emptyset\gamma}, & \text{if } (\delta,\gamma) \notin E \ \forall\delta \in \Delta. \\ \{\eta_\delta(i_\delta)_\gamma\}_{i_\delta\in\mathcal{I}_\delta} = \{\eta_\delta(1)_\gamma, \eta_\delta(2)_\gamma\}, & \text{if } (\delta,\gamma) \in E,\ \delta \in \Delta. \end{cases}$$

Without loss of generality, the values $\eta_{\emptyset\gamma}$ are set to zero. To set the values $\eta_\delta(1)_\gamma$ and $\eta_\delta(2)_\gamma$ so that both (4.3) and (4.20) are satisfied we proceed as follows. Assume that a genotype represented by a r.v. I_δ has a pleiotropic effect (see Chapter 2, pg. 33) on a set of genes corresponding to r.v.'s $\{Y_\gamma\}_{\gamma\in A_\delta}$, where $A_\delta = \{\gamma \in \Gamma : (\delta,\gamma) \in E\}$. By combining (4.3) and (4.20), for each $\gamma \in A_\delta$, we have that

$$\begin{aligned} a_{\delta\gamma} &= \frac{1}{|\mathcal{I}|/2} \sum_{i':i_\delta=1} \sum_{\zeta\in\Gamma} \sigma_{\gamma\zeta} h_\zeta(i') - \frac{1}{|\mathcal{I}|/2} \sum_{i':i_\delta=2} \sum_{\zeta\in\Gamma} \sigma_{\gamma\zeta} h_\zeta(i') = \\ &= \frac{1}{|\mathcal{I}|/2} \sum_{\zeta\in\Gamma} \sigma_{\gamma\zeta} \left\{ \sum_{i':i_\delta=1} h_\zeta(i') - \sum_{i':i_\delta=2} h_\zeta(i') \right\}. \end{aligned}$$

It follows that all r.v.'s X_ζ such that $\zeta \notin A_\delta$ are not associated to I_δ, and therefore, if $j, k \in \mathcal{I}$ are two discrete levels such that $j_\delta = 1$, $k_\delta = 2$ and $j_{\Delta\setminus\{\delta\}} = k_{\Delta\setminus\{\delta\}}$, we have that $h_\zeta(j) = h_\zeta(k)$. Hence, for all $\zeta \notin A_\delta$, the terms $h_\zeta(i')$ from both summations cancel out and we obtain

$$\begin{aligned} a_{\delta\gamma} &= \frac{2}{|\mathcal{I}|} \sum_{\zeta\in A_\delta} \sigma_{\gamma\zeta} \left\{ \sum_{i':i_\delta=1} h_\zeta(i') - \sum_{i':i_\delta=2} h_\zeta(i') \right\} = \\ &= \frac{2}{|\mathcal{I}|} \sum_{\zeta\in A_\delta} \sigma_{\gamma\zeta} \left\{ \frac{|\mathcal{I}|}{2} \eta_\delta(1)_\gamma - \frac{|\mathcal{I}|}{2} \eta_\delta(2)_\gamma \right\} = \sum_{\zeta\in A_\delta} \sigma_{\gamma\zeta} \left\{ \eta_\delta(1)_\gamma - \eta_\delta(2)_\gamma \right\}. \end{aligned}$$

Let $\eta_{\delta\gamma} = \eta_\delta(1)_\gamma - \eta_\delta(2)_\gamma$ and the vectors $\boldsymbol{a}_{\delta A_\delta} = \{a_{\delta\gamma}\}_{\gamma \in A_\delta}$, $\boldsymbol{\eta}_{\delta A_\delta} = \{\eta_{\delta\gamma}\}_{\gamma \in A_\delta}$, $\boldsymbol{\eta}_{1 A_\delta} = \{\eta_\delta(1)_\gamma\}_{\gamma \in A_\delta}$ and $\boldsymbol{\eta}_{2 A_\delta} = \{\eta_\delta(2)_\gamma\}_{\gamma \in A_\delta}$. We write the matrix form of the previous expression as

$$\begin{aligned}
\boldsymbol{a}_{\delta A_\delta} &= \Sigma_{\{A_\delta, A_\delta\}} \boldsymbol{\eta}_{\delta A_\delta}\,; \\
\boldsymbol{\eta}_{\delta A_\delta} &= \Sigma^{-1}_{\{A_\delta, A_\delta\}} \boldsymbol{a}_{\delta A_\delta}\,; \\
\boldsymbol{\eta}_{1 A_\delta} &= \Sigma^{-1}_{\{A_\delta, A_\delta\}} \boldsymbol{a}_{\delta A_\delta} + \boldsymbol{\eta}_{2 A_\delta}\,.
\end{aligned}$$

Finally, once the values of $h_\gamma(i)$ are determined for each $\gamma \in \Gamma$, we use (4.3) to obtain the $\mu(i)$ values by:

$$\mu(i) = \Sigma \cdot h(i)\,.$$

Note that although we have previously simulated a covariance matrix independently from the discrete r.v.'s, here we interpret it as a conditional covariance given the levels of $\mathcal{I}$ to generate the mean vector $\mu(i)$.

Simulation of eQTL network models of experimental crosses

We have integrated the algorithms presented above with the functions provided by the `R/qtl` package (Broman et al., 2003) to simulate eQTL network models of experimental crosses and data from them in the following way:

- First, we simulate a genetic map with a given number of chromosomes and markers using the `sim.map()` function of the `R/qtl` package.

- Second, we simulate a homogeneous mixed GMM: we determine the underlying regulatory model of *cis*-eQTLs, *trans*-eQTLs and gene-gene associations. Then, we simulate the structure and the parameters of this homogeneous mixed GMM according to the previous procedures.

- Third, we simulate data from the previous eQTL network model with the function `sim.cross()` from the `R/qtl` package. This function is overloaded in the `qpgraph` to plug the eQTL associations into the corresponding genetic loci.

The function `sim.cross()` defined in the `qpgraph` package proceeds as follows. First, the genotype data is simulated by the procedures implemented in the `R/qtl` package. Genotypes are sampled at each marker from a Markov chain with transition probabilities that depend on the distance between markers and a mapping function (the Haldane function is used by default). eQTLs are placed at the markers and, in particular, if eQTLs are located sufficiently apart from each other, we can assume that the discrete r.v.'s are marginally independent between them. Finally, `qpgraph` adds the simulation of the continuous gene expression values according to the homogeneous mixed GMM by selecting the corresponding parameters of the homogeneous CG-distribution $\mathcal{N}_{|\Gamma|}(\mu(i), \Sigma)$, given a sampled genotype i from all eQTLs.

An eQTL network example

In this section we illustrate how to simulate an eQTL network with the package `qpgraph` in conjunction with the `R/qtl` package (Broman et al., 2003) by using the procedures described in the previous section.

```
library(qtl)
library(qpgraph)
```

In this example, we simulated a genetic map formed by one chromosome of 100 cM long and 10 markers (Figure 4.4a).

```
map <- sim.map(len=100, n.mar=10)
```

Then, we built an eQTL network (Figure 4.4b) of $p = 6$ genes where gene 4 has a *cis*-eQTL (see Figure 4.4a to see the location of the eQTL and the genes). The undirected graph representing the gene network is a d-regular graph with $d = 3$. We also simulated the parameters of a mixed GMM representing the eQTL network such that the correlation between the genes is set to $\rho = 0.5$ and the additive effect of the eQTL on gene 4 is $a = 1$.

```
set.seed(123)
eQTLnet <- reQTLcross(eQTLcrossParam(map=map,
                                     genes=6,
                                     cis=1/6,
                 networkParam=dRegularGraphParam(d=3)),
                      rho=0.5, a=1)
```

From this mixed GMM (`eQTLnet`) we simulated 100 observations where 6 of them are shown below.

```
cross <- sim.cross(map, eQTLnet, n.ind=100)
head(round(cbind(cross$qtlgeno, cross$pheno), digits=2))
```

```
  QTL1   g1   g2   g3   g4   g5   g6
1    2 2.92 3.20 5.58 3.38 6.79 6.09
2    2 3.61 3.37 4.77 2.66 7.10 6.80
3    1 3.38 2.05 3.40 1.12 6.11 5.58
4    1 2.96 2.82 4.23 1.60 5.77 5.67
5    1 4.02 3.42 5.34 2.41 7.94 6.90
6    1 3.42 2.63 3.90 1.10 5.61 5.53
```

Here, we show the parameters of the simulated mixed GMM. First, we show the inverse of the covariance matrix, $K = \Sigma^{-1}$. Note that the zero entries of this matrix correspond to missing edges in the eQTL network of Figure 4.4b.

```
round(solve(eQTLnet$model$sigma), digits=1)
```

```
     g1   g2   g3   g4   g5   g6
g1 11.9  0.0  0.0  5.5 -5.5 -2.5
g2  0.0 15.8 -2.1  0.0 -2.4 -3.2
g3  0.0 -2.1 10.0 -5.4 -3.8  0.0
g4  5.5  0.0 -5.4  7.1  0.0 -1.6
g5 -5.5 -2.4 -3.8  0.0  6.5  0.0
g6 -2.5 -3.2  0.0 -1.6  0.0  3.7
```

The following matrix represents the mean vector $\mu(i)$ of the 6 genes (columns) given the genotype of the eQTL (rows). As we enforced previously, the additive effect a of the fourth gene is 1.

```
eQTLnet$model$mean()
```

```
        g1       g2       g3       g4       g5       g6
1 3.107509 2.424557 3.529890 1.428412 5.702648 5.043203
2 2.960638 2.706197 4.281949 2.428412 6.118407 5.626964
```

Finally, the following matrix represents the canonical parameter $h(i)$. Since none of the genes except the fourth have an eQTL, their values are constant across the levels of the eQTL.

```
round(solve(eQTLnet$model$sigma) %*%
                    t(eQTLnet$model$mean()), digits=2)
```

```
  g1 g2 g3   g4 g5 g6
1  1  1  1 0.00  1  1
2  1  1  1 1.32  1  1
```

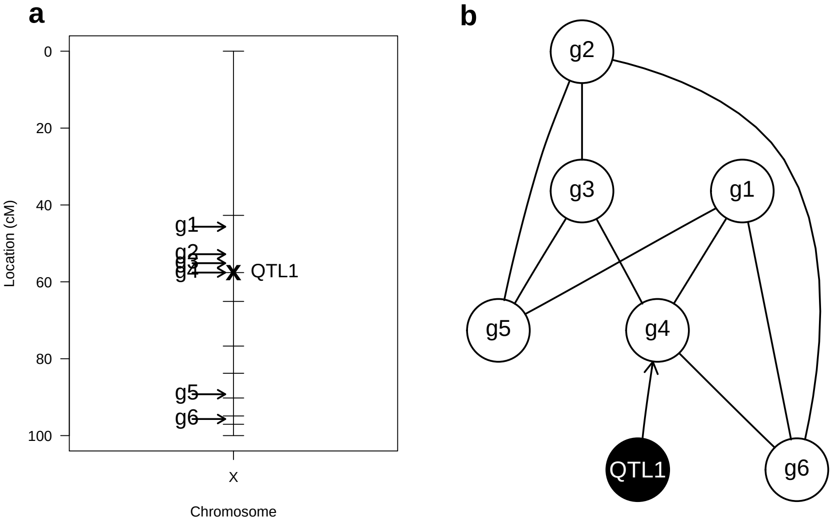

Figure 4.4: eQTL network example. Panel (a) shows a genetic map formed by one chromosome of 100 cM and the location of 10 simulated markers. Arrows indicate the location of the 6 genes that form the eQTL network depicted in panel (b). In this eQTL network, gene 4 has an eQTL which, as it is shown in (a), is a *cis*-eQTL since it is located at the same position of the gene.

4.6 Flow of genetic additive effects through gene expression

The purpose of this section is to gather insight into how mixed GMMs represent the underlying associations between genetic variants, genes and phenotypes in an eQTL network. To this end, we simulate eQTL models according to the procedures explained in the previous section.

Using the `R/qtl` package (Broman et al., 2003) we simulated a genetic map formed by one single chromosome 100 cM long and 10 equally-spaced markers. We built an eQTL network of $p = 5$ genes forming a chain, where the first of them had one eQTL placed randomly among the ten markers (Figure 4.5A). We simulated 10 mixed GMMs with the eQTL network structure shown in Figure 4.5A, under increasing values of the marginal correlation between the genes ($\rho = \{0.25, 0.5, 0.75\}$) and of the additive effects from the eQTL on gene labeled as "g1" ($a = \{0.5, 1, 2.5, 5\}$). We sampled 1,000 data sets from each of these 10 models. Each data set contained simulated genotypes and expression values for $n = 100$ individuals. We estimated the additive effect of the eQTL on each of the 5 genes, averaged over the 10,000 data sets at each combination of additive effect and marginal correlation. Note that only the additive effect on gene "g1" is direct (Figure 4.5A).

Grey lines in panels B-D of Figure 4.5 show the estimated average additive effects for the three different marginal correlation values on the gene-gene associations. These plots demonstrate that additive effects propagate as function of gene-gene correlations ρ. More concretely, when $\rho \geq 0.5$ moderate to large additive effects may easily show up as indirect eQTL associations when inspecting the margin of the data formed by one genetic variant and one gene expression profile.

4.7 Estimation of additive effects under selection bias

In the previous section we have shown how the additive effect of eQTLs propagates through the gene network as a function of their marginal correlation. In this section, we go one step further and show that the estimated effect of eQTLs often varies from its true effect, a phenomenon

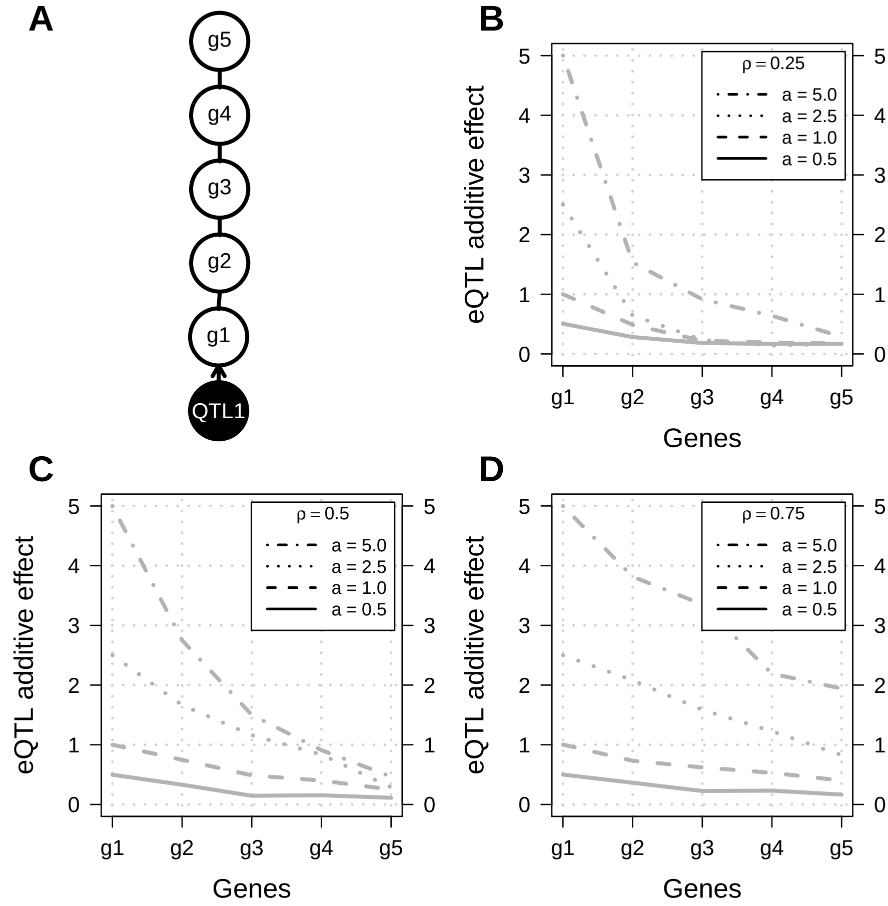

Figure 4.5: Propagation of indirect eQTL additive effects. Panel (A) shows the structure of the eQTL network underlying the mixed GMM employed to simulate the data shown on the other panels. All additive effects of the eQTL labeled “QTL1” on every gene g1, ..., g5, are indirect, except for g1. Panels (B) to (D) show average estimated additive effects of the eQTL on each gene across 10,000 data sets simulated from different combinations of nominal gene-gene correlations (ρ) and additive effects (a).

called selection bias (see Chapter 2, pg. 35).

To this end, we considered the data simulated in the previous section. In order to show the effect of selection bias, we estimated the additive effects of the eQTL only in those data sets for which we have an evidence of association between the gene and the eQTL. Concretely, we have calculated the LOD score associated to each gene and the eQTL using the single marker regression approach. Black lines in panels B, C and D of Figure 4.6 show the estimated additive effect of the eQTL of those cases in which the LOD score is above 3, a value that has been traditionally used as a threshold to declare linkage (Lander and Schork, 1994).

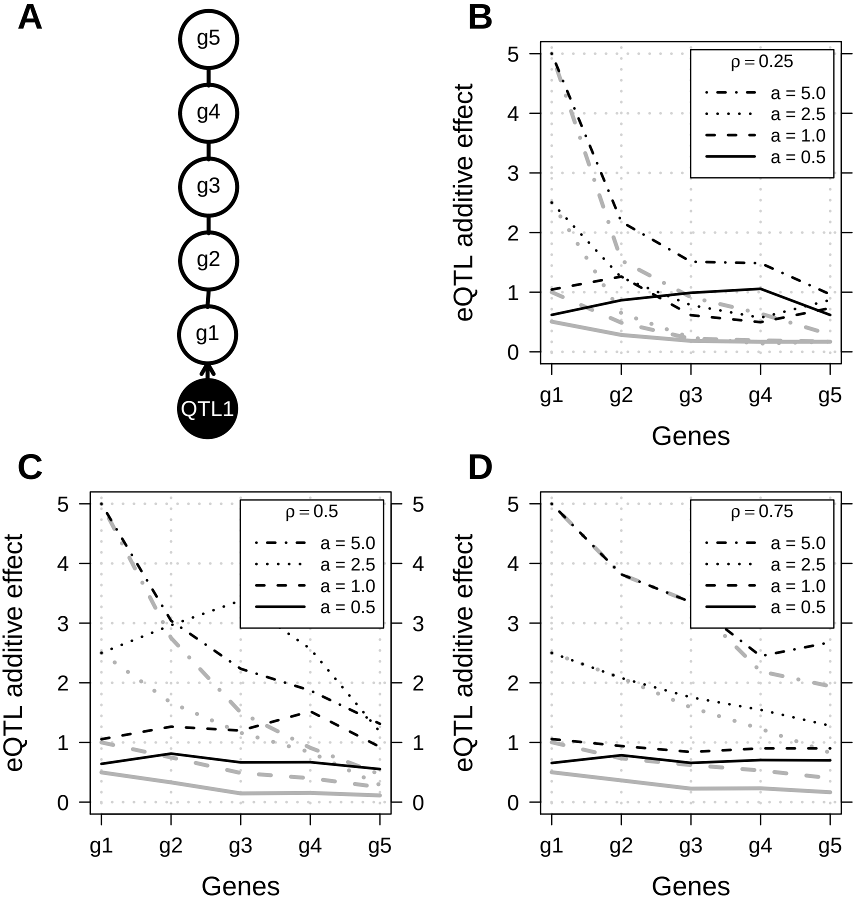

Figure 4.6: Estimation of eQTL additive effects under selection bias. Panel (A) shows the structure of the eQTL network underlying the mixed GMM employed to simulate the data shown on the other panels. Panels (B) to (D) show average estimated additive effects of the eQTL on each gene across 10,000 data sets simulated from different combinations of nominal gene-gene correlations (ρ) and additive effects (a). Grey lines were calculated from all data while black lines recreate the effect of selection bias by using only eQTL associations with LOD scores larger than 3.

By comparing them to the corresponding grey lines, we observe that the selection bias is noticeable on “g1” only when the true additive effect of the eQTL is small ($a = 0.5$). Yet, selection bias can be particularly strong across the rest of the genes regardless of the strength of the true additive effect (a) or the marginal correlation between the genes (ρ).

These simulations illustrate that when marginal independences between a genetic variant and a gene are tested, not only spurious eQTLs may give rise but also their estimated effects are larger than the true effects. Moreover, the effect of selection bias may also lead to larger proportions of phenotypic variance due to the detected eQTL (see Equation 2.18) (Sambrook et al., 1999; Broman, 2001).

4.8 Concluding remarks

In this chapter we have provided the basic theory of mixed GMMs that will be used throughout the rest of the thesis. Moreover,

- We have described a procedure to simulate the structure and the parameters of homogeneous mixed GMMs with given marginal linear correlations on the pure continuous associations and given additive effects on the mixed ones. We have also provided a procedure to simulate data sets from these mixed GMMs. In particular, we have simulated the structure, parameters and data from eQTL network models that represent the underlying associations between genetic variants and continuous genes and phenotypes from a backcross.

- These procedures have been implemented in the R/Bioconductor package called `qpgraph` available at http://www.bioconductor.org/packages/release/bioc/html/qpgraph.html and can be used in conjunction with the functionalities of the `R/qtl` package to build eQTL network models.

- By simulating data from mixed GMMs, we have gathered insight into the basis of eQTL networks. In particular, we have demonstrated the propagation of the additive effects of genetic variants to gene expression and continuous phenotypes as function of linear correlations between the genes. If additive effects are sufficiently large, these simulations demonstrate that indirect relationships may emerge as direct ones when marginally assessing the association between a genetic variant and a gene.

- We also provide simulations that recreate the effect of selection bias. We have shown that indirect eQTL effects are amplified under this circumstance, specially if the true effects are small, favoring the detection of spurious eQTL associations when inspecting the marginal association between a genetic variant and a gene. Further, this may give rise to larger percentages of the phenotypic variance explained by the detected eQTLs.
- These simulations illustrate the necessity to adjust for indirect effects when testing associations between genetic variants and genes.

