Doctoral Thesis

Identification of Interactions in expression Quantitative Trait Data via the Minimum Description Length

（ 最小記載長を用いた定量的発現形質データからの相互作用の同定 ）

ジョージ　チャルキディス

Acknowledgements

I would really like to take this opportunity to heart-fully thank Professor Sugano for his guidance through the PhD-course, his advice on research, and the many interesting discussions about genetics. The ideas of investigating gene expression data and its functional impacts are due to Professor Sugano and his expertise in RNA-sequencing.

Apart from gaining a deeper understanding on genomics thanks to Professor Sugano, the most important and probably valuable lesson Professor Sugano taught me is about proper academic writing, good etiquette in explicitly highlighting and giving credit to the authors of contributing ideas, and expressing one's owns ideas in an original, precise manner. For this I am extremely grateful to Professor Sugano.

Professor Hagenauer revealed Claude Shannon's involvement in genetics to me by pointing out that the creator of information theory actually wrote his PhD-thesis an algebra that can be used to describe phenomena in genetics. Thus, Professor Hagenauer underlined the potential of information theoretic analysis approaches in genomics and opened this challenging interdisciplinary field of study to me. I am thankful to have had the opportunity to receive his guidance and work on applications of information theory in genetics.

I would also like to express my gratitude to Professor Miyano and Professor Nagasaki who supported my undertaking in Japan from the beginning and gave me valuable advice about the culture. It was exciting to learn how to operate a supercomputer and harness its power for analyzing and interpreting genomic data.

Professor Nagasaki always had an open door and never hesitated to give personal as well as professional assistance and advice. I was amazed by his ability to quickly identify future research trends. It was an amazing experience working together with Professor Nagasaki.

Regarding the utility of joint genotype and transcriptome analysis in personalized medicine, I would like to thank Professor Bartlett and Professor Ray for their insightful and fruitful discussions. Furthermore, I am glad they provided me with their synthetic eQTL dataset and helped me acquire a broader understanding of genetic diagnostics and personalized medicine. Collaborating with Professor

Bartlett and Professor Ray in an international research effort on the analysis of eQTL data was a very rewarding experience.

I am also very happy and thankful for the comments, suggestions, and improvement advice the thesis-committee members; Professor Kobayashi, Professor Shibuya, Professor Nakai, and Professor Morishita, gave me. Their advice really helped me to improve a lot and create a good thesis. To express my sincerest gratitude towards them, I will briefly mention how their valuable advice had a positive impact on me.

I would like to sincerely thank Professor Kobayashi who advised me about the importance of precise definitions and explanations when talking about biological phenomena. This really improved my writing skills when presenting research findings in genomics.

Many thanks to Professor Shibuya and Professor Morishita for checking the technical details and their insightful discussions about the algorithm.

Professor Nakai pointed out that extensive benchmarking is necessary before accepting a novel method, and I am very grateful for his assistance in conducting such a thorough benchmarking test.

Regarding quality control in sequencing data and validation of computational results, I would like to thank Professor Suzuki for sharing his expertise with me.

Last but not least, I would like to thank all lab members for the wonderful time; the interesting discussions, lab parties, and life in Japan in general. Especially the students and lab members of Sugano-Lab helped me in navigating through the PhD-course.