5. An exact test to probe higher-order eQTL associations

5.1 Introduction

In Figure 4.5 we have seen that indirect additive effects in genetical genomics data can lead to spurious associations when inspecting the margin of the data formed by one gene and one genotype marker. More generally, when indirect effects are systematic and affect a large fraction of genes, such as batch effects (Leek et al., 2010), one also speaks of confounding effects. A classical approach to this problem within the statistical framework is to condition on the factors that confound the association of interest. In a linear regression model that treats each expression profile as a response variable and one or more markers and genes as its regressors, conditioning on confounding factors requires including them as main (e.g., Leek and Storey, 2007) or mixed (e.g., Listgarten et al., 2010) effects in the set of regressors.

In the context of mixed GMMs, we would like to perform a conditional independence test for mixed continuous and discrete data with conditioning sets of arbitrary size that would enable adjusting for confounding factors and for the expression of intervening genes. In principle, this amounts to include the conditioning set as covariates in both the null and the alternative models of the likelihood ratio test (LRT) of independence for an association between a genetic marker and a Gaussian distributed phenotype or between two Gaussian distributed phenotypes. In this chapter, we shall see that the classical χ^2 asymptotic test is not appropriate for this purpose and, for this reason, we investigate the development of an exact test. We first describe the theoretical aspects of this test and then we assess its performance with synthetic and real data from yeast.

Genetical genomics data from yeast

Throughout this thesis we use a well-studied real genetical genomics data set from yeast Brem and Kruglyak (2005) from a study where two yeast strains, a wild-type (RM11-1a) and a lab strain (BY4716), were crossed to generate 113 segregants which were profiled in their gene expression and genotyped. The genotype data was provided by Rachel Brem in the form of a flat file but only for 112 segregants. We discarded SNP markers with a duplicated identifier. Gene expression data come from two-channel microarray and are $\log_2$(sample/BY reference). They were also provided by Rachel Brem in the form of raw .gpr files where half of them are a dye-swap of the other half. We removed the expression data corresponding to the segregant for which we do not have genotype data. Then, we applied background correction, discarded control probes and normalized the expression data within and between arrays using the `limma` package (Smyth, 2005). We also excluded expression values from duplicated genes and genes which could not be mapped to current annotations in yeast. Finally, we averaged the expression values of the two dye-swapped arrays in order to correct a possible dye effect. This first step of normalized data consisted of 6,216 genes and 2,906 genotype markers throughout 112 samples, i.e., a total of $p = 9,122$ features, by $n = 112$ samples.

To proceed with the rest of the analysis, we took some further technical considerations. We observed that there were sets of SNP markers clustered in some regions of the genome that had identical genotypes. We sought the genes with highest LOD scores to each of these clusters of markers. When the gene was located in a different chromosome a marker was arbitrarily selected within its cluster. When the gene was located in the same chromosome, the closest marker to the gene was selected. The rest of identical markers were discarded from further analysis. We also removed genes whose annotation did not map to the April 2011 version of the yeast genome (sacCer3) at the UCSC Genome Browser (http://www.genome.ucsc.edu). This resulted in a final data set of 2,150 markers and 6,104 genes, i.e., a total of $p = 8,254$ features, by $n = 112$ samples.

5.2 An exact test of conditional independence

The approaches to learning the structure of a mixed GMM using higher-order correlations require testing for conditional independence between any two r.v.'s X_α and X_β given a set of conditioning ones X_Q, denoted as $X_\alpha \perp\!\!\!\perp X_\beta | X_Q$. To this end, we use a likelihood ratio test (LRT) between two models: a saturated model $\mathcal{M}_1$ and a constrained model $\mathcal{M}_0$ (Figure 5.1). Note that the classical LOD score (see Equation 2.6) corresponds to the LRT between $\mathcal{M}_1$ and $\mathcal{M}_0$ with $X_Q = \emptyset$.

The saturated model $\mathcal{M}_1$ is determined by the complete graph $G^1 = (V, E^1)$, where $V = \{\alpha, \beta, Q\}$ and $E^1 = V \times V$.

The constrained model $\mathcal{M}_0$ is determined by $G^0 = (V, E^0)$ with exactly one missing edge formed by the two vertices α, β representing the r.v.'s we wish to test, and thus $E^0 = \{V \times V\} \setminus (\alpha, \beta)$ and $Q = V \setminus \{\alpha, \beta\}$.

Note that both models are decomposable (see Chapter 4, pg. 66). For the model $\mathcal{M}_1$, $C_1 = V$ is the unique clique of G^1 and the separator set is $S = \emptyset$. On the other hand, the triple (α, β, Q) is a proper decomposition of $\mathcal{M}_0$ so that sets $C_1 = \{\alpha, Q\}$, $C_2 = \{\beta, Q\}$ are cliques of G^0 and $S = \{Q\}$ is the separator. Since $\mathcal{M}_0$ and $\mathcal{M}_1$ are decomposable, they admit explicit MLEs (see Equations 4.16-4.18).

We restrict the models used in the tests to homogeneous mixed GMMs, where $\Sigma(i) \equiv \Sigma$. As we already said in the previous chapter, in the context of eQTL models, this assumption implies that genotypes do not affect the variance of gene expression profiles but only their mean. In this case, the sample size requirements for the existence of MLEs are easier to meet with data where $p \gg n$.

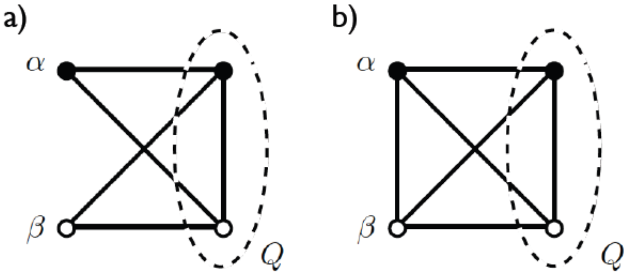

Figure 5.1: Example of two decomposable marked graphs representing a constrained model $\mathcal{M}_0$ (a) and a saturated model $\mathcal{M}_1$ (b). In (a), $X_\alpha \perp\!\!\!\perp X_\beta | X_Q$ where the cliques of $\mathcal{M}_0$ are $C_1 = \{\alpha, Q\}$ and $C_2 = \{\beta, Q\}$ and the separator is $S = \{Q\}$ while in (b), $X_\alpha \not\perp\!\!\!\perp X_\beta | X_Q$ and there is only one clique $C_1 = V$ and $S = \emptyset$.

The likelihood function of a homogeneous mixed GMM is written as

(Lauritzen, 1996, pg. 168):

$$\begin{aligned}\mathcal{L} &= \prod_{\nu=1}^{n} f\{x^\nu;(p,\mu,\Sigma)\} = \prod_{\nu=1}^{n} p(i^\nu)\phi\{y^\nu;\mu(i^\nu),\Sigma\} \\ &= \prod_{i\in\mathcal{I}} p(i)^{n(i)} \prod_{i\in\mathcal{I}} \prod_{\nu:i^\nu=i} \phi\{y^\nu;\mu(i),\Sigma\}\,,\end{aligned}$$

where ϕ stands for the probability density function of the multivariate Gaussian distribution.

In particular, for the saturated model $\mathcal{M}_1$, the likelihood function reduces to:

$$\mathcal{L}_1 = (2\pi)^{-n|\Gamma|/2} n^{n(|\Gamma|/2-1)} e^{-n|\Gamma|/2} |ssd|^{-n/2} \prod_{i\in\mathcal{I}} n(i)^{n(i)}\,, \tag{5.1}$$

where we have applied standard results about sampling distributions of the multivariate Gaussian distribution.

If the homogeneous mixed GMM is decomposable with cliques C_j and separators S_j, $j = 1,\ldots,k$, Equation (5.1) is re-written as (Lauritzen, 1996, pg. 191):

$$\mathcal{L} = (2\pi)^{-n|\Gamma|/2} n^{n(|\Gamma|/2-1)} e^{-n|\Gamma|/2} \times \prod_{j=1}^{k} \frac{|ssd_{S_j}|^{n/2} \prod_{i_{C_j}\in\mathcal{I}_{C_j}} n(i_{C_j})^{n(i_{C_j})}}{|ssd_{C_j}|^{n/2} \prod_{i_{S_j}\in\mathcal{I}_{S_j}} n(i_{S_j})^{n(i_{S_j})}}\,. \tag{5.2}$$

For the constrained model $\mathcal{M}_0$, where we have $C_1 = \{\alpha, Q\}$, $C_2 = \{\beta, Q\}$, $S_1 = \{Q\}$ and we set $S_2 = \emptyset$ (see Figure 5.1), we have

$$\begin{aligned}\mathcal{L}_0 &= (2\pi)^{-n|\Gamma|/2} n^{n(|\Gamma|/2-1)} e^{-n|\Gamma|/2} \\ &\times \frac{|ssd_{S_1}|^{n/2} \prod_{i_{C_1}\in\mathcal{I}_{C_1}} n(i_{C_1})^{n(i_{C_1})} \prod_{i_{C_2}\in\mathcal{I}_{C_2}} n(i_{C_2})^{n(i_{C_2})}}{|ssd_{C_1}|^{n/2}|ssd_{C_2}|^{n/2} \prod_{i_{S_1}\in\mathcal{I}_{S_1}} n(i_{S_1})^{n(i_{S_1})}}\,,\end{aligned} \tag{5.3}$$

whereas for the saturated model $\mathcal{M}_1$, where $C_1 = \{\alpha, \beta, Q\}$ and $S_1 = \emptyset$, Equation (5.2) coincides with Equation (5.1).

Given that $V = \Delta \cup \Gamma$, we denote by (γ, ζ) a pair of continuous r.v.'s (i.e., $\gamma, \zeta \in \Gamma$), by (δ, γ) a pair of mixed r.v.'s with $\delta \in \Delta$ and $\gamma \in \Gamma$, so that either $Q = V\backslash\{\gamma, \zeta\}$ or $Q = V\backslash\{\delta, \gamma\}$ are the conditioning subsets.

In the context of homogeneous mixed GMMs, the null hypothesis of conditional independence for the pure continuous case, $\gamma \perp\!\!\!\perp \zeta | Q$, corresponds to a zero value in the (γ, ζ) and (ζ, γ) cells of the canonical parameter K (see pg. 66). The log-likelihood ratio statistic, which is twice the difference of the log-likelihoods of models $\mathcal{M}_0$ and $\mathcal{M}_1$ given by Equations (5.3) and (5.1), respectively, is reduced to (Lauritzen, 1996, pg. 192):

$$D_{\gamma\zeta.Q} = -2\ln\left(\frac{\mathcal{L}_0}{\mathcal{L}_1}\right) = -2\ln\left(\frac{|ssd_\Gamma||ssd_{\Gamma\backslash\{\gamma,\zeta\}}|}{|ssd_{\Gamma\backslash\{\gamma\}}||ssd_{\Gamma\backslash\{\zeta\}}|}\right)^{n/2} = -2\ln\left(\Lambda_{\gamma\zeta.Q}\right)^{n/2} . \tag{5.4}$$

The null hypothesis of conditional independence in the mixed case, $\delta \perp\!\!\!\perp \gamma | Q$, corresponds to an expansion of the canonical parameter $h_\gamma(i)$ where the terms corresponding to δ are zero, $\eta_\delta(i)_\gamma = 0$ (see pg. 66). In this case, the log-likelihood ratio statistic is (Lauritzen, 1996, pg. 194):

$$D_{\delta\gamma.Q} = -2\ln\left(\frac{|ssd_\Gamma||ssd_{\Gamma^*}(\Delta^*)|}{|ssd_{\Gamma^*}||ssd_\Gamma(\Delta^*)|}\right)^{n/2} = -2\ln\left(\Lambda_{\delta\gamma.Q}\right)^{n/2} , \tag{5.5}$$

where $\Gamma^* = \Gamma\backslash\{\gamma\}$ and $\Delta^* = \Delta\backslash\{\delta\}$.

Under the null hypothesis, $D_{\gamma\zeta.Q}$ and $D_{\delta\gamma.Q}$ follow asymptotically a χ^2_{df} distribution with *df* degrees of freedom, where *df* is the difference in the number of free parameters of $\mathcal{M}_0$ and $\mathcal{M}_1$, as we shall see below.

Since the saturated $\mathcal{M}_1$ and the constrained $\mathcal{M}_0$ models are decomposable, they are collapsible onto the same set of variables $X_{V\backslash\{\gamma\}}$ (see Edwards, 2000, pg. 86-87 for a definition of collapsibility) and (Didelez and Edwards, 2004). This property implies that the density functions f of $\mathcal{M}_1$ and $\mathcal{M}_0$ can be factorized as

$$f_V = f_{V\backslash\{\gamma\}} \cdot f_{\gamma|V\backslash\{\gamma\}} , \tag{5.6}$$

such that the marginal and conditional densities, $f_{V\backslash\{\gamma\}} \in \mathcal{M}_{V\backslash\{\gamma\}}$ and $f_{\gamma|V\backslash\{\gamma\}} \in \mathcal{M}_{\gamma|V\backslash\{\gamma\}}$, respectively, can be parametrized separately.

From Equation (5.6), it follows that the likelihood function of $\mathcal{M}_1$ can be computed as the product of the likelihood of the marginal and the conditional models:

$$\mathcal{L}_1 = \mathcal{L}^1_{\gamma|V\backslash\{\gamma\}} \cdot \mathcal{L}^1_{V\backslash\{\gamma\}} ,$$

and, analogously, for the constrained model $\mathcal{M}_0$,

$$\mathcal{L}_0 = \mathcal{L}^0_{\gamma|V\backslash\{\gamma\}} \cdot \mathcal{L}^0_{V\backslash\{\gamma\}} .$$

The second term of these factorizations corresponds to the same saturated model induced by the complete subgraph composed by the vertices in $V\backslash\{\gamma\}$ (e.g., the subgraphs induced by $\{\alpha, Q\}$ in Figure 5.1 (a) and (b) are identical). Therefore, since $\mathcal{L}^1_{V\backslash\{\gamma\}} = \mathcal{L}^0_{V\backslash\{\gamma\}}$, we have that $D_{\gamma\zeta.Q}$ and $D_{\delta\gamma.Q}$ can be re-written as

$$-2\ln\left(\frac{\mathcal{L}_0}{\mathcal{L}_1}\right) = -2\ln\left(\frac{\mathcal{L}^0_{\gamma|V\backslash\{\gamma\}}}{\mathcal{L}^1_{\gamma|V\backslash\{\gamma\}}}\right) = -2\ln\left(\frac{\hat{\sigma}^0_{\gamma|V\backslash\{\gamma\}}}{\hat{\sigma}^1_{\gamma|V\backslash\{\gamma\}}}\right)^{-n/2} , \qquad (5.7)$$

where $\hat{\sigma}^0_{\gamma|V\backslash\{\gamma\}}$ and $\hat{\sigma}^1_{\gamma|V\backslash\{\gamma\}}$ stand for the estimation of the conditional variance of the r.v. X_γ given the rest of the r.v.'s under the null and the alternative conditional models $\mathcal{M}^0_{\gamma|V\backslash\{\gamma\}}$ and $\mathcal{M}^1_{\gamma|V\backslash\{\gamma\}}$, respectively.

In particular, these conditional models are equivalent to the ANCOVA models (see Edwards, 2000, pg. 91) in which the continuous r.v. $\gamma \in \Gamma$ is the response variable and the rest are explanatory variables. In this context, we have that the conditional variances in Equation (5.7) are equivalent to the residual sum of squares (RSS) of the corresponding ANCOVA models divided by the sample size n, that is, $\hat{\sigma}^0_{\gamma|V\backslash\{\gamma\}} = \mathrm{RSS}_0/n$ and $\hat{\sigma}^1_{\gamma|V\backslash\{\gamma\}} = \mathrm{RSS}_1/n$.

The ANCOVA model corresponding to $\mathcal{M}^1_{\gamma|V\backslash\{\gamma\}}$ for a backcross is

$$X_\gamma = \mu + \beta_\delta Z_\delta + \sum_{\kappa\in\Delta^*} \beta_\kappa Z_\kappa + \sum_{\kappa_1=2}^{|\Delta|} \sum_{\kappa_2=1}^{\binom{|\Delta|}{\kappa_1}} \left[\beta_{\kappa_1\kappa_2}\left(\prod_{\kappa=1}^{\kappa_1} Z_\kappa\right)\right] + \sum_{\lambda\in\Gamma^*} \beta_\lambda X_\lambda + \epsilon\,. \qquad (5.8)$$

In this model, μ is the phenotype's mean, the term $\beta_\delta Z_\delta$ represents the effect of the discrete variable $\delta \in \Delta$ that we are testing and we assume that $\epsilon \sim \mathcal{N}\left(0, \sigma^2_\gamma\right)$. The continuous r.v.'s in Q, Γ^*, are modeled as a linear combination of r.v.'s (third summation of the equation). On the other hand, the joint levels of δ and of the discrete r.v.'s in Q, Δ, are encoded through $(|\mathcal{I}| - 1)$ terms where some of them represent the main effects of each discrete r.v. (first summation). The rest of the terms encode all the interacting effects between the discrete r.v.'s (second summation). Each variable Z_κ is an indicator variable that takes value 0 or 1. For a backcross, each Z_κ takes value 0 if the genotype of κ is AA and takes value 1 if the genotype is AB; see Equation (2.7). For an intercross, each discrete r.v. has two corresponding indicator variables, one encoding the additive effect and the other encoding the dominance effect; see Equation (2.8).

Let's count the number of parameters of the model of Equation (5.8): $|\Delta|$ parameters come from the term encoding the main effect of Z_δ and the first summation; the second summation involves $2^{|\Delta|} - 1 - |\Delta|$ terms whereas the third one involves $|\Gamma| - 1$ terms. Thus, the model of Equation (5.8) has $n - 2^{|\Delta|} - |\Gamma| + 2$ free parameters in total where n is the sample size of the data.

Pure continuous case

For the pure continuous case, the conditional model $\mathcal{M}^0_{\gamma|V\setminus\{\gamma\}}$ is the same as the one in Equation (5.8) except that we remove the term of the third summation that corresponds to the variable X_ζ. In this case, this model has $n - 2^{|\Delta|} - |\Gamma| + 3$ free parameters. Therefore, under the null hypothesis, the statistic $D_{\gamma\zeta.Q}$ follows asymptotically a χ^2_{df} distribution with $df = 1$ degree of freedom.

In general, we can derive the degrees of freedom for the pure continuous case by writing explicitly the conditional expectation of X_γ given $X_{V\setminus\{\gamma\}}$. Under the conditional saturated model $\mathcal{M}^1_{\gamma|V\setminus\{\gamma\}}$, this corresponds to

$$\mathrm{E}\left(X_\gamma|\Delta, \Gamma\setminus\{\gamma\}\right) = \alpha(i_\Delta) + \sum_{\lambda\in\Gamma\setminus\{\gamma\}} \beta_{\gamma\lambda|\Gamma\setminus\{\gamma\}} X_\lambda\,, \qquad (5.9)$$

where $\alpha(i_\Delta) = \mu_\gamma(i_\Delta) - \sum_{\lambda\in\Gamma\setminus\{\gamma\}} \beta_{\gamma\lambda|\Gamma\setminus\{\gamma\}} \mu_\lambda(i_\Delta)$ and $\beta_{\gamma\lambda|\Gamma\setminus\{\gamma\}}$ is the partial regression coefficient that is found through the canonical parameter $K = \{k_{\gamma\zeta}\}, \forall\gamma, \zeta \in \Gamma$, as $\beta_{\gamma\lambda|\Gamma\setminus\{\gamma\}} = -k_{\gamma\lambda}/k_{\gamma\gamma}$ (Lauritzen, 1996, pg. 130). This model has $n - |\mathcal{I}| - |\Gamma| + 1$ free parameters since it has $|\mathcal{I}|$ parameters that come from the first term in Equation (5.9) and $|\Gamma| - 1$ from the second term.

Under the model $\mathcal{M}^0_{\gamma|V\setminus\{\gamma\}}$, the conditional expectation of X_γ given $X_{V\setminus\{\gamma\}}$ is written as

$$\mathrm{E}\left(X_\gamma|\Delta, \Gamma\setminus\{\gamma, \zeta\}\right) = \alpha(i_\Delta) + \sum_{\lambda\in\Gamma\setminus\{\gamma,\zeta\}} \beta_{\gamma\lambda|\Gamma\setminus\{\gamma,\zeta\}} X_\lambda$$

which, in an analogous way as the previous case, leads to $n - |\mathcal{I}| - |\Gamma| + 2$ parameters.

By computing the difference in the number of free parameters of both models, we see again that $D_{\gamma\zeta.Q}$ follows asymptotically a χ^2_{df} distribution with $df = 1$ degree of freedom.

Mixed case

In the mixed case, the likelihood ratio statistic $D_{\delta\gamma.Q}$ of Equation (5.5) is related to the LOD score used in QTL mapping:

$$\mathrm{LOD} = \log_{10}\left(\frac{\mathcal{L}_1}{\mathcal{L}_0}\right) ,$$

through the following transformation of the LOD score:

$$D_{\delta\gamma.Q} = 2\ln(10)\mathrm{LOD} \, , \tag{5.10}$$

as we have seen in Equation (2.10) for the case in which $Q = \emptyset$. In fact, since the ratio between $\mathcal{L}_1$ and $\mathcal{L}_0$ is equivalent to the ratio between $\mathcal{L}^1_{\gamma|V\setminus\{\gamma\}}$ and $\mathcal{L}^0_{\gamma|V\setminus\{\gamma\}}$ we have that

$$\mathrm{LOD} = \log_{10}\left(\frac{\mathcal{L}_1}{\mathcal{L}_0}\right) = \log_{10}\left(\frac{\mathcal{L}^1_{\gamma|V\setminus\{\gamma\}}}{\mathcal{L}^0_{\gamma|V\setminus\{\gamma\}}}\right) . \tag{5.11}$$

In this case, the conditional model $\mathcal{M}^1_{\gamma|V\setminus\{\gamma\}}$ for a backcross is the same as the one in Equation (5.8) whereas in the conditional model $\mathcal{M}^0_{\gamma|V\setminus\{\gamma\}}$, we delete all the terms of Equation (5.8) that involve the r.v. X_δ. This constrained model has $n - 2^{|\Delta|-1} - |\Gamma| + 2$ free parameters. Thus, for a backcross, the likelihood ratio statistic $D_{\delta\gamma.Q}$, and therefore, the transformation of the LOD score (Equation 5.10) follow a χ^2 distribution with $df = 2^{|\Delta|-1}$ degrees of freedom (as we already noted in Equation (2.10) for the case in which $Q = \emptyset$).

Again, we can derive the degrees of freedom of the χ^2 distribution for the general case by looking at the conditional expectation of the models $\mathcal{M}^1_{\gamma|V\setminus\{\gamma\}}$ (Equation 5.9) and $\mathcal{M}^0_{\gamma|V\setminus\{\gamma\}}$:

$$\mathrm{E}\left(X_\gamma|\Delta\setminus\{\delta\}, \Gamma\setminus\{\gamma\}\right) = \alpha(i_{\Delta\setminus\{\delta\}}) + \sum_{\lambda\in\Gamma\setminus\{\gamma\}} \beta_{\gamma\lambda|\Gamma\setminus\{\gamma\}} X_\lambda \, . \tag{5.12}$$

Here, the first term involves $|\mathcal{I}_{\Delta^*}|$ parameters and the second $|\Gamma| - 1$, so that the constrained model has $n - |\Gamma| + |\mathcal{I}_{\Delta^*}| - 1$ free parameters. Finally, we have that $D_{\delta\gamma.Q}$ and the transformed LOD score follow a χ^2 distribution with $df = |\mathcal{I}_{\Delta^*}|(|\mathcal{I}_\delta| - 1)$ degrees of freedom.

Exact distribution

However, Lauritzen (1996, pg. 192 to 194) observes that, for decomposable mixed GMMs, the likelihood ratios $\Lambda_{\gamma\zeta.Q}$ in Equation (5.4) and $\Lambda_{\delta\gamma.Q}$ in Equation (5.5) follow exactly a beta distribution. In order to enable the exact test for homogeneous mixed GMMs, we proceed to derive their corresponding parameters.

Due to the decomposability and collapsibility of the saturated $\mathcal{M}_1$ and constrained $\mathcal{M}_0$ models, we have seen that the analysis of the joint densities is equivalent to the study of the univariate conditional densities of X_γ given the rest of the variables. Concretely, for the pure continuous case, the likelihood ratio statistic $\Lambda_{\gamma\zeta.Q}$ is equivalent to the ratio $\mathrm{RSS}_1/\mathrm{RSS}_0$ where RSS_0 and RSS_1 are the residual sum of squares of the constrained and the saturated univariate models, respectively, and both follow a χ^2_k distribution, where k is the number of free parameters of each model.

Let $\mathrm{RSS}_{1.0}$ denote the difference $\mathrm{RSS}_0 - \mathrm{RSS}_1$. Following (Rao, 1973, pg. 166), if a r.v. X follows a χ^2_k with k degrees of freedom it also follows a gamma distribution $\Gamma(k/2, 2)$. Hence,

$$\mathrm{RSS}_1 \sim \Gamma\left(\frac{n - |\Gamma| - |\mathcal{I}| + 1}{2}, 2\right), \ \mathrm{RSS}_0 \sim \Gamma\left(\frac{n - |\Gamma| - |\mathcal{I}| + 2}{2}, 2\right),$$

and therefore, $\mathrm{RSS}_{1.0} \sim \Gamma\left(1/2, 2\right)$. Moreover, if X and Y are two independent r.v.'s such that $X \sim \Gamma\left(k_1, \theta\right)$ and $Y \sim \Gamma\left(k_2, \theta\right)$, then (Rao, 1973, pg. 165),

$$\frac{X}{X+Y} \sim \mathcal{B}(k_1, k_2),$$

where $\mathcal{B}(k_1, k_2)$ denotes the beta distribution with shape parameters k_1 and k_2. If we let $X = \mathrm{RSS}_1$ and $Y = \mathrm{RSS}_{1.0}$, it follows that

$$\Lambda_{\gamma\zeta.Q} = \frac{\mathrm{RSS}_1}{\mathrm{RSS}_0} \sim \mathcal{B}\left(\frac{n - |\Gamma| - |\mathcal{I}| + 1}{2}, \frac{1}{2}\right).$$

In the mixed case, the conditional expectation of $\gamma \in \Gamma$ given the rest of variables under the saturated model $\mathcal{M}_1$ coincides with the continuous case in Equation (5.9) and, under the constrained model $\mathcal{M}_0$, it is written in Equation (5.12). By an argument analogous to the pure continuous case, the likelihood ratio statistic raised to the power $2/n$ for the null hypothesis of a missing mixed edge follows a beta distribution with these parameters:

$$\Lambda_{\delta\gamma.Q} \sim \mathcal{B}\left(\frac{n-|\Gamma|-|\mathcal{I}|+1}{2}, \frac{|\mathcal{I}_{\Delta^*}|(|\mathcal{I}_\delta|-1)}{2}\right) . \tag{5.13}$$

Therefore, we have that the following transformation of the LOD score

$$\Lambda_{\delta\gamma.Q} = 10^{-\frac{2}{n}\mathrm{LOD}}$$

follows a beta distribution with parameters given in Equation (5.13).

Estimation of the phenotypic variance explained by an eQTL

Note that the percentage of the phenotypic variance, that is, the variance of X_γ, explained by an eQTL X_δ while conditioning on the rest of r.v.'s, denoted by η^2, can be estimated as the difference between the estimated conditional variances of X_γ given the rest of r.v.'s under the saturated and the constrained models divided by the total variance of X_γ:

$$\eta^2 = \frac{\hat{\sigma}^0_{\gamma|V\backslash\{\gamma\}} - \hat{\sigma}^1_{\gamma|V\backslash\{\gamma\}}}{\hat{\sigma}_{\gamma\gamma}} = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{n \cdot \mathrm{var}(X_\gamma)} . \tag{5.14}$$

In particular, when $V\backslash\{\gamma\} = \{\delta\}$, that is, $Q = \emptyset$, the estimated conditional variance under the constrained model, RSS_0/n, is equal to the unconditional variance of X_γ ($\mathrm{RSS}_0/n = \mathrm{var}(X_\gamma)$). In this case, the percentage of the phenotypic variance explained by the eQTL reduces to

$$\eta^2 = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_0} = 1 - \Lambda_{\delta\gamma.Q} \, , \tag{5.15}$$

as we have already seen in Equation (2.19).

5.3 Testing higher-order eQTL associations

In this section, we perform simulations that show that the exact conditional independence test described in Section 5.2 provides an accurate control of its significance level and yields a correct uniform distribution of its p-values under the null hypothesis, as opposed to the classical asymptotic counterpart. We also investigate how widespread confounding effects can be implicitly adjusted with higher-order conditional independences.

Control of Type-I error as function of sample size

Using a genetic map formed by only one chromosome 100 cM long and 10 markers equally-spaced, we built an eQTL network with 100 genes, where one of them, denoted g hereafter, has a *cis*-eQTL and where every gene is randomly connected to two other genes in the eQTL network. Given this network, and using procedures described in the previous chapter (Section 4.5), we simulated 10 models with parameter sets randomly drawn such that the mean marginal correlation between the genes is $\rho = 0.5$ and the additive effect from each eQTL is $a = 1$. From each model, 1,000 data sets are simulated with decreasing sample sizes $n = \{100, 75, 50, 25\}$.

In each data sets two conditional independence tests, the asymptotic and the exact one, were performed between the simulated genotypes from the eQTL and the simulated expression profile from a gene g' connected to g. Note that the eQTL and g' are indirectly associated by a path in the eQTL network but they are no directly connected, thereby recreating a null hypothesis of conditional independence between the eQTL and gene g' given the genes connected to g', which are g and some other gene.

The asymptotic test was performed by using the `scanone()` function from the `R/qtl` package specifying the genes connected to g' as additive covariates through the `addcovar` argument. The resulting `R/qtl` LOD scores were transformed to their χ^2-distributed counterparts (see Equation 5.10). The number of rejected tests at $\alpha = 0.05$ within each sample size constitutes an estimate of the type-I error rate of the test as function of the sample size.

Figure 5.2A shows that as the sample size decreases, the type-I error rate for the asymptotic test increases while the exact test yields a proper error rate around the nominal $\alpha = 0.05$ across all different sample sizes.

Control of Type-I error as function of network degree

We altered the previous simulation setup fixing the sample size at $n = 25$ and using eQTL networks of increasing connectivity between the genes. More concretely, we generated gene networks as random d-regular graphs (Harary, 1969) with $d = \{3, 4, 5, 6, 7\}$, where one of the genes g had a *cis*-eQTL. We used the same previous definition of null hypothesis, but increasing the size of the conditioning set to the d other genes connected to g', where g' is connected to g.

By counting again the number of rejected tests at $\alpha = 0.05$ across simulated data sets, we inspected the empirical type-I error rate as function of the gene network degree d, which determines the conditioning set size. Figure 5.2B shows that the type-I error rate grows with the degree of the underlying gene network for the asymptotic test, while the exact test controls properly the nominal level of $\alpha = 0.05$ across conditioning sets of increasing size (Leek and Storey, 2007).

Distribution of *p*-values under the null hypothesis

The previous two simulations revealed that the use of the LOD score in a hypothesis test for conditional independence can be problematic when taking its classical interpretation as an asymptotically distributed χ^2 statistic (Equation 5.10). To assess on real data the potential impact of this observation we took a yeast genetical genomics data set (Brem and Kruglyak, 2005) and explored thousands of null hypotheses of conditional independence between genotypes and expression profiles. Under each of these null hypotheses p-values should be uniformly distributed (Lehman and Romano, 2005) and discrepancies to this baseline may potentially inflate the downstream rate of false discoveries after adjusting for multiple testing (Leek and Storey, 2007).

We recreated null hypotheses from real data by first selecting 1,000 pairs of genotype markers and gene expression profiles uniformly at random. Second, we bootstrapped 1,000 data sets by sampling with replacement a number n of observations from the original full data of 112 segregants, permuting the expression values to break any possible existing correlation between the marker and the gene.

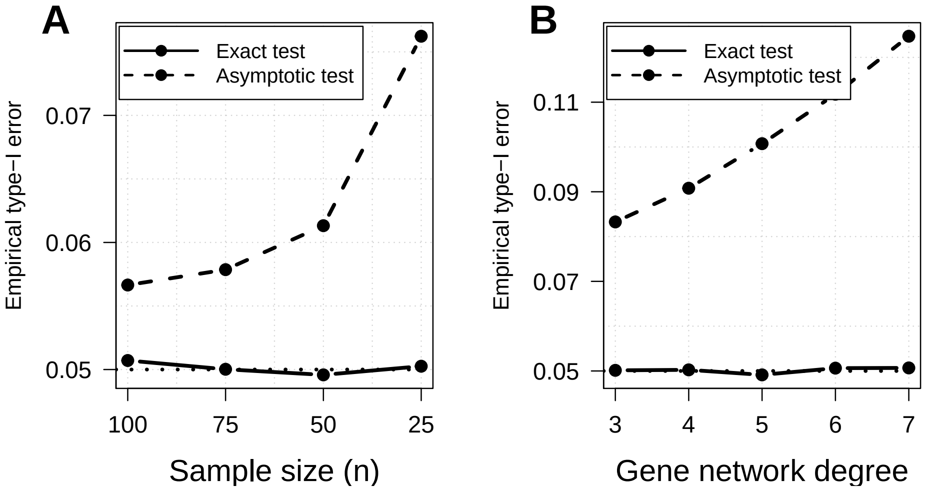

Figure 5.2: Empirical type-I error rate of asymptotic and exact tests for conditional independence. Plots on (A) and (B) show the empirical type-I error rate from simulated data for conditional independence tests at a nominal level $\alpha = 0.05$ (dotted horizontal line) as function of the sample size n (A), and as function of the underlying gene network degree (B). Both plots show that the exact test controls correctly the type-I error rate while the asymptotic one does not.

In every bootstrapped data set, we performed a conditional independence test between each pair of marker and gene with a fixed conditioning set of q randomly chosen genes among the ones outside the 1,000 marker-gene pairs. The resulting p-value should be a *null* p-value, in the sense that it has been calculated under the null hypothesis of conditional independence. We considered different combinations of sample $n = \{25, 75, 100\}$ and conditioning-order $q = \{0, 5, 15\}$ values where $q = 0$ means that we performed a hypothesis test of marginal independence. For each of these combinations of (n, q) values in a particular pair of marker and gene, we assessed the goodness of fit to a uniform distribution using a Kolmogorov-Smirnov (KS) test, of the null p-values calculated from the 1,000 bootstrapped data sets. In turn, as illustrated in (Leek and Storey, 2007), the p-values of the KS tests of the 1,000 pairs should be themselves uniformly distributed. This can be easily verified by means of quantile-quantile plots shown in Figure 5.3. In these plots, uniformly distributed p-values should lead to lines close to the diagonal. This is the case for all the exact tests ran in every combination of (n, q) values, and depicted with solid lines. On the other hand, the distribution of null p-values obtained with asymptotic tests increasingly deviate from the uniform distribution as $n - q$ decreases.

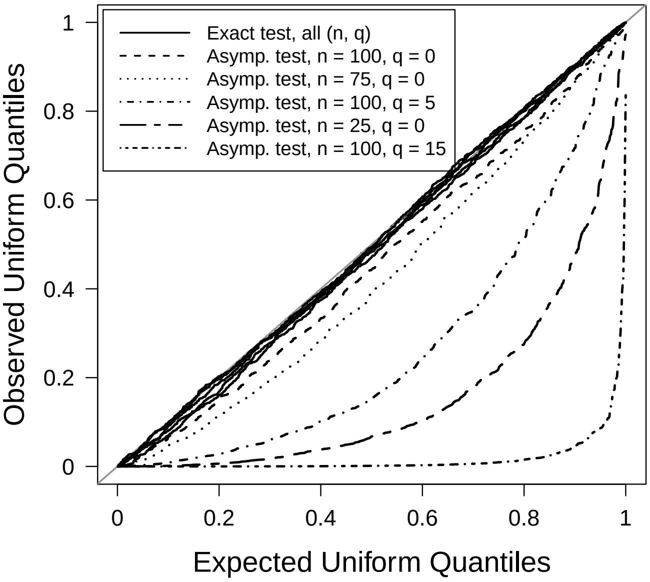

Figure 5.3: Distribution of p-values under the null hypothesis calculated with the asymptotic and exact tests for conditional independence. Quantile-quantile plots of Kolmogorov-Smirnov p-values obtained by testing asymptotic and exact p-values, from eQTL independence associations in a real genetical genomics yeast data set, against their expected null uniform distribution, under different sample and conditioning sizes n and q, respectively. The exact test (solid lines) produces correctly distributed null p-values while the asymptotic test displays an increasing discrepancy to the uniform distribution of null p-values as $n - q$ decreases.

Higher-order conditioning adjusts for confounding effects

Confounding effects in gene expression data affect most of the genes being profiled. Sometimes the sources of confounding are known, or can be estimated with methods such as SVA (Leek and Storey, 2007) or PEER (Stegle et al., 2010), and may be explicitly adjusted by including them as main effects into the model. Often, however, these sources are unknown and it may be difficult to adjust or remove them without affecting the biological signal and underlying correlation structure that we want to estimate. Using simulations, here we show that confounding effects affecting all genes can be implicitly adjusted by conditioning on higher-order associations.

We used the same simulation setup as when previously assessed type-I error rates with the following changes: first, we did not include any association between genes; second, the single eQTL present in the network had a fixed additive effect $a = 2.5$; finally, a continuous confounding factor with $\rho = 0.5$ was included under two models, a systematic one in which the confounding factor affects all genes, and a specific one in which it affects only the two genes, or the gene and marker, being tested. We considered a fixed sample size $n = 100$ and conditioning orders $q = \{0, 1, \ldots, 50\}$, where $q = 0$ corresponds to the marginal association without conditioning.

We tested for the presence of a gene-gene association and of an eQTL association between the marker containing the simulated eQTL and one of the genes not associated to that eQTL. Note that none of these associations were present in the simulated eQTL network. For every q order with $q > 0$, a subset Q of size q was sampled uniformly at random among the rest of the genes not being tested, and used for conditioning. When considering the explicit adjustment of the confounding factor, this one was added to Q except when $q = 0$ since then $Q = \{\emptyset\}$. Once Q was fixed, 100 data sets were sampled from the corresponding mixed GMM and two conditional independence tests were conducted in each data set for the presence of both, the eQTL and the gene-gene association, given the sampled genes in Q.

Figure 5.4 shows the empirical type-I error rate as function of the conditioning order q, where panel (A) corresponds to the eQTL association and (B) to the gene-gene association. This figure shows that, as expected, the explicit inclusion of the confounding factor in the conditioning subset Q (dotted lines) adjusts the confounding effect immediately with $q > 0$ in both situations, when either all genes are affected or only the tested ones. When the confouding effect is not included in Q and affects only the tested genes (solid lines), it yields high type-I error rates that only decrease linearly with $n - q$, quantity on which statistical power depends. However, when confounding affects all genes (dashed lines) the type-I error rate has an exponential decay and for $q > 20$ the confounding effect is effectively adjusted in these data.

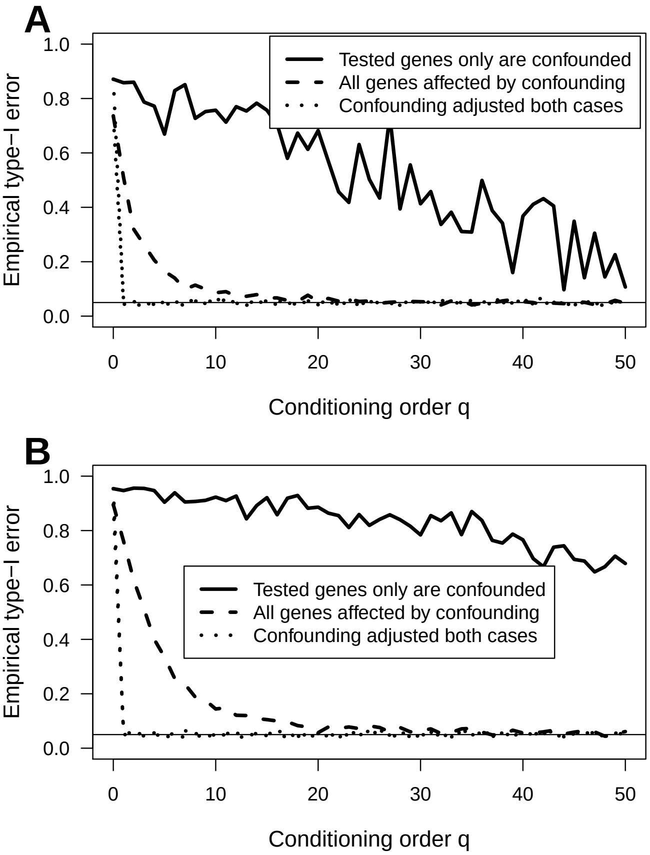

Figure 5.4: Explicit and implicit adjustment of confounding with higher-order conditional independence tests. Empirical type-I error rate for conditional independence tests from simulated data at a nominal level $\alpha = 0.05$ (dotted horizontal line) as function of the conditionin order q. Panel (A) shows results on testing for an absent eQTL association while panel (B) shows them for an absent gene-gene association. Solid lines correspond to the model under which confounding affects only the tested genes while dashed lines correspond to a confounding effect on all genes. In the latter situation, higher-order conditioning implicitly adjusts for the confounding effect when $q > 20$ in these data. Dotted lines from both confounding models overlap because they correspond to the inclusion of the confounding effect in the conditioning subsets, thereby explicitly adjusting for it.

5.4 Concluding remarks

In this chapter we have derived an exact conditional independence test for mixed continuous and discrete data which enables the investigation of higher-order eQTL and gene-gene associations. To this end, we perform a likelihood ratio test between two decomposable mixed GMMs. We next expose our main contributions:

- We have derived an exact charaterization of the null hypothesis of conditional independence between a genetic variant and a gene expression profile or between two gene expression profiles, given other genes, genetic variants or phenotypic variables.
- The exact test specially outperforms the traditional χ^2 asymptotic test for a small number of samples (n) and when the complexity of the underlying gene network increases. Moreover, the exact test provides a correct uniform distribution of p-values under the null hypothesis of conditional independence.
- By conditioning on a sufficiently large set of variables, we verify that our test is able to discriminate direct from indirect relationships between a genetic variant and a gene expression profile or between two gene expression profiles and correct for confounding effects simultaneously affecting all gene expression profiles.

6. Learning eQTL networks with mixed GMMs

In this chapter we present a methodology to map eQTLs and learn the structure of eQTL networks from genetical genomics data (where the number of variables p is much larger than the number of samples n) based on mixed graphical model theory. We describe the theoretical aspects of this new approach and assess its performance with real data from yeast. Our method requires testing higher-order conditional independences between each genetic variant and each gene expression profile and between two gene expression profiles, given other genes, genetic variants or phenotypic variables. We extend the concept of limited-order graphical model and the strategy to learn Gaussian GMMs introduced by Castelo and Roverato (2006) to the case in which data is composed of discrete and continuous r.v.'s (Section 6.1 and Section 6.2).

Then, by applying our learning approach on a real genetical genomics data set from yeast (Brem and Kruglyak, 2005), we assess the benefits of applying an exact conditional independence test when trying to dissect true direct eQTL and gene-gene associations. In particular, we focus on the study of distant linkages in order to identify the *trans*-eQTLs that have a direct effect on their target genes. Although many *trans*-eQTLs have been reported (e.g. Yvert et al., 2003; Brem and Kruglyak, 2005), there is little understanding of their functional role and, often, it is difficult to infer the underlying regulatory mechanism of these distant associations. Further, many spurious *trans*-associations may show up by applying classical QTL techniques as a consequence of strong gene-gene correlations and confounding effects. To this end, in the rest of the chapter we conduct different experiments and compare the results with those obtained from classical LOD scores computed with the single marker regression approach of the `R/qtl` package.