George Chalkidis

Contents

1 Introduction **1**

2 Paving the way towards better understanding of disease: From GWAS to eQTL studies **5**

2.1 Introducing the road from gene mapping to integrated genomic interaction studies 5

2.1.1 "Data deluge" in next-generation sequencing 7

2.1.2 Insights into Mendelian disease with next-generation sequencing 9

2.1.3 Achievements of GWAS: Uncovering the genetic map of complex disease 10

2.1.4 The road ahead: eQTL and beyond 13

3 Applications of information theory in eQTL analysis **18**

3.1 Information theoretic approaches to the gene mapping problem . . 18

3.1.1 Mathematical representation of eQTL data 18

3.1.2 Claude Shannon's information theory and its relation to genetics 21

3.2 Related work 23

3.2.1 Mutual information relevance networks 23

3.2.2 Gene mapping with Shannon's mutual information 24

3.2.3 ARACNE: Reconstruction of gene regulatory networks . . 26

3.2.4 Maximal information coefficient 27

4 Introducing the MDL-principle to eQTL analysis **29**

4.1 Analyzing genomic data with the MDL-principle 29

4.1.1 MDL eQTL analysis 30

4.1.2 Basics of MDL 32

4.1.3 Kontkanen and Myllymäki (KM) calculation method for NML codes . . . 34
4.1.4 The KM-dynamic programming algorithm for optimizing the stochastic complexity of a grid Ξ . . . 38
4.2 Extending the KM-method for eQTL applicability . . . 42
4.2.1 The mKM-algorithm for associating quantitative traits with discrete genotypes . . . 43
4.2.2 MDL-score for assessing association strengths between genotype and transcriptome in eQTL data via the mKM-method 47

5 Making of qMAP - The MDL-analysis software for eQTL 49
5.1 Implementing qMAP in python . . . 49
5.1.1 PLINK input file format . . . 49
5.1.2 Flowchart of the Python program . . . 50
5.1.3 Data initialization module . . . 50
5.1.4 $1D$ grid optimization module for obtaining NML-codelengths of gene expression quantitative traits . . . 53
5.1.5 $2D$ grid optimization module for obtaining the NML-code of the joint description of genotype and transcriptome . . . 57
5.1.6 MDL association score module . . . 59

6 Performance evaluation results 62
6.1 Evaluating qMAP using a synthetic simulated eQTL dataset . . . 62
6.1.1 Simulating eQTL data . . . 63
6.1.2 Evaluation approach . . . 69
6.1.3 Evaluated analysis tools . . . 72
6.1.4 Measurement of detection rates for correct SNP-gene transcript associations . . . 75
6.1.5 Measurements of assigned ranks for correct SNP-gene transcript associations . . . 83
6.1.6 Recovered interaction networks . . . 91
6.1.7 Analysis of a real human cortical eQTL dataset . . . 93

7 Conclusion and Outlook 96

A Figures & Tables 99

1 Introduction

The DNA sequence is a fascinating and inherently complicated information carrier with more functional features than originally anticipated.

We already know that genomic variants in a person's DNA sequence, with single nucleotide polymorphisms (SNPs) being the most intensely studied ones, regulate the emergence of phenotypic traits.

When speaking of phenotypic traits we can mean readily observable traits like eye color, measurable traits like the blood cholesterol level, or even experimentally determinable traits like the expression activity of a gene.

Due to the Human Genome Project and the accompanying sequencing technologies that were specifically developed for that enormous undertaking, the first layer of genomic information that has become accessible is the DNA sequence itself and the variations it contains.

From person to person the DNA sequence varies slightly and it is this observation that researchers believe to be responsible for the biological diversity we see in this world. The manifold of variations in the DNA sequence can encode a huge variety of phenotypes.

As a consequence, a great deal of research has gone into discovering the various links between genome variants and phenotypes. Genome wide association studies (GWAS) created an extensive map which depicts the relationships between SNPs and phenotypic traits. Among the most heavily investigated phenotypic traits are disease related GWAS.

Yet, despite the great efforts of GWAS, there remains a gap in our understanding of how the trait associated SNPs actually form the phenotype or influence disease pathogenesis. Unless the disease related SNP falls into a coding region of the genome and as a non-synonymous mutation alters the protein coding sequence of that gene resulting in a defective protein product which could possibly have existential impacts on an organism, the effect of a disease or phenotype related SNP is not so clear. Much less can be said about the molecular mechanisms that particular SNP is involved in.

Unfortunately, most of the SNPs discovered by GWAS fall into non-coding regions of the genome. Thanks to advances in sequencing technology, a second layer of genetic information has become accessible, namely the transcriptomic landscape.

Recent scientific studies have switched their focus of attention from genome variation studies to the analysis of the transcriptome, which captures the gene expressions of a person. It soon became clear that the space of potential phenotypic traits for which the transcriptomic landscape can code for is not only larger than that of the genotype, but also the connection between a phenotype and its transcriptomic landscape is stronger. According to several recent studies, the transcriptomic landscape is much more informative regarding a phenotype than the genotype data alone.

As a consequence, a new analytical paradigm is being advocated which tries to link both layers of genomic data, namely the genotype and the transcriptome. The many SNPs discovered by GWAS which were located in non-coding regions of the genome might actually play a vital role in controlling and regulating the gene expression patterns of the transcriptomic landscape.

These novel insights spurred the emergence of a new field of genetic study that tries to investigate the functional relations between genetic variations, transcriptome variability, and disease susceptibility by identifying the associations between those three entities which make up the molecular mechanisms in the form of an interaction network that underlies disease susceptibility. This emerging field is named *expression quantitative trait locus* (eQTL) study.

Identifying the regulatory SNPs that are responsible for altering the transcriptomic landscape leading to the onset of disease might prove very helpful in expanding our knowledge about the underlying mechanisms of phenotype formation and disease pathogenesis.

The knowledge we gain about the intrinsic systems related to the onset of disease by discovering the relationships between genotype, transcriptome, and phenotype in eQTL data, can be used to facilitate the development of new diagnostic strategies and optimize the treatment of disease, a goal that is broadly understood to fall under the domain of personalized medicine.

The contribution of this thesis is the analysis program qMAP (*quantitative MDL Association Program*) which extracts the various associations between genotype variants and gene expressions present in eQTL data based on normalized maximum likelihood (NML) encoding, the latest instantiation of Jorma Rissanen's information theoretic minimum description length principle (MDL).

As it is well known in information theory that data compression and knowledge acquisition in machine learning are equivalent concepts, we discover the various connections of the interaction network by encoding the eQTL data using the normalized maximum likelihood code and then look for associations between genomic and gene expression variants that yield short codelengths, i.e. minimum description lengths of the eQTL data.

This is accomplished by proposing an MDL-score for reporting the interaction association strengths. The MDL-score is constructed by building NML-codes for

eQTL data based on dynamic grid optimization techniques originally put forward by Kontkanen and Myllymäki.

Associations between SNPs and gene expressions are assigned an MDL-score that indicates the strength of the statistical functional relation between that feature pair.

The way of reasoning according to the MDL principle is that a SNP which regulates the expression of a gene can be used as a model to explain the phenomenon of the gene expression pattern. If the SNP turns out to be a good ground for accurately predicting the gene expression pattern, then the resulting NML-codelength of the gene expression, which uses the SNP as a statistical predictive model, will be short.

Interactions in eQTL data are identified by qMAP on the premise that prime candidates exhibiting functional and regulatory capability are those SNPs which produce short NML-codes during the encoding procedure of gene expressions and simultaneously have a short NML-code for themselves resulting in a minimum total description length for that interaction pair.

qMAP was benchmarked against the popular genome analysis toolkit PLINK, state-of-the-art information theoretic data exploration approach MIC, and Shannon's mutual information.

The ability to accurately and completely extract all the interactions in an eQTL datasets was assessed using a simulated eQTL study based on a synthetic dataset that was created and provided to the author by Bartlett and Ray. This was achieved in the form of measuring the detection rates, i.e. the success in identifying correct SNP-gene associations in the data, and the reconstruction of the interaction network.

Compared to MIC and PLINK, qMAP was able to considerably improve the detection rates. Existing methods delivered detection rates for correct genotype-transcript associations of the magnitude 57.3% for PLINK and 52% for MIC. qMAP raised those detection rates to 78%, an improvement of roughly 20 percentage points. The especially adapted approach using mutual information and kernel density estimates, abbreviated as MI-KDE, could be tuned to deliver detection rate results comparable to those of qMAP.