6.1 *q*-Order correlation graphs

The ability to test for conditional independences of arbitrary order opens up a wide spectrum of strategies that can be followed to learn a mixed

GMM from data, similarly as with Gaussian GMMs for pure continuous data (see, e.g., Castelo and Roverato, 2006; Kalisch and Bühlmann, 2007). In the context of estimating eQTL networks from genetical genomics data the number of genes and genotype markers p exceeds by far the sample size n, i.e., $p \gg n$. This fact precludes conditioning directly on the rest of the genes and markers $X_{V\setminus\{i,j\}}$ when testing for an eQTL association (i, j) while adjusting for all possible indirect effects. In other words, we cannot directly test for full-order conditional independences $X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}}$.

We approach this problem using limited-order correlations, an strategy successfully applied to Gaussian GMMs (Castelo and Roverato, 2006). It consists of testing for conditional independences of order $q < (p - 2)$, i.e., $X_i \perp\!\!\!\perp X_j | X_Q$ with $|Q| = q$, expecting that many of the indirect relationships between i and j can be explained by subsets Q of size q. The extent to which this can happen depends on the sparseness of the underlying network structure G and on the number of available observations. The mathematical object that results from testing q-order correlations is called a q-order correlation graph, or qp-graph (Castelo and Roverato, 2006), and it is defined as follows (Castelo and Roverato, 2006, Def. 1):

Definition 12. Let P_V be a probability distribution which is Markov over an undirected graph $G = (V, E)$ with $|V| = p$ and an integer $0 \leq q \leq (p - 2)$. A qp-graph of order q with respect to G is the undirected graph $G^{(q)} = (V, E^{(q)})$ where $(i, j) \notin E^{(q)}$ if and only if there exists a set $U \subseteq V\setminus\{i, j\}$ with $|U| \leq q$ such that $X_i \perp\!\!\!\perp X_j | X_U$ holds in P_V.

Assuming there are no additional independence restrictions in P_V than those in G, it can be shown (Castelo and Roverato, 2006) that $G \subseteq G^{(q)}$ in the sense that every edge that is present in the true underlying network G is also present in the qp-graph $G^{(q)}$. In fact, $G^{(q')} \subseteq G^{(q)}$ if and only if $q' \geq q$. Hence, as the conditioning sets become larger, the qp-graph gets closer to the true structure G, and therefore, $G^{(q)}$ can be seen as an approximation to G; see (Castelo and Roverato, 2006) for further details.

Because separation in undirected marked graphs with mixed discrete and continuous vertices works the same as in undirected pure graphs with either one of these two types of vertices, it follows that the definition of qp-graph also holds for mixed vertices and CG-distribution P_V. In

Figure 6.1, we illustrate this with an example. Let G be the true underlying eQTL network depicted on the left of Figure 6.1, where the vertex 5 represents an eQTL and the rest of the vertices represent gene expression profiles. The rest of the graphs of this figure are the qp-graphs $G^{(q)}$ associated to q-order correlations of order 0, 1 and 2, respectively, from left to right.

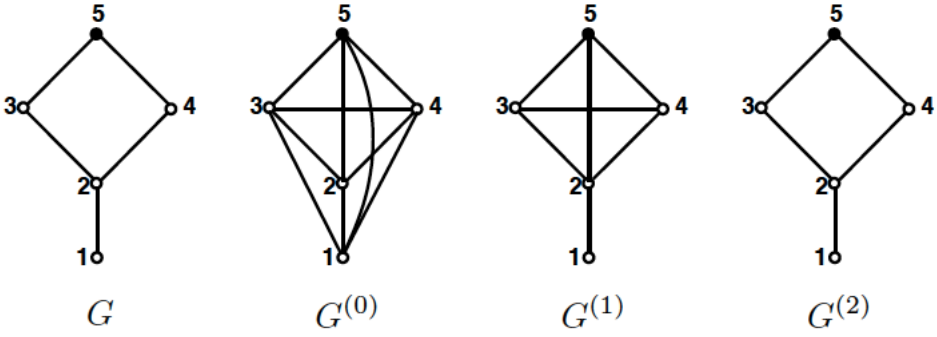

Figure 6.1: Examples of qp-graphs. This figure illustrates the undirected marked qp-graphs $G^{(q)}$ for different values of q associated to the true undirected marked graph G. Note that the qp-graph of order $q = 2$, $G^{(2)}$, already coincides with the true graph G although the maximum order q of G is 3. *Adapted from Castelo and Roverato (2009).*

6.2 Estimation of eQTL networks as qp-graphs

Instead of directly approaching the problem of inferring the graph structure G of the underlying eQTL network from genetical genomics data with $p \gg n$, we propose to calculate a qp-graph estimate $\hat{G}^{(q)}$. For this purpose, we use a measure of association between two r.v.'s called the *non-rejection rate* (NRR), which is defined as follows.

Let $\mathcal{Q}_{ij}^{q} = \{Q \subseteq V \backslash \{i,j\} : |Q| = q\}$ and let T_{ij}^{q} be a binary r.v. associated to the pair of vertices (i,j) that takes values from the following three-step procedure:

1. an element Q is sampled from $\mathcal{Q}_{ij}^{q}$ according to a (discrete) uniform distribution;

2. test the null hypothesis of conditional independence

$$H_0 : X_i \perp\!\!\!\perp X_j | X_Q \,,$$

which in the case of pure continuous data it corresponds to test the null hypothesis of zero partial correlation controlling for the subset Q such that $|Q| = q$ ($H_0 : \rho_{ij.Q} = 0$); and

3. if the null hypothesis H_0 is rejected then T^q_{ij} takes value 0, otherwise takes value 1.

We have that T^q_{ij} follows a Bernoulli distribution and the non-rejection rate, denoted as ν^q_{ij}, is defined as its expectancy

$$\nu^q_{ij} := \mathrm{E}[T^q_{ij}] = \Pr(T^q_{ij} = 1) \,.$$

In particular, we have that the probability that H_0 is not rejected given a set $Q \in \mathcal{Q}^q_{ij}$, denoted as $\Pr(T^q_{ij} = 1|Q)$, is $(1 - \alpha)$ if the set Q separates i and j in G, where α is the probability of the type-I error of the test. If Q does not separate the two vertices i and j in G, then $\Pr(T^q_{ij} = 1|Q) = \beta_{ij.Q}$, where $\beta_{ij.Q}$ is the probability of the type-II error of the test. Then, if we denote by π_{ij} the proportion of elements of $\mathcal{Q}^q_{ij}$ which separates i and j in G, Castelo and Roverato (2006) found that the theoretical proportion of non-rejected null hypotheses for each pair of variables (X_i, X_j) depends on α, β_{ij} and π_{ij}:

$$\nu^q_{ij} = \beta_{ij}(1 - \pi_{ij}) + (1 - \alpha)\pi_{ij} \,, \tag{6.1}$$

where β_{ij} is the average value of the type-II error for the pair of vertices i, j over $\mathcal{Q}^q_{ij}$. In fact, we have that $1-\beta_{ij}$ is the average statistical power of the hypothesis tests associated to the r.v.'s (X_i, X_j).

This expression helps understanding the information conveyed by the NRR in the following way. For a pair of adjacent vertices i, j in G, we have that $\pi_{ij} = 0$ and the theoretical NRR is equal to $\nu^q_{ij} = \beta_{ij}$. It follows that the NRR for i, j adjacent in G is 1 minus the statistical power to detect the association. Hence, for a finite sample size n, the NRR depends on the strength of the linear association of X_i, X_j over all marginal distributions of size $(q + 2)$. In particular, note that from Figure 6.2 and Equation (6.1) it follows that a NRR value ν^q_{ij} close to zero implies that both β_{ij} and π_{ij} are close to zero. This means that either (i, j) is in G or q is too small.

On the contrary, a value of ν_{ij}^q close to one either means that π_{ij} or β_{ij} are large (see Equation 6.1 and Figure 6.2), and we can conclude that either i and j are not adjacent in G or, otherwise, there is no sufficient statistical power to detect that association. As the statistical power to reject the null hypothesis H_0 depends on $n - q$, the latter circumstance may be due to an insuficient sample size n, a value of q that is too large, or both.

From these observations it follows that a NRR value ν_{ij}^q close to zero indicates that $(i, j) \in G^{(q)}$ while a value close to one points to the contrary, $(i, j) \notin G^{(q)}$.

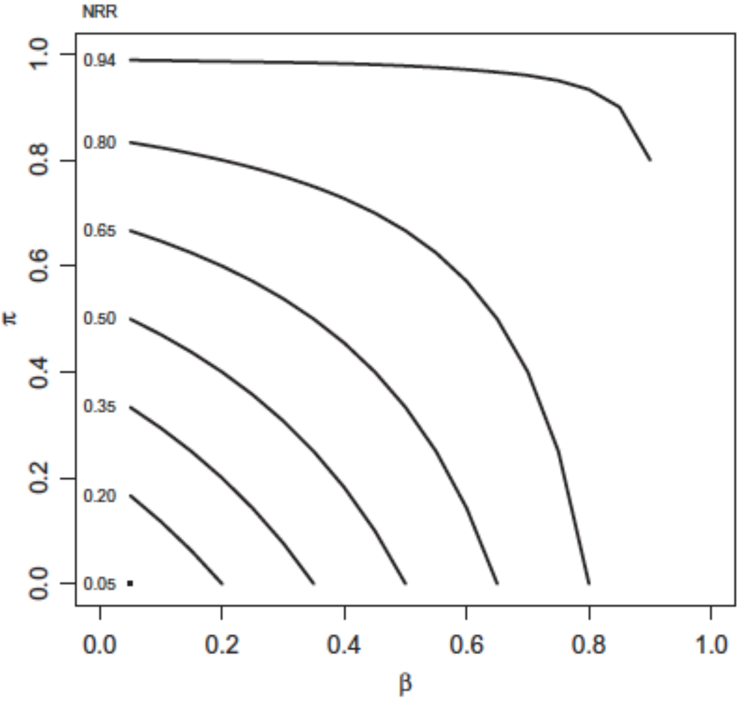

Figure 6.2: This figure shows the relationship between the NRR values and the values of π_{ij} (y-axis) and β_{ij} (x-axis). *Taken from Roverato and Castelo (2012).*

An unbiased estimation of the NRR for a pair of r.v.'s (X_i, X_j) is obtained by testing the conditional independence $X_i \perp\!\!\!\perp X_j | X_Q$ for every $Q \in \mathcal{Q}_{ij}^q$. However, the number of subsets Q in $\mathcal{Q}_{ij}^q$ can be prohibitively large. An effective approach to address this problem (Castelo and Roverato, 2006) consists of calculating an estimate $\hat{\nu}_{ij}^q$ on the basis of a limited number subsets $Q \in \mathcal{Q}_{ij}^{q}$, such as one-hundred, sampled uniformly at random.

We may be interested in explicitely adjusting for confounding factors and other covariates $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$. It is straightforward to incorporate them into a NRR $\nu_{ij.\mathcal{C}}^{q}$ by sampling subsets Q from

$$Q_{ij}^{q} = \{Q \subseteq \{V \backslash \{i, j\}\} \cup \mathcal{C} : |Q| = q\} .$$

Note that covariates in $\mathcal{C}$ can be known or, in the case of unknown confounding factors, estimated with algorithms such as SVA (Leek and Storey, 2007) or PEER (Stegle et al., 2010).

Finally, an estimation of the underlying eQTL network G is obtained by learning the qp-graph $\hat{G}^{(q)}$. Starting from a complete graph, the strategy consists in disconnecting those pairs of variables (i, j) whose corresponding r.v.'s have a NRR value above a certain threshold ϵ, that is,

$$\hat{G}^{q} := \{(V, E^{(q)}) : (i, j) \in E^{(q)} \Leftrightarrow \nu_{ij}^{q} < \epsilon\} .$$

6.3 qp-Graphs estimates of eQTL networks are enriched for *cis*-eQTL associations

In Section 5.3 we proved the importance of testing conditional independences with an exact test against the traditional asymptotic test when looking for an eQTL-association between a genetic marker and a gene. Probing higher-order eQTL associations in a genetical genomics data set can be exploited in a number of ways. One of them, consists of systematically testing for conditional independences of order q and using the expected number of non-rejections ν_{ij}^{q}, known as non-rejection rate (NRR) (Castelo and Roverato, 2006), to estimate a qp-graph $\hat{G}^{(q)}$, as an approximation to the underlying eQTL network.

Expression QTL acting in *cis* have more direct mechanisms of regulation than those acting in *trans* (Rockman and Kruglyak, 2006; Cheung and Spielman, 2009). This view is supported by the observation that *cis*-acting eQTLs, on average, explain a larger fraction of expression variance and show larger additive effects, than those acting in *trans* (Rockman and Kruglyak, 2006; Petretto et al., 2006; Cheung and Spielman, 2009).

On the other hand, spurious eQTL associations tend to inflate the discovery of *trans*-acting eQTLs (Breitling et al., 2008). From this perspective, it makes sense to expect an enrichment of *cis*-eQTLs when indirect associations are effectively discarded (Kang et al., 2008a; Listgarten et al., 2010).

We estimated NRR values ν_{ij}^{q} on every pair (i, j) of marker and gene from the yeast data set of $n = 112$ segregants for different $q = \{25, 50, 75, 100\}$ orders, restricting conditioning subsets to be formed by genes only. The resulting estimates $\hat{\nu}_{ij}^{q_k}$, $q_k \in q$, were averaged, $\hat{\nu}_{ij}^{\bar{q}} = \frac{1}{|q|} \sum_{q_k} \hat{\nu}_{ij}^{q_k}$, to account for the uncertainty in the choice of the conditioning order q (Castelo and Roverato, 2009).

We ranked marker-gene pairs (i, j) by average NRR values $\hat{\nu}_{ij}^{\bar{q}}$ and made a comparison against the ranking by p-value of the (exact) LRT for marginal independence (i.e., where $q = 0$) to directly assess the added value of higher-order conditioning under the same type of statistical test. We considered conservative and liberal cutoff values $\epsilon = \{0.1, 0.5\}$ on the average NRR $\hat{\nu}_{ij}^{\bar{q}}$ and obtained two different qp-graph estimates of the underlying eQTL network, denoted by $\hat{G}_{\epsilon}^{(\bar{q})} = (V, E_{\epsilon}^{(\bar{q})})$, each of them having $|E_{0.1}^{(\bar{q})}| = 4,667$ and $|E_{0.5}^{(\bar{q})}| = 81,360$ edges.

We then selected the top-k number of marker-gene pairs (i, j) with the lowest p-value in the marginal independence test, where $k = |E_{\epsilon}^{(\bar{q})}|$, which led to two other estimates of the eQTL network, denoted by $\hat{G}_{\epsilon}^{(0)}$. Note that both, $\hat{G}_{\epsilon}^{(\bar{q})}$ and $\hat{G}_{\epsilon}^{(0)}$, have the same number of edges, in this case, pairs (i, j) of eQTL associations between a marker and a gene, thereby enabling a direct comparison of the fraction of *cis* and *trans*-acting selected eQTL associations.

In Figure 6.3 we can see dot plots of the eQTL associations present in qp-graphs $\hat{G}_{\epsilon}^{(\bar{q})}$ (panels A, C), and those present in $\hat{G}_{\epsilon}^{(0)}$ by using the marginal approach (panels B, D). Given the same number of eQTL associations, Figure 6.3 shows that qp-graph estimates of the underlying eQTL network have a higher number of *cis*-acting eQTLs, with an increase between 25% and 58% over the marginal test (see Table 6.1). Moreover, with the marginal approach many more vertical bands of *trans*-acting associations remained present among the strongest selected eQTLs (panel D), than with the qp-graph estimates (panel C). We interpret this observation as evidence of the propagation of additive effects due to strong gene-gene correlations either present in the underlying eQTL network or created by confounding effects, and possibly aggravated by selection bias, as previously shown in Figure 4.6.

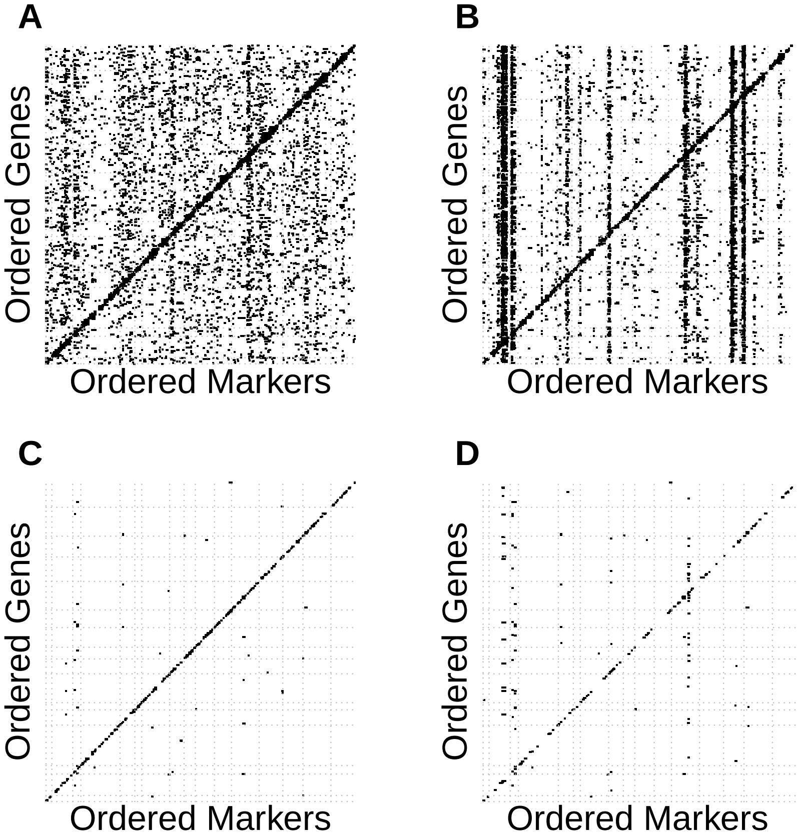

Figure 6.3: Enrichment of *cis*-acting eQTL associations. Dot plots of eQTL associations in yeast, where the x-axis and y-axis represent positions along the genome of markers and genes, respectively. Diagonal bands arise from *cis*-eQTLs while vertical ones from *trans*-eQTLs. Each row shows the top-k eQTLs with largest strength in terms of non-rejection rates (A, C) and p-values for the null hypothesis of marginal independence (B, D), where k was the number of eQTLs meeting a liberal (A) and conservative (C) cutoff on the non-rejection rate. Hence, panels in each row contain the same number of eQTLs. Conditioning with the approach introduced in this chapter (A, C) leads to more *cis*-acting eQTLs than using marginal tests (B, D); see Table 6.1.

Finally, we wanted to investigate which qp-graphs $\hat{G}^{(q)}$ are obtained when the conditioning sets Q are small compared to the biggest possible order, for instance, $q = 5$. As it can be shown in Figure 6.4A and

Figure 6.4C, when the q order is too small, the estimated qp-graph is not able to adjust for indirect associations even when a conservative NRR cutoff ϵ is applied. In this case, we cannot observe a *cis*-enrichment of the eQTL associations and, furthermore, the marginal and the conditional approaches do not show significant differences.

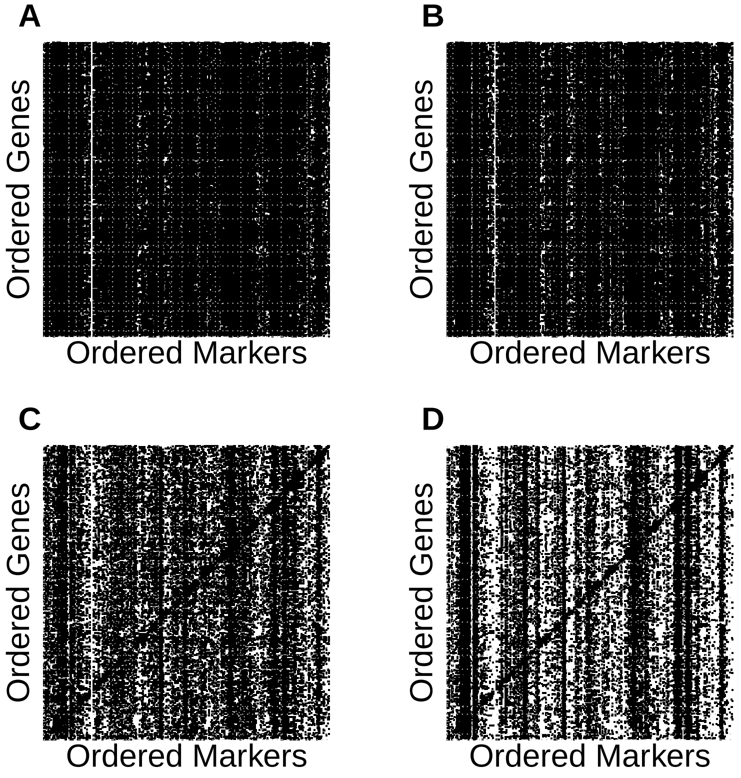

Figure 6.4: qp-Graphs of small order $q = 5$. Dot plots of eQTL associations in yeast, where the x-axis and y-axis represent physical genome positions of markers and genes, respectively. Each row shows the top-k eQTLs with largest strength in terms of NRR values (A, C) and p-values for the null hypothesis of marginal independence (B, D), where k was the number of eQTLs meeting a liberal (A) and conservative (C) cutoff on the NRR. Conditioning sets of small order do not adjust for indirect eQTL associations.

Table 6.1: Enrichment of *cis*-eQTL associations. Number of *cis*-eQTL associations in yeast found by the method introduced in this chapter (qp-graph) and by a marginal test of independence, indicated as row names. The third row reports the enrichment of *cis*-eQTLs of qp-graph over the marginal approach. Different columns correspond to different cutoffs (conservative, liberal) employed by qp-graph to select eQTLs and different distances (500bp and 10kb) around genes to call eQTL as *cis*-acting. Using higher-order conditional independences (qp-graph) yields between 26% and 58% more *cis*-eQTL associations in the yeast data set, depending on the minimum strength of eQTLs and their maximum distance to their associated genes.