The main strength of qMAP which surfaced during our study was its robustness and reconstruction capability of the interaction network that was contained in the synthetic eQTL data. Among all tested approaches, qMAP delivered the most complete image of associations between SNPs and gene expressions of the interaction network.

Consequently, with qMAP the reconstructed interaction networks are more accurate and complete than would be possible if using existing analysis tools.

Hopefully, qMAP will prove useful to the biomedical community and physicians who aim to learn more about disease through the study of big eQTL datasets.

2 Paving the way towards better understanding of disease: From GWAS to eQTL studies

2.1 Introducing the road from gene mapping to integrated genomic interaction studies

What kind of new knowledge about the mechanisms of disease pathogenesis can be revealed by analyzing the huge amounts of genomic data which are being generated by next-generation sequencing machines? The answer to that question is that the development of our understanding of genetic mechanisms responsible for human diseases is an ongoing process and that technological advances help us reveal one piece at a time of that interesting puzzle.

Of particular interest is the functional understanding of the information encoded in the DNA and the various mechanisms it influences in an organism. Therefore, a lot of effort has been invested in developing new sequencing technologies that enable us to open new frontiers in genomic analysis.

After completion of the first big milestone, namely the assembly of a human reference genome by the Human Genome Project [1, 2], it became clear that the phenotypic diversity of traits in a population can be traced back to differences in the DNA sequence.

A phenotype is a broad concept that describes an observational trait like for example eye or hair color, the body mass index, or even a disease like diabetes or asthma. It is well known that our genetic code is composed of the 4 bases A,T,G,C which stand for *Adenine*, *Thymine*, *Guanine*, and *Cytosine*. When two genome sequences are compared with each other it is possible to detect genome variation.

There are several types of genome variation present in a DNA sequence; when only one base is altered we speak of *single nucleotide polymorphism*, in short SNP. Apart from SNPs, many other variations have been observed in the human genome, like insertions and deletions, referred to as InDels, tandem repeats,

translocations, inversions, and copy number variations (CNV) [3]. All those alterations in the genomic sequence of the DNA could have an effect on the phenotype and in particular play an important role in disease pathogenesis.

With the costs for genetic sequencing steadily falling, it becomes possible to obtain many genomic datasets which can be used in an analysis to identify the associations between a genetic variant and a specific phenotype. In many studies the phenotype under investigation is a disease, like e.g. cancer or Alzheimer, and the aim is to identify genetic markers, usually SNPs, that are linked to the disease. The finding and identification of such markers through analytical methods is referred to as genetic mapping and *genome wide association studies* (GWAS) try to elucidate the underpinnings of disease [4].

Manolio [4] gives an excellent review about the current status of genome wide association studies. After the success of the Human Genome project, researchers around the world made use of sequencing technologies in order to gain a better understanding of the relationship between genetic variants and disease.

When a single nucleotide polymorphism occurs in the exonic, i.e. protein coding sequences, region of a genome, it can have severe effects depending on whether is a synonymous or non-synonymous mutation. While synonymous mutations do not alter the transcribed protein for which the gene codes for, a non-synonymous mutation can have a devastating effect, because it changes or even stops the protein sequence which is encoded by the gene. This can have adverse effects on the organism and even result in disease pathogenesis as outlined in the review by Lathrop [5] and Manolio [4] .

Although countless studies have revealed associations between SNPs and disease phenotypes over and over again [4], the functional complexity of the DNA has been largely underestimated. Despite the fact that many studies successfully proved the impact of SNPs [4, 6] on phenotypic traits, questions about the molecular interaction mechanisms which lead to the onset of disease could not be answered by GWAS alone [4, 5, 7, 8].

As pointed out by Manolio in [4] most SNPs identified by GWAS fall into regions in the DNA sequence which are not known to code for any protein. According to Manolio [4], the present distribution of GWAS reported SNPs is as follows: 12% seem to occur in exonic regions which means that there potentially exists a direct influence on the protein sequence that could lead to disease pathogenesis resulting in a pathogenic state. In those cases, the underlying disease mechanisms can readily be traced to malicious effects of the SNPs on protein coding sequences. The remaining SNPs fall into intergenic and non-coding regions of the genome, a fact which points towards a more indirect effect of SNPs on disease, namely that the route of effect transmission might be as follows; genetic variants in non-coding regions of the genome have an impact on the regulation of transcription activity, which is the gene expression activity. In turn, modifications in gene expression

patterns result in disease onset [5, 8]. This observation hints on the importance of expanding genome-wide association studies into studies that also take other traits, like for example variation in gene expression, into account.

Several recent studies have revealed that SNPs originating from non-coding regions of the genome can have a strong impact on the transcription activity of a gene (an excellent overview is given in [5, 7]). In other words, the DNA's functional repertoire can use SNPs as a tool to regulate gene expression.

Pritchard [7] suggested that the magnitude of divergence observed in gene expressions results in a much larger set of phenotypic traits, which could not be attained by genomic variation alone. It seems that according to several authors [5, 7, 8], differences in transcript abundances have a stronger association with disease and that interaction networks of SNPs and gene expressions play an important role in disease pathogenesis. This statement is enforced by the observation that SNPs which fall into non-coding regions in GWAS, are shown to have a regulatory effect on genes which have strong evidence of association with a disease [9].

This has lead to the emergence of a new type of study, namely *expression Quantitative Trait Loci* (eQTL) which tries to bridge the gap in understanding between GWAS results and the actual functional repertoire of the DNA. Lathrop [5] and Bartlett [8] suggest that perturbations in gene expression activity are involved in disease susceptibility and therefore the study of interaction networks might reveal further insights into disease pathogenesis.

Interaction networks are composed of regulatory SNPs and genes whose transcript abundance is modified by the presence of genetic variation. The aim of an eQTL study is to identify those SNPs which are associated with gene expression patterns, thus hinting at a regulatory mechanism.

When combined with GWAS data, eQTL becomes a very powerful tool for discerning potential interaction forces involved in disease susceptibility and pathogenesis [5, 8]. This feat was demonstrated in [10] showing that cholesterol levels are affected by a gene whose activity is influenced by a SNP. Consequently, eQTL studies can deliver intriguing new insights regarding our understanding of disease.

2.1.1 "Data deluge" in next-generation sequencing

One peculiar observation regarding the development of next-generation sequencing is that the reduction in cost for obtaining a sequenced sample and the explosion in the quantity of data obtained from a sample has outstripped Moore's Law [11–13].

This technological development has both advantages and disadvantages. Let us begin with mentioning the benefits that cost reduction in sequencing technology bring. First, it becomes possible for more laboratories to utilize this technology

in order to study the mechanics of genomic systems. Hopefully, the more laboratories make use of sequencers the faster we will get a deeper understanding of the functioning of the DNA and use genetic data to combat disease. Genetic markers identified to be associated to disease pathogenesis or disease susceptibility might be used to improve clinical diagnostic tests and even aid in disease treatment.

Another bright side of this development is that it becomes feasible to study rare diseases [14,15]. Although the study of rare disease might not get enough financial support because it falls below the radar of public interest, reduced sequencing costs enable moderately funded institutions to perform genetic studies on rare diseases. Despite the fact that the pool of affected persons might be small and that it was previously prohibitively costly to perform such a study, nowadays laboratories can perform sequencing studies on subjects with a rare disease. As a consequence, the circle of beneficiaries is expanded in society. Individuals whose disease status could not be genetically examined in the past, might benefit from the results of new studies that are underway.

Recently, genetic analysis services have started to enter the consumer market with 23andMe [16] and DeNA in Japan offering genetic variant screening tests whose results show known risk factors for a variety of conditions. The implications of these genetic testing companies are vividly discussed in academic and business circles as ethical as well as practical issues surrounding the topic still need to be resolved. Since genetic analysis studies point out that genomic variants are not the only source that contribute to disease and that lifestyle and epigenetic factors also play an important role, screening results from the above mentioned companies should be interpreted with caution. In my opinion it is advisable to consult a genetics expert or a medical doctor when trying to make sense out of the results delivered by commercial services.