	Conservative cutoff (4,667 eQTLs)		**Liberal cutoff** (81,360 eQTLs)	
Method	*cis* **dist. 500bp**	*cis* **dist. 10kb**	*cis* **dist. 500bp**	*cis* **dist. 10kb**
qp-graph	278	1,998	911	8897
marginal	221	1,367	646	5626
Enrichment	26%	46%	41%	58%

6.4 Higher-order conditioning leads to sparser eQTL networks with more direct *trans*-acting associations

Gene expression is often influenced by several *trans*-acting regulators. The genetic variability associated to each of these regulators normally makes a small contribution to the overall genetic effect that modulates the transcriptional throughput of the target gene (Cheung and Spielman, 2009). This makes it even harder to find genuine direct *trans*-acting eQTLs because small additive effects may also result from the propagation of large effects through gene-gene correlations and selection bias (see Figure 4.6).

We explored the use of higher-order conditioning to filter out spurious *trans*-acting eQTLs selected by classical QTL mapping with LOD scores. To this end, we first conducted single marker regression analysis with the `R/qtl` package (Broman and Sen, 2009) to identify *trans*-eQTLs located at least 500bp away from the linked gene. Using permutation tests from `R/qtl`, p-values were calculated and 31,478 eQTLs met a genome-wide cutoff of $p < 0.01$, corresponding to a minimum LOD score of 4.32.

Among these eQTLs, 535 were *cis*-acting and the remaining *trans*-eQTLs were associated to 2,416 different genes. Note that this estimate of the eQTL network corresponds to a qp-graph estimate with $q = 0$, $\hat{G}^{(0)}$, based on a permutation test with a null hypothesis for each gene of no eQTL anywhere in the genome.

Using average NRR estimates $\hat{\nu}_{ij}^{\bar{q}}$ calculated in the previous subsection and a conservative NRR cutoff value of $\epsilon = 0.1$, we selected a qp-graph estimate $\hat{G}_{0.1}^{(\bar{q})} \subseteq \hat{G}^{(0)}$ which only had 361 genes with at least one *trans*-eQTL, from the initial set of 2,416. Recall from previous sections, that this means that for each eQTL in $\hat{G}_{0.1}^{(\bar{q})}$, on average through the different $q = \{25, 50, 75, 100\}$, at least 90% of LRTs reject the null hypothesis of q-order conditional independence.

Among the 361 genes in $\hat{G}_{0.1}^{(\bar{q})}$ with at least one *trans*-eQTL, only 12 had exactly the same eQTLs in the initial estimate $\hat{G}^{(0)}$ obtained by single marker regression. For each of the remaining 349, we compared the following two linear models with the expression profile of each gene, denoted by Y_γ, as response variable and the linked *trans*-acting eQTLs, denoted by $I_1, \ldots, I_l$, as explanatory factors:

$$\begin{array}{rcl} H_1: \; Y_\gamma & = & \beta_0 + \beta_1 I_1 + \cdots + \beta_k I_k + \beta_{k+1} I_{k+1} + \cdots + \beta_l I_l + \epsilon \,, \\ H_0: \; Y_\gamma & = & \beta_0 + \beta_1 I_1 + \cdots + \beta_k I_k + \epsilon \,. \end{array}$$

Since the linear models derived from $\hat{G}_{0.1}^{(\bar{q})}$, and corresponding to H_0, are nested into those derived from $\hat{G}^{(0)}$, we can test for each gene Y_γ whether the models from $\hat{G}^{(0)}$ explain a significantly larger amount of variance than the ones from $\hat{G}_{0.1}^{(\bar{q})}$ using an F-test.

Figure 6.5A shows the distribution of the resulting 349 p-values. For the vast majority of them (83% with $p > 0.05$) we cannot reject the null hypothesis at a reasonable significance level, and therefore, the sparser model derived from the qp-graph $\hat{G}_{0.1}^{(\bar{q})}$ should be preferred. We repeated again this exercise replacing the eQTLs in each gene from $\hat{G}_{0.1}^{(\bar{q})}$ by randomly chosen ones among those that form part of the larger model for the same gene in $\hat{G}^{(0)}$. The result, in Figure 6.5B, reveals that in comparison with the eQTLs selected in $\hat{G}_{0.1}^{(\bar{q})}$, fewer of the null (random) models (57% with $p > 0.05$) fit the data as good as the alternative larger models. Note that these null random models are still built with eQTLs significantly linked to their gene by single-marker regression, which may explain a substantial fraction of the gene expression variability in the 43% of the models with $p < 0.5$.

The qp-graph estimate $\hat{G}^{(\bar{q})}_{0.1}$ had 2,055 genes for which all their *trans*-acting eQTLs were removed from the initial $\hat{G}^{(0)}$. This means that at least 10% of higher-order conditional independence tests could not reject the null hypothesis between every of these 2,055 genes and any of the *trans*-eQTLs significantly linked to them by single-marker regression at a genome-wide $p < 0.01$. The explanation for this discrepancy is that conditioning on their associated genes, the linked eQTLs do not explain a significantly larger fraction of the variability of the target gene. We verified this hypothesis using an analogous strategy to the previous case with the 349 genes. However, this time we could only consider the fraction of the 2,055 eQTL genes (465/2,055) that had at least one gene-gene association in $\hat{G}^{(\bar{q})}_{0.1}$. For every of these 465 genes, we added to the null and alternative models, those genes $Y_1, \ldots, Y_k$ in $\hat{G}^{(\bar{q})}_{0.1}$ connected to the target gene Y_γ, i.e., with $\hat{\nu}_{\gamma j} < 0.1, j = 1, \ldots, k$:

$$\begin{aligned} H_1: \; Y_\gamma &= \beta_0 + \beta_1 Y_1 + \cdots + \beta_k Y_k + \beta_{k+1} I_{k+1} + \cdots + \beta_l I_l + \epsilon \,, \\ H_0: \; Y_\gamma &= \beta_0 + \beta_1 Y_1 + \cdots + \beta_k Y_k + \epsilon \,. \end{aligned}$$

Figure 6.5C shows the distribution of p-values of the F-test between the previous two models for the 465 genes and 67% of them had $p > 0.05$. We performed an analogous control to the one used before, this time replacing the genes $Y_1, \ldots, Y_k$ by randomly select ones among the rest in the data set. The resulting p-value distribution shown in Figure 6.5D reveals that most of tests could be rejected, and therefore, the *trans*-eQTLs identified by single-marker regression do explain a significantly larger fraction of the variability of the target gene, than just using gene expression from randomly selected genes.

Broader cis-window

Lastly, we verified if the stringent definition of the *cis*-window used in the previous section affected somehow the obtained results. Therefore, we repeated the analysis but using a less strict definition of *cis*-eQTL. In particular, we assumed that an eQTL is acting in *cis* if it is located within a region of 10 kb from the gene as it was defined in (Brem et al., 2002).

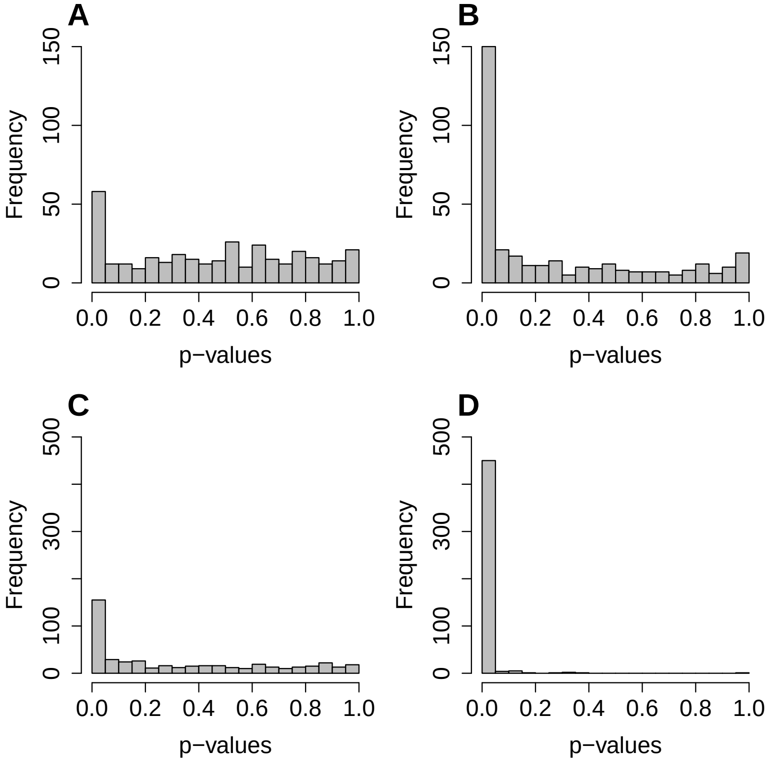

Figure 6.5: Fit of gene models with *trans*-eQTLs to yeast data. Distribution of p-values for the F-test between sparser qp-graph $\hat{G}^{(\bar{q})}_{0.1}$ (null) models and larger single-marker (alternative) models. Null qp-graph models were derived from genes γ with *trans*-eQTLs δ found significant by single-marker regression at a genome-wide $p < 0.01$ and discarding those with NRR $\hat{\nu}^{\bar{q}}_{\gamma\delta} > 0.1$. Panel (A) contains p-values from 349 genes with at least one *trans*-eQTL in $\hat{G}^{(\bar{q})}_{0.1}$. Panel (B) results from using the same alternative $\hat{G}^{(0)}$ models as in (A) but replacing eQTLs δ in null models by different ones δ' randomly selected among those in the alternative models. Panel (C) contains p-values from 465 genes γ with no *trans*-eQTL in $\hat{G}^{(\bar{q})}_{0.1}$ but connected to at least one other gene in $\hat{G}^{(\bar{q})}_{0.1}$. In this case, genes η in $\hat{G}^{(\bar{q})}_{0.1}$ whose $\hat{\nu}^{\bar{q}}_{\gamma\eta} < 0.1$ where included in both, the null and the alternative models. Panel (D) results from using the same alternative $\hat{G}^{(0)}$ models as in (C) but replacing genes η in null models by different ones η' randomly selected among the rest of the genes in the data set. The vast majority of tests in (A, C) have $p > 0.05$, thus indicating that denser alternative gene models derived from single-marker regression $\hat{G}^{(0)}$ do not fit the data significantly better than the sparser null models derived from qp-graph estimates $\hat{G}^{(\bar{q})}_{0.1}$. This fact changes substantially in the control experiments shown in (B, D).

According to this definition, we found 3,072 *cis*-eQTLs and 28,406 *trans*-eQTLs that met a genome-wide cutoff of $p < 0.01$ by using permutation tests from `R/qtl`. *Trans*-eQTLs are associated to 2,357 different genes.

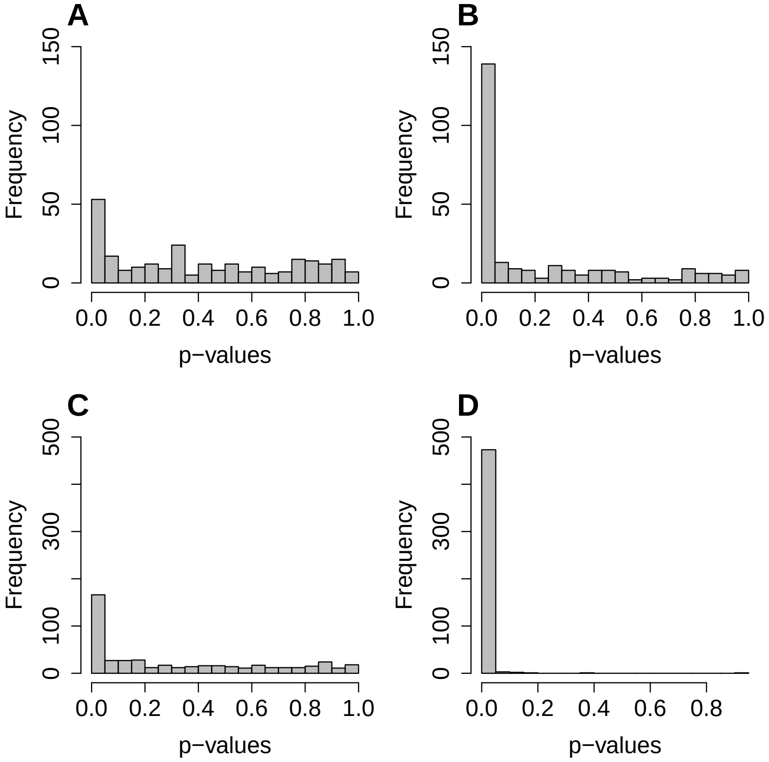

Figure 6.6: Fit of gene models with trans-eQTLs to yeast data and broader *cis*-region. Distribution of p-values obtained from F-tests comparing sparser qp-graph $\hat{G}_{0.1}^{(\bar{q})}$ (null) models and larger single-marker (alternative) models. Here, the *cis*-window contains eQTLs up to 10 kb far from the target gene. A) Distribution of p-values from the F-tests associated to 269 genes that have at least one *trans*-eQTL with a NRR estimate $\nu_{ij}^{\bar{q}} < 0.1$. C) Distribution of p-values from the F-tests associated to 481 genes that do not have any *trans*-eQTL with a NRR estimate $\nu_{ij}^{\bar{q}} < 0.1$. B-D) Analogously to panels C and D of Figure 6.5, these panelos show the distribution of p-values from the control F-tests.

We proceeded as before and, in this case, only 6 out of 2,357 genes of the estimated qp-graph $\hat{G}_{0.1}^{(\bar{q})}$ had the same *trans*-eQTLs as in $\hat{G}_{0.1}^{(0)}$ obtained by single marker regression. For 263 genes, the number of *trans*-eQTLs associated to them was smaller in $\hat{G}_{0.1}^{(\bar{q})}$ than in $\hat{G}_{0.1}^{(0)}$ and, for the remaining

2,088 genes, all the *trans*-eQTLs significantly linked to them according to the single marker regression approach had NRR estimates $\nu_{ij}^{\bar{q}}$ larger than the cutoff $\epsilon = 0.1$.

We performed F-tests as in the previous section. In the histogram of Figure 6.6A, the p-values obtained from the 263 F-tests comparing the null and the alternative models are shown. In this case, the alternative model fit better the data ($p < 0.05$) only in 20% of the cases. Further, in Figure 6.6C, we observe that 35% of 481 genes (the ones that had at least one gene-gene association with a NRR estimate below $\epsilon = 0.1$) had $p < 0.05$. Analogously to the previous section, we performed control F-tests. From Figures 6.6B-D, we can observe that the results do not change significantly from those reported in the previous section and the percentages of significant p-values ($p < 0.05$) are clearly much larger than the ones of Figures 6.6A-C, respectively.

These results mainly indicate that, in general, *trans*-eQTLs selected by the single marker regression do not explain a significantly larger fraction of the variability of the target gene expression profiles than the eQTLs identified by the non-rejection rate. Further, the *trans*-eQTLs identified by the non-rejection rate explain better the variation of the gene expression profiles regardless of the definition of the *cis*-window.

Functional analysis of trans-eQTLs

The striking differences in Figure 6.5 between panels (A, C) and (B, D) confirm, from a purely statistical standpoint, that higher-order conditioning can effectively help to discard indirect *trans*-acting eQTL associations. However, they tell little about differences in biological information conveyed by sparser eQTL network estimates which, *a priori*, fit the data better.

We attempted to address this question in a systematic way by adapting an approach previously used with transcriptional networks (Castelo and Roverato, 2009) for this same purpose. This approach estimates the degree of coherence between the function of a transcription-factor coding gene and its putative targets using Gene Ontology (Ashburner et al., 2000) (GO) annotations. This degree of functional coherence (FC) takes values between 0 and 1, where 1 implies identical biological function and 0 completely different. Assuming genes acting closer in pathway should exert more similar functions than those acting far apart, one should expect that FC estimates of direct associations are closer to 1 than those calculated from indirect ones.

To enable this kind of analysis with eQTL associations we restricted it to those overlapping a gene and assumed that the GO terms of this gene describe the functions most directly affected by the genetic variability of the eQTL. We also fetched the GO terms of the gene *trans*-associated to the eQTL. Finally, an FC estimate for each *trans*-eQTL was obtained by comparing GO hierarchies growing, in one hand, at the GO terms from the eQTL-overlapping gene with, on the other hand, the GO terms annotated at the gene. This comparison was done as the ratio between the intersection of GO terms between the two hierarchies divided by their union Castelo and Roverato (2009). Note that GO annotations themselves are partial and, in some cases, inaccurate, but one may expect that the large number of available annotations for an organism such as yeast still enable this approach.

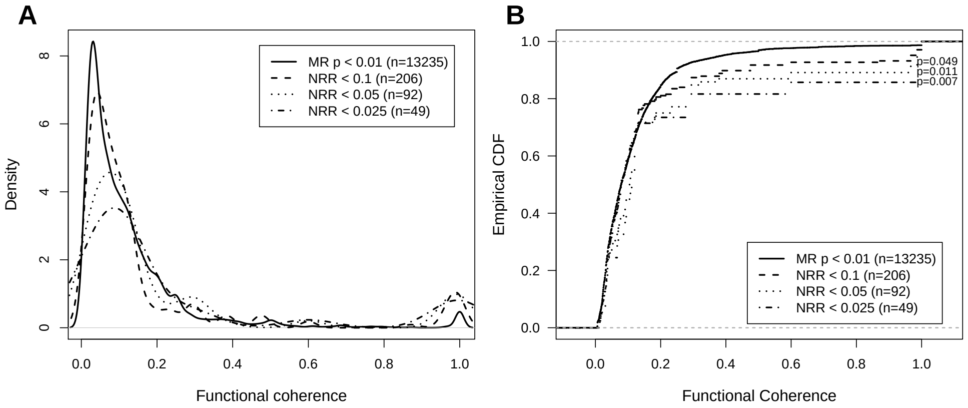

Figure 6.7: Functional coherence of *trans*-eQTLs. Densities (A) and empirical cumulative distribution functions -CDFs- (B) of estimated values of functional coherence (FC) of *trans*-acting eQTL associations selected by single-marker regression (MR) at a genome-wide $p < 0.01$ and by different cutoffs of the non-rejection rate -NRR- (0.1, 0.05, 0.025). The legend indicates the number n of eQTLs for which FC could be estimated. In panel (B), at the height of each line for FC=0.8, the p-value for the Kolmogorov-Smirnov (KS) test on whether the corresponding CDF of NRR values is stochastically larger than the one of MR values, is reported. It follows that FC values increase significantly ($p < 0.05$) when restricting MR *trans*-eQTL associations to those selected by NRR and higher-order conditioning.

We proceeded to calculate FC values on *trans*-eQTLs from the eQTL network estimated by single-marker regression with genome-wide $p < 0.01$, $\hat{G}^{(0)}$, and from qp-graph estimates derived from three different NRR cutoffs, $\hat{G}^{(\bar{q})}_{0.1}, \hat{G}^{(\bar{q})}_{0.05}, \hat{G}^{(\bar{q})}_{0.025}$. To avoid having results depending on the minimum distance employed to call an eQTL as *trans*-acting, we further selected only those eQTLs whose target genes were located in a different chromosome.

Figure 6.7 shows the densities and empirical cumulative distribution functions (ECDF) of FC values from each eQTL network estimate. It can be seen that as the NRR cutoff decreases, the FC distribution shifts towards larger values. This shift is significant with respect to the single-marker regression estimate, when testing the difference in ECDFs by a Kolmogorov-Smirnov test ($p < 0.05$). This implies that larger fractions of rejected conditional independence tests lead to larger FC values which, in turn, suggest that eQTLs with small (stronger) NRR values are in some sense functionally "closer" to the target gene than eQTLs with weaker (larger) NRR values.

From the statistical and biological evidence gathered in this subsection we may conclude that higher-order conditioning leads to sparser eQTL networks with more direct *trans*-acting associations.

6.5 An estimate of the yeast eQTL network

We took a closer look to the qp-graph estimate $\hat{G}^{(\bar{q})}_{0.1}$ of the yeast eQTL network. First, we focused on the genetic connected components involving at least one *cis* or *trans*-acting eQTL association. These components involved 379 genes and 288 eQTLs, with a median of 4 eQTls per gene. A significant percentage of genes (20%) had more than 10 eQTLs on the same chromosome of the linked gene. Since eQTLs in $\hat{G}^{(\bar{q})}_{0.1}$ were independently selected from each other, a fraction of those targeting a common gene may be tagging the same causal variant. We removed redundant eQTLs by the following forward selection procedure. For each gene, we ordered its linked eQTLs by increasing NRR values $\nu^{(\bar{q})}_{ij}$ and proceeded over the ranked eQTLs to test the conditional independence of the gene and the eQTL, given the eQTLs ocurring before in the ranking. An eQTL association was retained if the test was rejected at $p < 0.05$ and the selection procedure stopped whenever $p > 0.05$ to continue on the next gene.

The genetic connected components were substantially pruned and the vast majority of genes (328/379) were left with just one eQTL, 50 genes with two, and only one gene had 3 eQTLs. This final eQTL network comprised 288 eQTLs and 1,295 genes, the vast majority of them (916) forming gene-gene associations without any eQTL.

Using the previous forward selection strategy, this time without testing and dropping any eQTL, on the 379 genes with at least one eQTL, we applied Equation (5.14) to estimate the percentage of variance explained by eQTLs at each gene, adjusting for the presence of multiple eQTLs in the case of the 51 genes with more than one. The distribution of resulting values is shown in Figure 6.8A. About half of the genes had eQTLs explaining 50% or less of their expression variability, and only in about 10% their eQTLs explained more than 70% of it. The small fraction of genes (6%) with eQTLs explaining less than 20% of their variance is consistent with the adjustment of small effects due to indirect associations (Figure 4.5).