A disadvantage of the "data deluge" [12, 13] is on the other side, that sequencing machines output more and more data while data processing systems and analytical methods cannot keep pace [12]. This not only leads to rising demands in disk drive capacity for storing all those genomic datasets but also a sharp increase in computation time for processing and analyzing genomic data. Without the use of supercomputers like the one at the Human Genome Center [17], it would be virtually impossible to extract useful information out of the vast amounts of data.

Supercomputers in combination with proper analytical tools can help researchers obtain answers to their questions regarding the complex functional interactions exerted by the DNA.

As a consequence, this has led Hagenauer [18, 19] to suggest that new analytical algorithms are needed to deal with genomic data and because genomic analysis resembles analysis of information, why not try to introduce analytical tools based on information theory for the analysis of next-generation sequencing data to the

biomedical research society.

2.1.2 Insights into Mendelian disease with next-generation sequencing

Several studies employing DNA-sequencing, especially the more cost effective exome-sequencing method, are already uncovering a plethora of information about disease associated SNPs.

In [15] Ng et al. used a method called *exome-sequencing* in order to clarify for the first time the genetic causes of Miller syndrome, which is a rare Mendelian disease. The attribute "rare" is used to signify that the number of affected persons among the general population is quite low. The disease manifests itself as a series of multiple physiological deformations which can be examined along with further details about the Miller disease in Ng's study [15].

The achievement of Ng's group was to pinpoint the genetic causes of the Miller syndrome to mutations in the DHODH gene, a feat the was not possible according to [15] with discovery approaches not relying on DNA-sequencing. The responsible mutations were found in a protein coding gene that produces an enzyme which acts in the biosynthesis pathway [15]. These results explain the motivation behind using exome-sequencing instead of the much more expensive whole-genome sequencing.

With exome-sequencing it is possible to capture most of the protein coding genes of a human genome and since non-synonymous mutations in protein coding genes can have severe impacts on the organism, exome-sequencing is a cost effective method for identifying causes of Mendelian disease as outlined in [14]. Current toolkits capture only a fraction, hovering around 5%, of the entire human genome, thus making it possible to spend the saved money on increasing the read coverage of the sequenced regions in order for the SNP calling algorithms to deliver more reliable results.

After sequencing the exons and performing SNP calling on the obtained sequence data, the analysis pipeline of Ng et al. consisted of consecutive filtering against databases of known variants like dbSNP [20] and HapMap [21] and then using an inheritance model to pinpoint the remaining SNPs as the causes of Miller Syndrome. The identified mutations were then confirmed using another sequencing method called Sanger sequencing, which produces much more reliable results than shotgun sequencing methods, but has a lower throughput.

2.1.3 Achievements of GWAS: Uncovering the genetic map of complex disease

Now we leave the world of Mendelian disease and focus on the class of *complex disease* [22] like Diabetes. In contrast to Mendelian disease, *complex disease* are said to be the outcome of several factors occurring together like genetic disposition combined with environmental influences and lifestyle [22].

As a consequence, the analysis of complex disease is much more difficult and challenging than that of a Mendelian disease, but the advent of genome-wide association studies (GWAS), an important tool for identifying associations between phenotypic traits and genomic variants [24], paved the way towards gaining a better understanding of the genetic workings responsible for the development of traits like hair and eye colour, a person's body mass index (BMI), and diseases like Asthma and Diabetes [4, 23].

Most of the traits that we can observe in a human being cannot be traced back to one single gene or SNP in our genome. They are rather the result of a combination of several genetic and non-genetic factors. Hair color for example is not determined by a single mutation in a person's genome and it has been shown that many SNPs and genes are responsible for determining the hair color of a person.

For this reason in a GWAS study, a multitude of SNPs is tested for association with a specific trait. In a simple setup, a case-control study is performed where the trait is present in one group but not in the other.