Using a method based on linear mixed modeling and the relatedness matrix between all pairs of segregants (Lee et al., 2011; Bloom et al., 2013) we estimated the narrow-sense heritability h^2 for these 379 genes and compare it against the percentage of variance explained by the eQTLs (Figure 6.8B). Setting the percentage explained to the expected maximum h^2 when the former was larger, the fraction of missing heritability ranges from 0 to 73%. We labeled genes in this figure by their connectivity degree to other genes in the eQTL network. We observed that this degree correlated positively with both, h^2 and percentage of variance explained. In fact, as Figure 6.9 shows, genes whose eQTLs explained more than 70% of their variability were connected to 9 or more other genes in the eQTL network. This means that the larger genetic control on gene expression takes place on those genes acting as hubs in the network. These highly-connected genes, shown in Table 6.2, are involved in regulatory processes that help yeast cells finding mating partners (*STE2*, *STE3*, *STE6*), react upon nitrogen starvation (*ASP3-1*, *ASP3-2*, *ASP3-3*) and participate in the leucine biosynthesis pathway (*LEU1*, *BAT1*, *OAC1*). These are fundamental pathways for yeast growth, and therefore, we are able to recapitulate with genetical genomics data previous evidence from yeast genetic interaction networks derived with double-mutant screens (Costanzo et al., 2010; Baryshnikova et al., 2013), where highly-connected genes were involved in primary cellular functions.

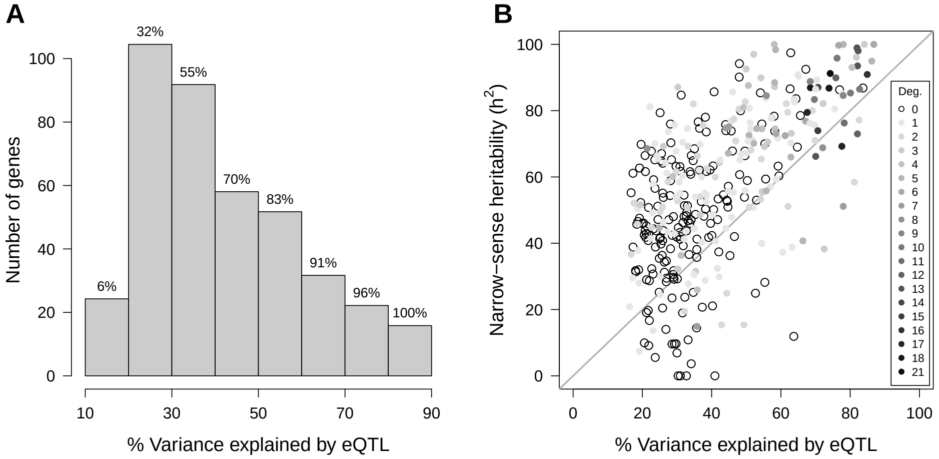

Figure 6.8: Variance explained in the eQTL network. (A) Distribution of the percentage of gene expression variance explained by eQTLs. The cumulative percentage of genes is reported on top of each bar. A majority of genes has eQTLs that together explain less than 40% of the expression variance. Only about 10% of the genes have eQTLs that explain more than 70% of their expression variance. (B) Scatter plot of the narrow-sense heritability h^2 as function of the percentage of variance explained by eQTLs. The diagonal line is drawn at values where this percentage equals h^2, and it is only shown as a visual guide. Open circles correspond to genes with exclusively eQTL associations while solid ones indicate also the presence of at least one association to other gene. The grayscale in solid circles correlates with the connectivity degree of gene-gene associations in the eQTL network, as indicated in the legend. Both, h^2 and variance explained by eQTLs, is higher for genes with more gene-gene associations.

We also investigated how genetic variation affects gene expression differently across the yeast chromosomes. For this purpose, we produced hive plots (Krzywinski et al., 2012), one for each chromosome, shown in Figure 6.10. A first observation is that eQTLs ocurring within the same chromosome (edges between the markers and *cis* genes axes) mostly lead to concentric edges, pointing to *cis* regulatory mechanisms acting at different distances. A remarkable exception is chromosome III where many of those edges cross through each other. This chromosome is also distinctive in that it has a lower density of *cis*-acting eQTLs than the rest of the genome.

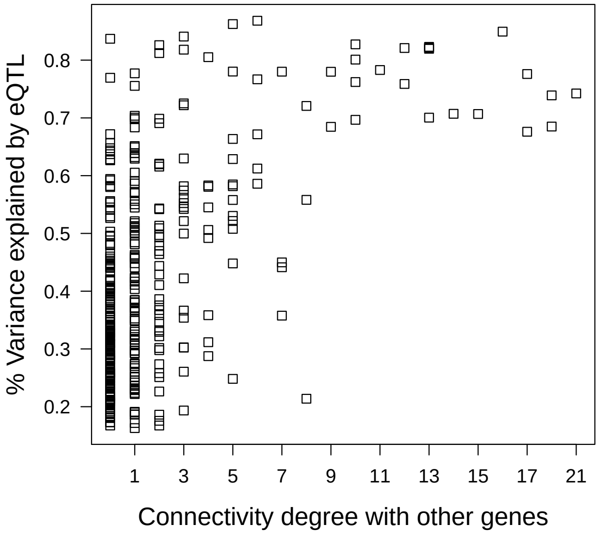

Figure 6.9: Percentage of variance explained by eQTLs as function of the connectivity degree with other genes in the eQTL network. Degrees 19 and 20 had no genes and are omitted from the x-axis. Genes whose eQTLs explain more than 70% of their expression variance are connected to at least 9 other genes in the eQTL network.

More concretely, around 200 kb of chromosome III, we identify the *MAT* locus whose genetic composition determines the mating type of yeast. This eQTL is *cis*-associated to the gene *MATALPHA1* which is expressed in haploids of the alpha mating type and which has been previously reported as a candidate regulator of the rest of genes associated to this locus (Yvert et al., 2003; Curtis et al., 2013). Concretely, we find this locus *trans*-associated to two other genes in the same chromosome (*HMLALPHA1*, *HMRA1*) and to a set of genes distributed throughout the genome (*STE2*, *STE3*, *STE6*, *AFB1*, *BAR1*, *MF(ALPHA)1*, *MFA2*) which are all involved in the regulation of mating-type specific transcription. We also find an eQTL located around 100 kb of chromosome III which is *cis*-associated to the gene *LEU2*. This eQTL is also *trans*-associated to the genes *BAT1*, *OAC1* and *BAP2*, which are downstream of a binding site of *LEU3*. This gene is a major regulatory switch in the pathway of *LEU2*, whose activity may be affected by a feedback loop in the pathway (Chin et al., 2008).

Table 6.2: Genes with more than 70% variance explained. Genes whose eQTLs explain 70% or more of their variance.

GeneSymbol	Num. eQTLs	h^2	η^2	Deg. Genes
STE6	1.00	0.91	0.74	21
STE2	1.00	0.87	0.69	18
STE3	1.00	0.87	0.74	18
BAR1	1.00	0.79	0.68	17
YKL177W	1.00	0.69	0.78	17
MATALPHA1	1.00	0.91	0.85	16
MF(ALPHA)1	1.00	0.74	0.71	15
YLR040C	1.00	0.87	0.71	14
YCR097W-A	1.00	0.66	0.70	13
ASP3-1	1.00	0.99	0.82	13
ASP3-2	1.00	0.94	0.82	13
ASP3-3	1.00	0.98	0.82	13
HMLALPHA1	1.00	0.73	0.82	12
LEU1	1.00	0.90	0.76	12
MFA2	1.00	0.76	0.78	11
DSE1	1.00	0.86	0.83	10
SCW11	1.00	0.85	0.80	10
BAT1	1.00	0.83	0.70	10
ASP3-4	1.00	0.96	0.76	10
DSE2	1.00	0.85	0.78	9
OAC1	1.00	0.89	0.68	9

Further, *LEU3* is known to regulate the expression of *BAP2* (Nielsen et al., 2001). Finally, we highlight the hotspot located on chromosome V around 116.6 kb which is *cis*-associated to *URA3* and which, among their *trans*-genes, we identify *URA1* (located on chromosome XI) and *URA4* (located on chromosome XII), all of them taking part in the biosynthesis of pyrimidines (Yvert et al., 2003; Curtis et al., 2013). Few of the eQTLs affect directly transcription factors, such as the *ARR1* gene in chromosome XI, or RNA-binding proteins, such as *NOP8* in chromosome V.

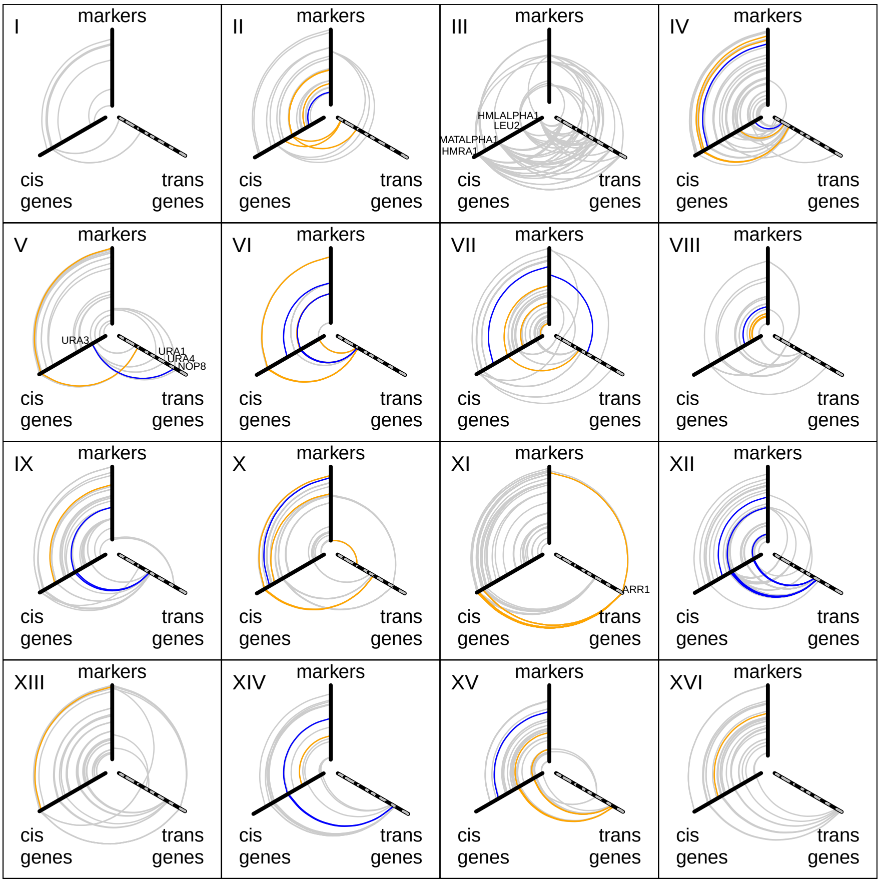

Figure 6.10: eQTL network of yeast. A qp-graph estimate $\hat{G}_{0.1}^{(\bar{q})}$ of the eQTL network in yeast, involving only connected components with at least one *cis* or *trans*-acting association, is shown by means of hive plots. For each chromosome, the hive plot shows three axes, where markers and genes are ordered from the centre according to their genomic location. Vertical and left axes represent the chromosome in the corresponding panel, while the right axis represents the entire yeast genome alternating black and gray along consecutive chromosomes. In these plots we label the left and right axes as *cis* and *trans* to merely indicate the same chromosome or the entire genome, respectively. Edges between axes labeled as markers and *cis* genes connect eQTLs and genes located in the same chromosome, whereas edges between the markers axis and the *trans* genes axis correspond to *trans*-eQTLs affecting genes in different chromosomes. Edges between *cis* genes and *trans* genes axes correspond to gene-gene associations in $\hat{G}_{0.1}^{(\bar{q})}$. Edges whose at least one of their endpoints correspond to a transcription factor or RNA-binding coding gene, are highlighted in orange and blue, respectively.

6.6 Concluding remarks

In this chapter we have presented a new approach to learn eQTLs networks based on mixed GMMs. The method per se, is based on testing conditional independences between genotypes and gene expression profiles and also between two gene expression profiles while controlling for other genotypes, gene expression profiles and other phenotypic variables. We have applied this approach to a real genetical genomics data set from yeast. These are the main conclusions of this work:

- In order to learn eQTL networks, we have extended a strategy introduced in the context of Gaussian GMMs to learn mixed GMMs. By applying this procedure, we obtain a class of undirected marked graphs, called qp-graphs, $\left(G^{(q)}\right)$ which are non-decomposable and whose edges represent q-order conditional dependencies.

- We have verified that qp-graphs estimates of the underlying eQTL network of yeast effectively discard spurious indirect associations. This is reinforced by the observation of an enrichment of *cis*-eQTLs.

- We have also verified that *trans*-eQTLs in our qp-graph estimates fit better the target gene expression profiles than *trans*-eQTLs identified by the single marker regression approach.

- As a consequence, we have proved that our multivariate approach based on an exact conditional independence test is better at identifying more direct *trans*-eQTL associations. Moreover, we have demonstrated that these *trans*-eQTL associations are also more functionally related than the ones identified by the single marker regression approach.

- We estimated the qp-graph representing the eQTL network from yeast and provide an intuitive representation of the eQTL network by means of a three-axis panel for each chromosome. From this estimated qp-graph, we have described some features of the genetic architecture of gene expression in yeast. We have shown that genes with a high estimated heritability and a high percentage of the variance explained by the eQTLs are acting as hubs in the gene network and are involved in fundamental pathways for yeast growth.

7. Learning eQTL networks from data with missing observations

7.1 Introduction

A common obstacle in eQTL mapping studies is that genetical genomics data often exhibit missing values. Two approaches are mostly used to learn eQTL networks from data with missing observations.

The complete-case analysis, which involves the deletion of observations with missing values, is a straightforward approach. However, the large number of SNPs that are typically genotyped in a genetical genomics data set increases the likelihood of having at least one missing value at each observation. Since in these high-dimensional data sets, the number of SNPs greatly exceeds the number of samples, too few samples may be left complete and ready to directly apply some multivariate eQTL mapping method. Such a decrease in the number of available samples lead to significant reductions in the detection power and mapping resolution.

On the other hand, imputation approaches such as the one proposed in Broman et al. (2003) for genotype data from experimental crosses, enable using existing learning algorithms for complete data. In this approach, genotypes are imputed conditioning on observed genotype data. Then, each imputed data set is analyzed separately and the results are averaged. A major challenge when directly applying this approach is that multiply-imputed data sets are required to preserve the sampling properties of the data and reduce the biases produced by imputed genotypes. However, this leads to a substantial increase in the computational cost of eQTL mapping methods.

In this chapter, we show that using limited-order correlations in the $p \gg n$ setting permits a straightforward and effective application of complete-case analysis (Section 7.2).

Moreover, we also provide an expectation-maximization (EM) algorithm to accurately estimate the likelihood ratio statistic. Obtaining accurate estimations of the likelihood ratio statistic may be of special interest to calculate the fraction of the phenotypic variance explained by an eQTL, denoted by η^2 (see Equation 5.15). The details of this algorithm are provided in Section 7.3.

In Section 7.4 we describe the procedure to simulate genetical genomics data with missing values. Concretely, we focus on the case where only discrete values are unobserved corresponding to missing genotype values. In fact, this is the most common situation when inferring associations between continuous gene expression profiles and discrete genotypes since gene expression values are never missing. In Section 7.5 we assess the accuracy of the estimations of the likelihood ratio statistic calculated with the complete-case approach and the EM algorithm. The performance of the complete-case analysis when calculating the non-rejection rate is illustrated in Section 7.7. Finally, the main conclusions of the research are provided in Section 7.8.

7.2 Complete-case analysis for the non-rejection rate

We consider data that follow a homogeneous CG-distribution where discrete and/or continuous r.v.'s may have missing values. We denote by $x_{\text{obs}} = (i_{\text{obs}}, y_{\text{obs}})$ the non-missing components of an observation from such distribution. We assume that at least one component of every observation $x^{(\nu)}$ is not missing so that $x^{(\nu)}_{\text{obs}} \neq \emptyset$. In particular, if all the components of an observation are not missing, we have that $x^{(\nu)} = x^{(\nu)}_{\text{obs}}$ and $x^{(\nu)}_{\text{mis}} = \emptyset$.

In order to calculate the non-rejection rate, we use marginal distributions of size $(q+2) < n$ which facilitates the use of complete-case analysis. This is done by replacing $\mathcal{I}_A$ in Equation (4.11) by $\mathcal{I}^{\text{obs}}_A$ such that $\mathcal{I}^{\text{obs}}_A \subseteq \mathcal{I}_A$ contains only the combined levels from $A \cap \Delta$ that are fully observed. Analogously, for missing continuous values, we replace $y^{(\nu)}$ by their observed components $y^{(\nu)}_{\text{obs}}$ in Equations (4.7), (4.8) and, (4.9).

Complete-case analysis is appropriate under the assumption that data is *missing completely at random* (MCAR). In this case, the occurrence of a missing value does not depend on other observed nor unobserved values. Complete-case analysis leads to biased estimates of the parameters otherwise (Little and Rubin, 2002). Since genotypes do not depend on gene expression profiles and the occurrence of missing genotypes does not depend on the genotypes at other genetic loci, it is reasonable to assume that missing genotype values occur under the mechanism of MCAR.

7.3 An EM algorithm to estimate the likelihood ratio statistic

The expectation-maximization (EM) algorithm (Dempster et al., 1977) is a method to find the MLEs in statistical models when these models depend on unobserved latent variables or missing data observations. The EM algorithm is an iterative method which alternates two steps until convergence is achieved:

1. Expectation (E) step. The E-step calculates the expectation of the sufficient statistics given the observed data and the current estimation of the parameters.

2. Maximization (M) step. The M-step determines the new estimates by maximizing the conditional expectation of the sufficient statistics calculated in the E-step.

In the context of learning the parameters of a mixed GMM from a data set with missing values and $n \gg p$, Didelez and Pigeot (1998) use the EM algorithm to provide the MLEs of the mixed GMM under the assumption that missing values are generated according to the *missing at random* (MAR) mechanism. Under this assumption, a value is missing or not depending on the other observed values but not on unobserved ones (Little and Rubin, 2002). Geng et al. (2000) developed a more efficient EM algorithm (PIEM) for decomposable mixed GMMs.

Here, we apply the EM algorithm to estimate the likelihood ratio between a saturated and a constrained decomposable mixed GMM (see Equations 5.4 and 5.5). Such a ratio involves the calculation of ssd matrices derived from the decomposition of both models. Before describing the application of the EM algorithm to the calculation of the likelihood ratio, let's define some concepts related to the patterns of observed and missing data.

Observed data pattern

Let's consider a data set that follows a homogeneous CG-distribution where discrete and/or continuous r.v.'s may have missing values. We denote the observed data pattern as $\mathrm{T} = \{t_1, t_2, ..., t_L\}$ where t_l is a subset of r.v.'s, $t_l \subseteq V$, such that they are observed for a group of individuals. This implies that the observed data is classified into L groups and for each group l, the set of observed variables is the one defined by t_l. $\{x_{obs}^{(\nu)} = x_{t_l}^{(\nu)}\}$ for $\nu = 1, ..., n_l$, where n_l is the number of observations in the lth group.

Geng et al. (2000) introduced a concept relating the decomposition structure of a graph G with the pattern T of a data set that follows a probability distribution that is Markov with respect to G.

Definition 13. A decomposition (A, B, C) of G is called a *lossless decomposition on* T if each t in T^B contains C, where $\mathrm{T}^B = \{t \in \mathrm{T} : t \cap B \neq \emptyset\}$.

If a triplet (A, B, C) is a lossless decomposition of a graph G on an observed data pattern T, then G can be decomposed losslessly into the subgraphs $G_{A \cup C}$ using data in T and $G_{B \cup C}$ using data in T^B.

The algorithm

The basic strategy consists in applying the EM algorithm to each subset of variables involved in the calculation of the *ssd* matrices composing the likelihood ratios of Equations (5.4) and (5.5).

On the one hand, we apply the EM algorithm to the saturated model to obtain the *ssd* matrix corresponding to the r.v.'s in $\{\alpha, \beta, Q\}$ (see Section 5.2).

Under the constrained model, we apply the EM algorithm to estimate the *ssd* matrix corresponding to the variables composing the clique $C_1 = \{\alpha, Q\}$. The EM algorithm is also applied to estimate the *ssd* matrix corresponding to the r.v.'s composing the clique $C_2 = \{\beta, Q\}$, and finally, the EM algorithm is applied to estimate the *ssd* matrix corresponding to the r.v.'s in the separator set $S = \{Q\}$. However, if for each observation where β is observed, the values in the separator set are also observed, Geng et al. (2000) state that the decomposition $\{\alpha, \beta, Q\}$ is lossless. In this case, the *ssd* matrices corresponding to C_2 and S can be directly estimated using only the samples where β is observed. The detailed algorithm is explained below.

Initialization

First, the moment parameters of the model $\{p_0(i), \mu_0(i), \Sigma_0\}$ are initialized. We use

$$\begin{aligned} p_0(i) &= |\mathcal{I}|/n \quad \text{for each} \quad i \in \mathcal{I}\,, \\ \mu_0(i) &= 0 \quad \text{for each} \quad \gamma \in \Gamma \quad \text{and,} \\ \Sigma_0 &= \mathrm{Id}_{|\Gamma|\times|\Gamma|}\,, \end{aligned}$$

where $\mathrm{Id}_{|\Gamma|\times|\Gamma|}$ stands for the identity matrix of dimension $|\Gamma|\times|\Gamma|$. In the context of genetical genomics data, the percentage of missing genotype values is low and the choice of the initial parameters does not have a critical incidence in the convergence of the algorithm. In our experience, other initial values lead to the same solution.

E-step

The E-step calculates the expectation of the sufficient statistics in Equation (4.6), Equation (4.7) and, Equation (4.9) given the observed data and the current estimation of the parameters for each model:

$$E\{n(i_d)|x_{\text{obs}}\} = \sum_{\nu=1}^{n} pr(I_d = i_d|x_{\text{obs}}^{(\nu)}) = \sum_{\nu=1}^{n} \sum_{i' \in \mathcal{I}: i'_d = i_d} pr(I = i'|x_{\text{obs}}^{(\nu)})\,,$$

where

$$pr(I = i'|x_{\text{obs}}^{(\nu)}) = \frac{\exp k(i')}{\sum_{s \in \mathcal{S}} \exp k(s)} \tag{7.1}$$

and

$$\begin{aligned} k(i) \quad = \quad & y_{\text{obs}}^T \Sigma_{\{\text{obs,obs}\}}^{-1} \mu(i)_{\text{obs}} - \\ & \frac{1}{2}\left[y_{\text{obs}}^T \Sigma_{\{\text{obs,obs}\}}^{-1} y + \mu(i)_{\text{obs}}^T \Sigma_{\{\text{obs,obs}\}}^{-1} \mu_{\text{obs}}(i)\right] + \log p(i). \end{aligned}$$

The set $\mathcal{S} = \{(i_{\text{obs}}, i_{\text{mis}}) | i_{\text{mis}} \in \mathcal{I}_{\text{mis}}\}$ in Equation (7.1) is the set of all combinations of discrete levels given the observed ones. Moreover,

$$\begin{aligned}
E\{s(i_d)_\gamma | x_{\text{obs}}\} &= \sum_{\nu=1}^{n} pr(I_d = i_d | x_{\text{obs}}^{(\nu)}) E(Y_\gamma | y_{\text{obs}}^{(\nu)}, i_d)\,, \\
E\{ss(i_d)_\gamma | x_{\text{obs}}\} &= \sum_{\nu=1}^{n} pr(I_d = i_d | x_{\text{obs}}^{(\nu)}) \times \left[E(Y_\gamma | y_{\text{obs}}^{(\nu)}, i_d)^2 + c_{\gamma\gamma} \right] \text{ and,} \\
E\{ss(i_d)_{\gamma,\eta} | x_{\text{obs}}\} &= \sum_{\nu=1}^{n} pr(I_d = i_d | x_{\text{obs}}^{(\nu)}) \times \\
&\quad \left\{ E(Y_\gamma | y_{\text{obs}}^{(\nu)}, i_d) E(Y_\eta | y_{\text{obs}}^{(\nu)}, i_d) + c_{\gamma\eta} \right\}\,,
\end{aligned}$$

where

$$E(Y_\gamma | y_{\text{obs}}^{(\nu)}, i_d) = \mu(i)_\gamma - \Sigma_{\{\gamma,\text{obs}\}} \Sigma_{\{\text{obs},\text{obs}\}}^{-1} \{y_{\text{obs}} - \mu(i)_{\text{obs}}\}$$

and

$$c_{\gamma\eta} = \text{cov}(Y_\gamma, Y_\eta | y_{\text{obs}}, i) = \Sigma_{\{(\gamma,\eta),(\gamma,\eta)\}} - \Sigma_{\{(\gamma,\eta),\text{obs}\}} \Sigma_{\{\text{obs},\text{obs}\}}^{-1} \Sigma_{\{\text{obs},(\gamma,\eta)\}}\,.$$

M-step

In this step, we update the parameters of the model. To this end, the algorithm maximizes the conditional expectations of the sufficient statistics calculated in the E-step. For the saturated model, this requires calculating the moment parameters $\hat{p}(i)$, $\hat{\mu}(i)$ and $\hat{\Sigma}$ of Equation (4.15) whereas for the constrained model we calculate $\hat{p}(i)$ and the canonical parameters $\hat{h}(i)$ and $\hat{K}$ of Equation (4.16), Equation (4.17) and, Equation (4.18), respectively.

Convergence criteria

In order to check the convergence, the canonical parameters of the constrained model are transformed into moment parameters:

$$\hat{\mu}(i) = \hat{K}^{-1}\hat{h}(i) \quad \text{and} \quad \hat{\Sigma} = \hat{K}^{-1},$$

and the E-step and M-step are iterated until the following condition is satisfied

$$\max_{i \in \mathcal{I}, \gamma, \eta \in \Gamma} \left\{ \frac{|\Delta \hat{m}_i|}{\sqrt{(\hat{m}_i + 1)}}, \frac{|\Delta \hat{\mu}_i^\gamma|}{\sqrt{\hat{\sigma}_i^{\gamma,\gamma}}}, \frac{|\Delta \hat{\sigma}_i^{\gamma,\eta}|}{\sqrt{\hat{\sigma}_i^{\gamma,\gamma} \hat{\sigma}_i^{\eta,\eta} + (\hat{\sigma}_i^{\gamma,\eta})^2}} \right\} < \epsilon\,.$$

The quantities in the numerators correspond to the difference of the moment parameters between two consecutive iterations. The tolerance value ϵ is predetermined and we fixed it at $\epsilon = 0.01$ which in our experience it provides sufficiently robust results.

Once the convergence criterion is met, what we are interested in, is the last update of the *ssd* matrices which, in fact, we calculate at each iteration in order to perform the M-step.

7.4 Simulation of genetical genomics data with missing values

Missing values are generated under the MCAR mechanism, that is, the occurrence of missing values does not depend on any other observed nor unobserved values.

Given a complete data set $\mathcal{X}$ sampled with the algorithms described in Chapter 4, we remove uniformly at random a fraction of τ observations of each discrete genetic marker $\delta \in \Delta$. Thus, the number of missing values for each genetic marker equals τn. We consider a maximum rate of missing values of $\tau = 0.2$. Genotype data with higher rates are generally discarded (The International HapMap Consortium, 2007).

7.5 Accuracy of likelihood ratio estimates

Here, we analyze the accuracy of the likelihood ratio estimates $\Lambda_{\delta\gamma.Q}$ defined in Equation (5.5) when the EM algorithm and the complete-case analysis are used.

To this end, we consider a genetic map consisting of one chromosome of 100 cM long and 3 equally-spaced genetic markers. Given this genetic map, we simulate a gene network of 50 genes each of them randomly connected to three other genes. We also simulate 3 eQTL hotspots such that each of them is also connected to three genes. The marginal Pearson correlations between the genes is set to 0.5 and the additive effect from eQTL associations are simulated from a normal distribution with mean value of 2.5 and standard deviation of 1. From this eQTL model, we simulate 1,000 data sets $\mathcal{X}$ of $n = 30$ observations each.

For each data set, we select a pair of mixed discrete and continuous r.v.'s (δ, γ) uniformly at random between present and absent edges. We remove 20% of values of δ according to the MCAR mechanism.

For each complete and missing data set, we calculate the likelihood ratio corresponding to the null hypothesis of $\delta \perp\!\!\!\perp \gamma | Q$ where Q is formed by the continuous variables indexed by vertices adjacent to γ in G.

In Figure 7.1, the likelihood ratios estimated from a complete data set are depicted against the ones estimated from a data set with missing values. We observe that likelihood ratios corresponding to present edges calculated with the complete-case analysis are more variable than those obtained with the EM algorithm (one-sided F-test, p-value < 0.05). Therefore, the EM algorithm provides more accurate estimates of the likelihood ratio.

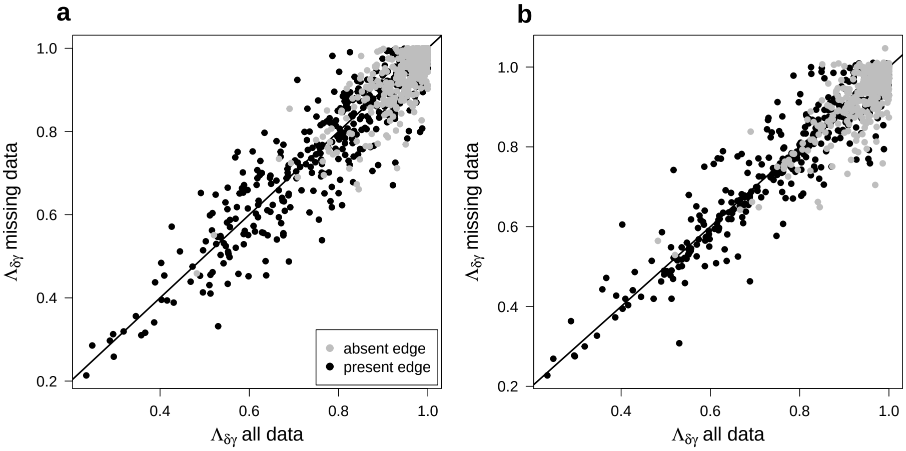

Figure 7.1: Comparison of likelihood ratio estimates from a complete data set and a data set with missing values calculated from (a) complete-case analysis and (b) EM algorithm.

7.6 Accuracy of phenotypic variance estimates

In Equation (5.15) we have seen that the percentage of phenotypic variance of a r.v. X_γ due to an eQTL, η^2, is a function of the likelihood ratio statistic $\Lambda_{\delta\gamma.Q}$ when $Q = \emptyset$. In this section we want to analyze the accuracy of this quantity η^2 when it is estimated from data with missing genotype values. To this end, we consider exactly the same previous simulation setup. However, in this case, for each complete and missing data set, we calculate the likelihood ratio corresponding to the null hypothesis of $X_\delta \perp\!\!\!\perp X_\gamma | X_Q$ where $Q = \emptyset$ and X_δ and X_γ correspond to a discrete and a continuous r.v.'s, respectively, such that α and β form an edge in the underlying graph G.

In Figure 7.2, the η^2 estimates from a complete data set, calculated as 1 minus the likelihood ratio statistic $\Lambda_{\delta\gamma.Q}$, are depicted against the ones estimated from a data set with missing values. We observe that the percentage of phenotypic variance explained by an eQTL calculated with the complete-case analysis are more variable than those obtained with the EM algorithm (one-sided F-test, p-value < 0.05). Thus, the EM algorithm provides more accurate η^2 estimates.

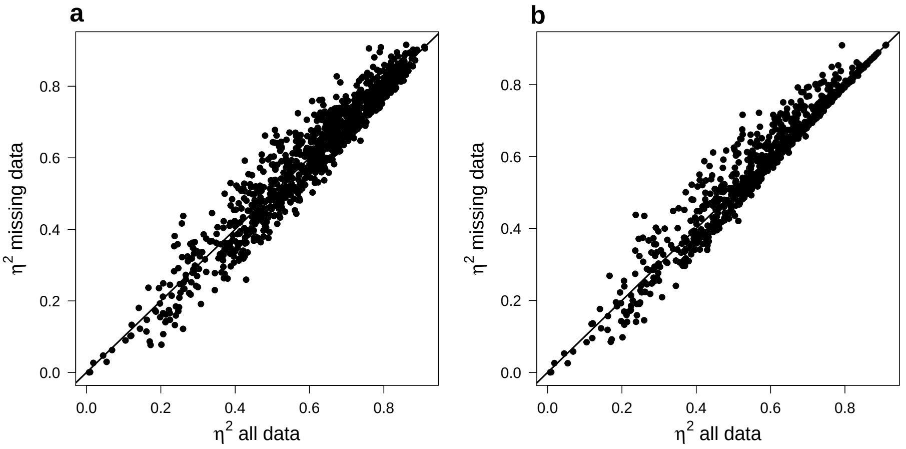

Figure 7.2: Comparison of η^2 estimates (Equation 5.15) from a complete data set and a data set with missing values calculated from (a) complete-case analysis and (b) EM algorithm.

7.7 Analysis of the non-rejection rate

Finally, we investigate the performance of different strategies to learn an eQTL network from data with missing genotype values and where the number of variables is much higher than the number of samples ($p \gg n$). To this end, we estimate the non-rejection rate for each pair of continuous gene expression profile and discrete genotype where the likelihood ratio statistic used in each hypothesis test is calculated using the complete-case analysis and an imputation approach.

Here, we consider a genetic map consisting of 10 chromosomes of 100 cM long each one and 10 genetic markers on each chromosome. Given this genetic map, we simulate a gene network of 200 genes such that each of these genes is randomly connected to 10 other genes. We also simulate 90 *cis*-eQTLs and 10 eQTL hotspots such that each of them is connected to 10 genes. The marginal Pearson correlations between the genes is set to 0.5 and the mixed linear eQTL associations are simulated from a normal distribution with mean value of 1 and standard deviation of 1.

From this eQTL model, we simulate a single data set $\mathcal{X}$ of $n = 50$ observations. We estimate non-rejection rate values for each pair of mixed vertices using conditioning sets of size $q = 11$.

From the previous data set $\mathcal{X}$ we remove a fraction $\tau = 0.2$ of the observations of each eQTL. We use two different methods to deal with missing genotype data: the complete-case analyis and the multiple imputation method of Broman et al. (2003), implemented through the function `sim.geno()` included in the `R/qtl` package. This method uses a hidden Markov model to impute missing genotypes in experimental crosses. We use it here to create 10 imputed data sets from the original one. Finally, the non-rejection rate is calculated in each of the imputed data sets and averaged at each pair of vertices.

For each method, we build a precision-recall curve from the non-rejection rate values from all the mixed pairs (δ, γ) with $\delta \in \Delta$ and $\gamma \in \Gamma$ (Figure 7.3). From these curves, we observe that the multiple imputation approach and the complete-case analysis do not display substantial differences in performance although the curve from the latter is slightly better than the former. However, the multiple imputation method involves the calculation of the non-rejection rate for each imputed data set. Therefore, it results that the complete-case analysis is, in this case, ten times computationally less expensive than the multiple imputation approach.

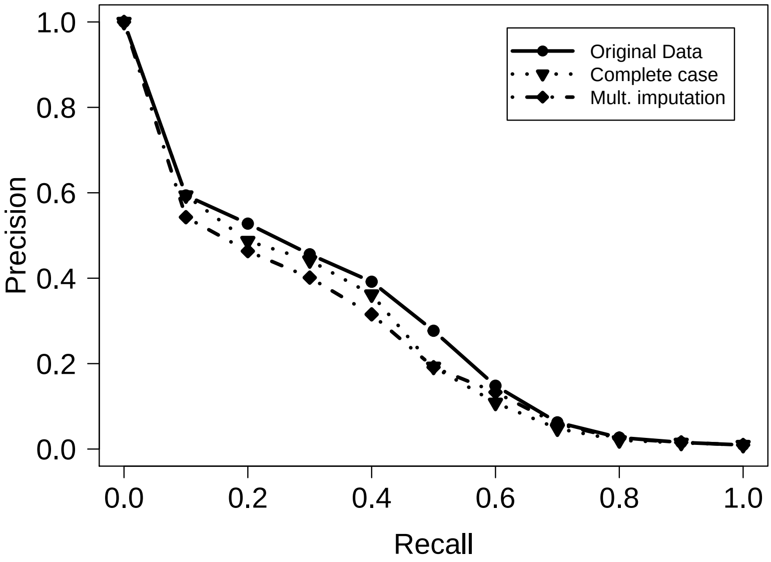

Figure 7.3: Precision-recall curves from non-rejection rate values estimated with the original data set and data with missing values. In the latter case, the complete-case analysis and a multiple-imputation method were used.

7.8 Concluding remarks

In this chapter we have introduced an strategy to learn a mixed GMM from data with $p \gg n$ and missing values by integrating it as part of our procedure to calculate the non-rejection rate. We have also provided an EM algorithm to estimate the likelihood ratio statistic. These are the main conclusions of this study:

- Under the assumption of data being MCAR, we provide an effective and straightforward implementation of the complete-case approach when calculating limited-order correlations.

- We have integrated an EM algorithm to accurately estimate the parameters of a decomposable mixed GMM and the corresponding likelihood ratio between a saturated and a constrained mixed GMMs. This algorithm provides robust estimations under the less stringent assumption of data being MAR.

- Both methods have been implemented in the R/Bioconductor package `qpgraph`.

- By performing an experiment with synthetic data, we show that the EM algorithm yields more accurate estimates of the likelihood ratio and of the percentage of phenotypic variance explained by an eQTL compared to those obtained from the complete-case analysis under the assumption of data being MCAR.

- The use of marginal distributions of size $(q + 2) < n$ to estimate the non-rejection rate favors the application of the complete-case analysis because it allows the use of all the complete observations of the corresponding margin and not of all the variables in the data set. Thus, the performance of the complete-case analysis and the multiple imputation approach is similar when using the non-rejection rate to learn an eQTL network. However, the multiple imputation approach requires running analyses algorithms multiple times in a high-dimensional setting, thereby increasing the computational cost of the analysis.

Conclusions and discussion

Conclusions

Nearly a century ago, Fisher established the theoretical framework to investigate the genetic heritability of complex traits (Fisher, 1918). Since then, quantitative genetics has emerged as a widely productive research area which tries to identify the genomic regions that contribute to the variability of quantitative traits.

With the arrival of the genomics era, an increasing amount of genome-scale biological data has become available. These data enable the exploration of the genetic basis of complex traits at an unprecedented resolution. However, this high-dimensionality challenges classical analysis methods that study one trait at a time, specially with multivariate cellular traits such as gene expression.

In this context, the work in this thesis has tackled the problem of eQTL mapping and, concretely, we have provided a multivariate approach to map eQTL networks based on mixed graphical Markov models. To this end, five main contributions which are discussed in detail in the previous chapters have been presented. Here, we summarize the main results:

We have developed and implemented algorithms to simulate eQTL network models and data from them by simulating mixed GMMs.

We have provided a procedure to simulate eQTL network models and genetical genomics data from these eQTL models. Concretely, we have developed algorithms to simulate the structure and the parameters of a mixed graphical Markov model representing the underlying eQTL and gene-gene associations with given marginal Pearson correlations between genes and additive effects of the discrete variables on the continuous ones. These algorithms, in conjunction with the functionalities of the `R/qtl` package, allow one to build an eQTL network model from experimental crosses such as a backcross.

The simulation of these models enables us to investigate the underpinnings of eQTL networks. In particular, we have examined the propagation of the additive effects of the eQTLs through the gene network. These simulations allow us to understand how indirect eQTL associations may arise as direct ones when inferring eQTL networks by testing the marginal independence between a genetic variant and a gene. This suggests the necessity to adjust these associations for other genes and confounding factors. We have also examined how this situation may be more problematic under the effect of selection bias and we illustrated how the estimated additive effect of a detected eQTL may vary from its true effect, specially if the true effect is small. Moreover, these spurious eQTL associations may lead to an overestimation of the percentage of variability of gene expression profiles explained by the identified eQTLs. Finally, the simulation of genetical genomics data from eQTL models helps us to assess and compare the performance of learning eQTL network methods as we have seen in Chapter 5.

We have provided an exact conditional independence test for mixed discrete and continuous data to identify higher-order eQTL associations.

In order to find non-spurious eQTL and gene-gene associations from genetical genomics data, we have derived a likelihood ratio test in which we test the association between a genetic variant and a gene expression profile or between two gene expression profiles while controling for a set of variables including other genes, genetic variants or phenotypic variables in addition to confounding factors. Under the null hypothesis of conditional independence, this test follows exactly a beta distribution with certain parameters.

By simulating eQTL network models with our algorithms, we have demonstrated that this exact test is specially suitable for testing conditional independencies from data with small sample sizes and with more complex gene networks compared to the traditional likelihood ratio test that follows a χ^2 distribution asymptotically. Moreover, we have used a real data set from yeast to verify that the exact test, unlike the asymptotic one, provides a correct uniform distribution of p-values under the null hypothesis of conditional independence. Furthermore, we have demonstrated that higher-order conditioning implicitly adjusts for confounding effects that simultaneously affect all gene expression profiles.

We have provided a multivariate approach to learn eQTL networks by estimating mixed GMMs.

We have introduced a multivariate statistical procedure to learn mixed GMMs representing an eQTL network from genetical genomics data in which the number of variables (p) is much greater than the number of samples (n), that is, $p \gg n$.

For this purpose, we have adapted the limited-order correlation approach of Castelo and Roverato (2006) for Gaussian GMMs to mixed GMMs. This strategy consists in testing conditional independencies between each genetic variant and each gene expression profile as well as between each pair of gene expression profiles by conditioning on other $q < (n-2)$ variables. By applying this strategy, we calculate a measure of linear association between each pair of variables called non-rejection rate from which we obtain an estimation of the underlying eQTL network, called qp-graph. The estimated qp-graphs are non-decomposable mixed GMMs whose edges represent q-order conditional dependencies.

We have demonstrated that limited-order correlations constitute an appealing framework to develop approaches that exploit the sparseness of the underlying eQTL network when trying to learn the structure of a mixed GMM from a genetical genomics data set in which the number of variables is much greater than the number of samples. The methodology presented in this thesis is implemented as part of the R/Bioconductor package called `qpgraph` available from http://www.bioconductor.org.

We have demonstrated that the use of higher-order conditioning is suitable for identifying non-spurious eQTL associations.

We have estimated a qp-graph representing the underlying eQTL network of a real genetical genomics data set from yeast (Brem and Kruglyak, 2005) and compared it to the eQTL network estimated from the traditional univariate approach called single marker regression. On the one hand, we have revealed an enrichment of *cis*-eQTLs identified by our method compared to those identified by the single marker regression approach.

On the other hand, we have found that, in general, the set of *trans*-eQTLs identified by the non-rejection rate is smaller than the set of *trans*-eQTLs identified by the single marker regression approach. However, our set of *trans*-eQTLs explains at least the same fraction of phenotypic variability as the set of eQTLs identified by the univariate approach. Furthermore, we have investigated the functional role of these *trans*-eQTLs and we verified that the non-rejection rate finds *trans*-eQTLs such that their overlapping and target genes have more coherent biological functions than those identified by the single marker regression.

These results suggest that our method is able to identify more direct eQTL associations and discard those eQTL associations that are indirect due to gene-gene correlations or because they are confounded by a factor simultaneously affecting all the gene network.

The estimated eQTL network by means of a qp-graph provides relevant information about the genetic control of gene expression in yeast.

We have provided a strategy to learn eQTL networks from genetical genomics data with missing genotype values.

The use of marginal distributions of size $(q + 2) < n$ to estimate the non-rejection rate have enabled an efficient application of the complete-case analysis when learning eQTL networks from data with $p \gg n$ and missing genotype values. We have compared the performance of this strategy with a multiple-imputation approach and both methods perform similarly when using the non-rejection rate to learn the structure of a mixed GMM. However, the multiple imputation approach is computationally more intense since it requires the application of learning algorithms multiple times in a high-dimensional setting.

We have also provided an expectation-maximization (EM) algorithm that yields more accurate estimates of the likelihood ratio for the presence of a mixed interaction and also provides more accurate estimations of the percentage of phenotypic variance explained by an eQTL.

Discussion

After a century of research, the field of quantitative genetics still relies on the assumption that genetic inheritance is mainly additive. Geneticists are aware that this model does not reflect the true nature of biological systems and that other than additive effects of genetic variants should contribute to the total variability of complex traits. Some of these non-additive genetic effects could be investigated with the approach presented in this thesis in the following way.

argued that complex traits could be affected by the interaction of two or more loci, that is, may be affected by loci showing epistasis. Mixed GMMs naturally accomodate interaction effects between discrete variables and the learning approach provided in this thesis could be used to detect epistatic effects affecting gene expression profiles. Furthermore, a natural extension of the work done in Chapter 4 could be the development of procedures that enable the simulation of other experimental crosses such as an intercross (F2) or recombinant inbred lines (RILs). If so, dominance or epistatic effects could be simulated and more general eQTL network models could be explored.

The contribution of genetic variance control, that is, the assumption that genetic factors not only affect the mean of complex traits but also their variance, could be approached in the context of mixed GMMs. Basically, we shoul derive the appropriate parameters of the null distribution of the exact test of conditional independence for heterogeneous mixed GMMs, in which the covariance matrix $\Sigma(i)$ of gene expression profiles does depend on the genotypes.

Finally, it has been suggested that many genetic variants with small additive effects also contribute to the total variability of complex traits. In particular, it is argued that the additive effects of genetic variants affecting the variability of complex traits follow approximately an exponential distribution. Concretely, few genetic variants with a strong additive effect on the phenotypes are usually detected whereas a large number of them have a very small contribution to the variability of the trait and most of these remain to be discovered. It is argued that data sets with larger sample sizes would facilitate the discovery of these genetic variants. However, we have seen that small effects also arise as a result of the propagation of additive effects through gene-gene correlations, the effect of confounding factors and selection bias, hampering the identification of genetic variants directly associated to the complex trait. Therefore, one of the challenges for researchers when trying to identify genetic variants with small effects is the design of methods that account for these obstacles. In fact, when we adjust the association between a marker and a gene for the rest of genes, we have shown in a data set from yeast that the number of genetic variants that have a direct small effect on the phenotype is low.

On the other hand, two further extensions of our work could be addressed in the near future. One is that in Chapter 7 we have implemented an EM algorithm to estimate the parameters of decomposable mixed GMM from data with missing values. Although the likelihood ratio statistics calculated from these estimations are very accurate, we could not derive their distribution under the null hypothesis of conditional independence and the exact test cannot be performed in this case. Therefore, it would be desirable to derive the exact distribution of these likelihood ratio statistics when they are computed with the parameters obtained from the EM algorithm so that they could be used to test conditional independencies.

The other is that one of the assumptions of our learning approach is that gene expression values follow a multivariate Gaussian distribution given the values of the discrete genotypes. While this assumption works reasonably well for gene expression data measured with microarray technology, it prevents the direct application of our learning procedure to data assayed from NGS technologies. Therefore, it could be interesting to explore the way in which this graphical Markov model approach could be extended and applied to sequence-based expression measurements.

In summary, quantitative genetics is still an evolving research field. In the next decades, the advances in genomics and the development of statistical methods addressing challenges as the ones exposed above, among others, will bring us to a deeper understanding of the biological function and the connection between genotypes and phenotypes.

Bibliography

Ackermann, M., Clment-Ziza, M., Michaelson, J. J., and Beyer, A. (2012). Teamwork: Improved eQTL Mapping Using Combinations of Machine Learning Methods. *PLoS ONE*, 7(7):e40916.

Albert, F. W., Treusch, S., Shockley, A. H., Bloom, J. S., and Kruglyak, L. (2014). Genetics of single-cell protein abundance variation in large yeast populations. *Nature*.

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322(5903):881–888.

Arends, D., Prins, P., Jansen, R. C., and Broman, K. W. (2010). R/qtl: high-throughput multiple QTL mapping. *Bioinformatics*, 26(23):2990–2992.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B., and Boone, C. (2013). Genetic interaction networks: Toward an understanding of heritability. *Annual review of genomics and human genetics*, 14:111–133.

Basten, C.J., B. W. and Zeng, Z.-B. (2002). *QTL Cartographer, Version 1.16.* Department of Statistics, North Carolina State University, Raleigh, NC.

Bateson, W., Saunders, E. R., and Punnett, R. (1905). Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society*, 2:1–55, 80–99.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., Urban, A. E., Montgomery, S. B., Levinson, D. F., and Koller, D. (2013). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*.

Bennett, S. T., Barnes, C., Cox, A., Davies, L., and Brown, C. (2005). Toward the $1000 human genome. *Pharmacogenomics*, 6:373–382.

Bing, N. and Hoeschele, I. (2005). Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics*, 170(2):533–42.

Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, 494:234–237.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32:314–331.

Breitling, R., Li, Y., Tesson, B. M., Fu, J., Wu, C., Wiltshire, T., Gerrits, A., Bystrykh, L. V., de Haan, G., Su, A. I., et al. (2008). Genetical genomics: spotlight on QTL hotspots. *PLoS Genetics*, 4(10):e1000232.

Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102:1572–7.

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755.

Broman, K. W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal*, 30(7):44–52.

Broman, K. W. and Sen, S. (2009). *A guide to QTL mapping with R/qtl.* Springer.

Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19:889–890.

Castelo, R. and Kočka, T. (2003). On inclusion-driven learning of Bayesian networks. *Journal of Machine Learning Research*, 4:527–74.

Castelo, R. and Roverato, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *Journal of Machine Learning Research*, 7:2621–50.

Castelo, R. and Roverato, A. (2009). Reverse engineering molecular regulatory networks from microarray data with qp-graphs. *Journal of Computational Biology*, 16(2):213–227.

Chen, L. S., Emmert-Streib, F., and Storey, J. D. (2007). Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol*, 8(10):R219.

Cheung, V. G. and Spielman, R. S. (2009). Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature Review Genetics*, 10:595–604.

Chickering, D. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–54.

Chin, C.-S., Chubukov, V., Jolly, E. R., DeRisi, J., and Li, H. (2008). Dynamics and design principles of a basic regulatory architecture controlling metabolic pathways. *PLoS Biol*, 6(6):e146.

Chun, H. and Keleş, S. (2009). Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, 182(1):79–90.

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138:963–971.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Review Genetics*, 10:184 – 194.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. *science*, 327(5964):425–431.

Cox, D. and Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Crick, F. (1958). On protein synthesis. *The Symposia of the Society for Experimental Biology*, 12:138–163.

Curtis, R. E., Kim, S., Woolford Jr, J. L., Xu, W., and Xing, E. P. (2013). Structured association analysis leads to insight into Saccharomyces cerevisiae gene regulation by finding multiple contributing eQTL hotspots associated with functional gene modules. *BMC Genomics*, 14(196):1–17.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B*, 39(1):1–38.

Didelez, V. and Edwards, D. (2004). Collapsibility of graphical CG-regression models. *Scandinavian Journal of Statistics*, 31(4):535–551.

Didelez, V. and Pigeot, I. (1998). Maximum likelihood estimation in graphical models with missing values. *Biometrika*, 85(4):960–966.

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Review Genetics*, 3:43–52.

Edwards, D. (2000). *Introduction to graphical modelling.* Springer.

Edwards, D., de Abreu, G. C. G., and Labouriau, R. (2010). Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC Bioinformatics*, 11:18.

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F. O., and Knight, J. C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nature genetics*, 44(5):502–510.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics (4th Edition).* Pearson Prentice Hall.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433.

Flassig, R. J., Heise, S., Sundmacher, K., and Klamt, S. (2013). An effective framework for reconstructing gene regulatory networks from genetical genomics data. *Bioinformatics*, 29(2):246–254.

Fodor, S., Read, J., Pirrung, M., Stryer, L., Lu, A., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995):767–773.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–41.

Fu, J., Wolfs, M. G. M., Deelen, P., Westra, H.-J., Fehrmann, R. S. N., te Meerman, G. J., Buurman, W. A., Rensen, S. S. M., Groen, H. J. M., Weersma, R. K., van den Berg, L. H., Veldink, J., Ophoff, R. A., Snieder, H., van Heel, D., Jansen, R. C., Hofker, M. H., Wijmenga, C., and Franke, L. (2012). Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet*, 8(1):e1002431.

Galton, F. (1889). *Natural Inheritance*, volume v. 42; v. 590 of *Natural Inheritance*. MacMillan.

Geng, Z., Wan, K., and Tao, F. (2000). Mixed graphical models with missing data and the partial imputation EM algorithm. *Scand J Stat*, 27(3):433–444.

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer.

Grone, R., Johnson, C., Sá, E., and Wolkowicz, H. (1984). Positive definite completions of partial Hermitian matrices. *Linear Algebra Applic.*, 58:109–124.

Grosveld, F., van Assendelft, G. B., Greaves, D. R., and Kollias, G. (1987). Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell*, 51(6):975–985.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E., and Martin, J. B. (1983). A polymorphic DNA marker genetically linked to huntington's disease. *Nature*, 306:234–238.

Haldane, J. and Bell, J. (1937). The linkage between the genes for colour-blindness and haemophilia in man. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 123(831):119–50.

Haley, C. S. and Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69:315–324.

Hansen, K. D., Wu, Z., Irizarry, R. A., and Leek, J. T. (2011). Sequencing technology does not eliminate biological variability. *Nature Biotechnology*, 29(7):572–573.

Harary, F. (1969). *Graph theory.* Addison-Wesley.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition.* Springer.

Huentelman, M. J., Craig, D. W., Shieh, A. D., Corneveaux, J. J., Hu-Lince, D., and Pearson, John V andStephan, D. A. (2005). SNiPer: Improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. *BMC Genomics*, 6(149).

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Irizarry, R. A., Hobbs, B., Collin, F., BeazerBarclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Jackson, D. A., Symons, R. H., and Berg, P. (1972). Biochemical method for inserting new genetic information into DNA of simian virus 40: Circular SV40 DNA molecules containing lambda phage genes and the galactose operon of escherichia coli. *Proceedings of the National Academy of Sciences*, 69(10):2904–2909.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–0356.

Jansen, R. C. and Nap, J.-P. (2001). Genetical genomics: the added value from segregation. *TRENDS in Genetics*, 17(7):388–390.

Jansen, R. C. and Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, 136(4):1447–1455.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 8:613–636.

Kang, E. Y., Ye, C., Shpitser, I., and Eskin, E. (2010). Detecting the presence and absence of causal relationships between expression of yeast genes with very few samples. *J Comput Biol*, 17(3):533–46.

Kang, H. M., Ye, C., and Eskin, E. (2008a). Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–25.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008b). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Kao, C.-H., Zeng, Z.-B., and Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics*, 152(3):1203–1216.

Kendziorski, C. and Wang, P. (2006). A review of statistical methods for expression quantitative trait loci mapping. *Mammalian Genome*, 17(6):509–517.

Kerem, B., Rommens, J., Buchanan, J., Markiewicz, D., Cox, T., Chakravarti, A., Buchwald, M., and Tsui, L. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922):1073–1080.

Kim, S. and Xing, E. P. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet*, 5(8):e1000587.

Kruglyak, L. (2008). The road to genome-wide association studies. *Nature Reviews Genetics*, 9:314–318.

Krzywinski, M., Birol, I., Jones, S. J., and Marra, M. A. (2012). Hive plots−rational approach to visualizing networks. *Briefings in Bioinformatics*, 13(5):627–644.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, 11:241–247.

Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199.

Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, 265:2037–2048.

Landsteiner, K. (1901). Uber agglutinationserscheinungen normalen menschlichen blutes. *Wienerklinische Wochenschrift*, 14:1132–1134.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):1–10.

Lauritzen, S. (1996). *Graphical Models.* Oxford University Press.

Lauritzen, S. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17(1):31–57.

Leduc, M. S., Blair, R. H., Verdugo, R. A., Tsaih, S.-W., Walsh, K., Churchill, G. A., and Paigen, B. (2012). Using bioinformatics and systems genetics to dissect HDL-cholesterol genetics in an MRL/MpJ SM/J intercross. *Journal of Lipid Research*, 53(6):1163–1175.

Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *American journal of human genetics*, 88:294–305.

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

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161.

Lehman, E. and Romano, J. (2005). *Testing statistical hypotheses*. Springer-Verlag.

Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics*, 14:168–178.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(323):1:16.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, J. and Burmeister, M. (2005). Genetical genomics: combining genetics with gene expression analysis. *Human Molecular Genetics*, 14(suppl 2):R163–R169.

Li, Q., Peterson, K. R., Fang, X., and Stamatoyannopoulos, G. (2002). Locus control regions. *Blood*, 100(9):3077–3086.

Listgarten, J., Kadie, C., and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38):16465–16470.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J. T., van Oudenaarden, A., Barton, D. B. H., Bailes, E., Nguyen, A. N., Jones, M., Quail, M. A., Goodhead, I., Sims, S., Smith, F., Blomberg, A., Durbin, R., and Louis, E. J. (2009). Population genomics of domestic and wild yeasts. *Nature*, 458:337–341.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. Probability and Statistics. Wiley, second edition.

Liu, B., de la Fuente, A., and Hoeschele, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, 178(3):1763–76.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680.

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S. A., James, M., Groot, N., MacFarlane, H., Jenkins, B., Anderson, M. A., Wexler, N. S., Gusella, J. F., Bates, G. P., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A.-M., Lehrach, H., Buckler, A. J., Church, D., Doucette-Stamm, L., O'Donovan, M. C., Riba-Ramirez, L., Shah, M., Stanton, V. P., Strobel, S. A., Draths, K. M., Wales, J. L., Dervan, P., Housman, D. E., Altherr, M., Shiang, R., Thompson, L., Fielder, T., Wasmuth, J. J., Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., Collins, F. S., Snell, R., Holloway, T., Gillespie, K., Datson, N., Shaw, D., and Harper, P. S. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971 – 983.

Mackay, T. F. (2001). The genetic architecture of quantitative traits. *Annual Review of Genetics*, 35:303–339.

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Review Genetics*, 10:565–577.

Manichaikul, A., Moon, J. Y., Sen, ., Yandell, B. S., and Broman, K. W. (2009). A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics*, 181(3):1077–1086.

Manichaikul, A., Palmer, A. A., Sen, S., and Broman, K. W. (2007). Significance thresholds for quantitative trait locus mapping under selective genotyping. *Genetics*, 177(3):1963–1966.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747–753.

Marco-Sola, S., Sammeth, M., Guigo, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, 9:1185–1188.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380.

Martínez, O. and Curnow, R. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics*, 85(4):480–488.

McCall, M. N., Bolstad, B. M., and Irizarry, R. A. (2010). Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2):242–253.

Michaelson, J. J., Alberts, R., Schughart, K., and Beyer, A. (2010). Data-driven assessment of eQTL mapping methods. *BMC genomics*, 11(1):502.

Morgan, T. H. (1911). Random segregation versus coupling in mendelian inheritance. *Science*, 34(873):384.

Morgan, T. H. (1916). *A critique of the theory of evolution*. Princeton University Press.

Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J., and Marra, M. A. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short−read sequencing. *BioTechniques*, 45(1):81–94.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5:621–628.

Nackley, A. G., Shabalina, S. A., Tchivileva, I. E., Satterfield, K., Korchynskyi, O., Makarov, S. S., Maixner, W., and Diatchenko, L. (2006). Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science*, 314(5807):1930–1933.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349.

Neto, E. C., Ferrara, C. T., Attie, A. D., and Yandell, B. S. (2008). Inferring causal phenotype networks from segregating populations. *Genetics*, 179(2):1089–100.

Neto, E. C., Keller, M. P., Attie, A. D., and Yandell, B. S. (2010). Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann Appl Stat*, 4(1):320–339.

Nielsen, P., van den Hazel, B., Didion, T., de Boer, M., Jrgensen, M., Planta, R., Kielland-Brandt, M., and Andersen, H. (2001). Transcriptional regulation of the Saccharomyces cerevisiae amino acid permease gene BAP2. *Molecular & general genetics*, 264(5):613–622.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Review Genetics*, 12(6):443–451.

Parts, L., Stegle, O., Winn, J., and Durbin, R. (2011). Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genet*, 7(1):e1001276.

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A.-L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406:747–752.

Petes, T. D. and Botstein, D. (1977). Simple mendelian inheritance of the reiterated ribosomal DNA of yeast. *Proceedings of the National Academy of Sciences*, 74(11):5091–5095.

Petretto, E., Mangion, J., Dickens, N. J., Cook, S. A., Kumaran, M. K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M., Hubner, N., and Aitman, T. J. (2006). Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genetics*, 2(10):e172.

Plomin, R., Haworth, C. M. A., and Davis, O. S. P. (2009). Common disorders are quantitative traits. *Nature Review Genetics*, 10:872–878.

Powell, J. E., Visscher, P. M., and Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, 11:800–805.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics*, 69(1):124 – 137.

Rao, C. (1973). *Linear Statistical Inference and Its Applications*. John Wiley & Sons.

Roberts, A., McMillan, L., Wang, W., Parker, J., Rusyn, I., and Threadgill, D. (2007). Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics*, 23(13):i401–i407.

Rockman, M. V. (2008). Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature*, 456:738–744.

Rockman, M. V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nature Review Genetics*, 7:862–872.

Roff, D. A. (2007). A centennial celebration for quantitative genetics. *Evolution*, 61:1017–1032.

Ronald, J., Brem, R. B., Whittle, J., and Kruglyak, L. (2005). Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet*, 1(2):e25.

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411.

Roverato, A. and Castelo, R. (2012). Learning undirected graphical models from multiple datasets with the generalized non-rejection rate. *International Journal of Approximate Reasoning*, 53(9):1326 – 1335.

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., Bhardwaj, N., Rubin, M., Snyder, M., and Gerstein, M. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology*, 7(1).

Sambrook, P. N., MacGregor, A. J., and Spector, T. D. (1999). Genetic influences on cervical and lumbar disc degeneration: A magnetic resonance imaging study in twins. *Arthritis & Rheumatism*, 42(2):366–372.

Sanger, F., Nicklen, S., and Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of The National Academy of Sciences of The United States Of America*, 74(12):5463–5467.

Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus Vulgaris*. *Genetics*, 8(6):552–560.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., Lum, P. Y., Leonardson, A., Thieringer, R., Metzger, J. M., Yang, L., Castle, J., Zhu, H., Kash, S. F., Drake, T. A., Sachs, A., and Lusis, A. J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37:710–717.

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Seber, G. (2007). *A matrix handbook for statisticians*. Wiley-Interscience.

Segal, E., Pe'er, D., Regev, A., Koller, D., and Friedman, N. (2006). Learning module networks. *J Mach Learn Res*, 6(1):557–88.

Sen, S. and Churchill, G. (2001). A statistical framework for quantitative trait mapping. *Genetics*, 159(1):371–387.

Sillanpää, M. J. and Coriander, J. (2002). Model choice in gene mapping: what and why. *Trends in Genetics*, 18:301–307.

Smyth, G. (2005). limma: Linear models for microarray data. In Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S., editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, pages 397–420. Springer New York.

Soller, M., Brody, T., and Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics*, 47(1):35–39.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Review Genetics*, 14:483–495.

Speed, T. (2005). Genetic map functions. In *Encyclopedia of Biostatistics 2nd Edition*. John Wiley & Sons.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology*, 6(5):e1000770.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14:43–59.

Tenesa, A. and Haley, C. S. (2013). The heritability of human disease: estimation, uses and abuses. *Nature Review Genetics*, 14:139–149.

Tesson, B. M. and Jansen, R. C. (2009). eqtl analysis in mice and rats. In DiPetrillo, K., editor, *Cardiovascular Genomics*, volume 573 of *Methods in Molecular Biology*, pages 285–309. Humana Press.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661 – 678.

Venter, J. C. et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era - concepts and misconceptions. *Nature Review Genetics*, 9:255 – 266.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., and Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet*, 2(3):e41.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Review Genetics*, 10(1):57–63.

Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid.

Wessel, J., Zapala, M. A., and Schork, N. J. (2007). Accommodating pathway information in expression quantitative trait locus analysis. *Genomics*, 90(1):132 – 142.

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., Zhernakova, A., Zhernakova, D. V., Veldink, J. H., Van den Berg, L. H., Karjalainen, J., Withoff, S., Uitterlinden, A. G., Hofman, A., Rivadeneira, F., 't Hoen, P. A. C., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A. B., Hernandez, D. G., Nalls, M. A., Homuth, G., Nauck, M., Radke, D., Volker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S. A., Enquobahrie, D. A., Lumley, T., Montgomery, G. W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R. C., Visscher, P. M., Knight, J. C., Psaty, B. M., Ripatti, S., Teumer, A., Frayling, T. M., Metspalu, A., van Meurs, J. B. J., and Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45:1238–1243.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Williams, R. B., Chan, E. K., Cowley, M. J., and Little, P. F. (2007). The influence of genetic variation on gene expression. *Genome Research*, 17(12):1707–1716.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceeding of the National Academy of Science of the United States of America*, 6(6):320–332.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215.

Wu, L., Candille, S. I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. *Nature*, 499:79–82.

Yandell, B. S., Mehta, T., Banerjee, S., Shriner, D., Venkataraman, R., Moon, J. Y., Neely, W. W., Wu, H., von Smith, R., and Yi, N. (2007). R/qtlbim: QTL with bayesian interval mapping in experimental crosses. *Bioinformatics*, 23(5):641–643.

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., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42:565 – 569.

Yi, N. and Shriner, D. (2007). Advances in bayesian multiple quantitative trait loci mapping in experimental crosses. *Heredity*, 100:240–252.

Yin, J. and Li, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *Annals of Applied Statistics*, 5(4):2630–2650.

Yvert, G., Brem, R. B., Whittle, J., Akey, J. M., Foss, E., Smith, E. N., Mackelprang, R., and Kruglyak, L. (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics*, 35:57–64.

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A. L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet*, 9(5):e1003520.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics*, 136(4):1457–68.

Zhang, W., Zhu, J., Schadt, E. E., and Liu, J. S. (2010). A bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Comput Biol*, 6(1):e1000642.

Zhu, J., Lum, P. Y., Lamb, J., GuhaThakurta, D., Edwards, S. W., Thieringer, R., Berger, J. P., Wu, M. S., Thompson, J., Sachs, A. B., and Schadt, E. E. (2004). An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res*, 105(2-4):363–74.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198.