A nice way to visualize the results of such an association study are Manhattan Plots. On the horizontal axis there are the SNPs, possibly sorted by chromosome and on the vertical axis their association strength with the trait. The plot resembles a city skyline, with strongly associated SNPs rising from the ground like skyscrapers. By setting a certain threshold value, or visually speaking, setting a certain minimum height to define a skyscraper, SNPs that are thought to be related to the trait are thus filtered from the dataset.

As is the case with traits like height, which is the outcome of an interplay between several genetic, environmental and lifestyle factors, the same is true for several disease. In such cases it is not possible to pinpoint the onset of the disease to just one or a few genetic markers. Examples of those cases are Alzheimer's, Scleroderma, Asthma, Parkinson's, Diabetes and many more other diseases. Because a majority of disease falls into this category, where multiple factors contribute to pathogenesis, these diseases are termed *complex disease* [4, 23].

Because a complex disease is characterized as an accumulation of various combinations of genetic factors plus environmental influences, physicians hardly ever speak of that a patient will get for example Alzheimer's when a genetic test that scans the genome for known Alzheimer associated genetic variants turned out to be positive [25]. Instead, phrases like "elevated risk" are used to describe the

uncertainty associated with the results of the genetic test and to assert the importance of lifestyle and environmental factors, both of which a patient can have more or less control of and thus has the possibility to influence his or her risk of contracting a disease.

To understand how GWAS extracts information from DNA-sequence data and how it helps us to better understand the development of physical traits and disease pathogenesis, we will take a look at recent findings of GWAS publications.

The basic methodology used in GWAS is called *gene mapping*. By using an association measure, may it be Spearman correlation or mutual information, the "strength" of an association between a genetic marker, usually a SNP, and a trait is quantified. Since there are so many genetic markers and each of them is tested for association, a sufficiently large sample size is required for producing good (i.e. correct or reliable) results.

An excellent review by Manolio [4] about GWAS appeared in The New England Journal of Medicine and the reader is referred to his article about a survey of intriguing recent GWAS discoveries in medicine. Only a few selected examples will be presented here, enough to give the reader an impression about the discovery power of GWAS.

With DNA-sequencing it is possible to capture the genetic variation present in an entire genome. Ongoing advancements in sequencing technology alleviate former problems of DNA-sequencing studies, namely of not being able to produce a large enough sample size due to cost restrictions. As a consequence, studies whose sample size is not large enough have difficulties outputting "meaningful" results, but technology improvements have leveled this barrier.

Outside the academic literature, it is often reported that certain mutations are causing a trait or disease. While this might be true in a common sense, there exists no mathematical framework of analysis for identifying causalities yet. Therefore, we will try to avoid talking about "disease causing genes or mutations" in this thesis, and if a sentence states or implicates that a causality exists, it is meant that only a statistical association between a genetic marker and a trait exists. This is to prevent any misunderstandings when reporting results in this thesis.

Some of the more recent interesting uncovering results of GWAS are a magnitude of associates between genetic markers and traits like intelligence, BMI, and body height.

A Nature Genetics study by Yang et al. [23] reports that it has found 87 multiple associated SNPs for body height out of which 49 have not been published before. The authors of [23] argue that the effect sizes of single associations are usually very weak because as we have already explained, complex traits are the outcome of a complex orchestration between many genetic factors (not to forget other non-genetic influences).

Consequently, on top of just looking for single associations, a SNP-trait association has to be analyzed in context. This means that the joint impact of SNPs has to be tested for their association with the trait.

The expression of a phenotypic trait can depend on the joint presence of multiple SNPs in a person's genome as well as a conditional combination of SNPs. If one were to perform association tests between all potential combinations of SNPs and a phenotypic trait, this would lead to an almost incomputable explosion of the potential search space.

As a countermeasure, several algorithms have been proposed in the literature to deal with this problem and some of them will be explained in subsequent part of the thesis in Section 3.2.

One interesting observation that was made by Yang et al. in [23] is that several SNPs are strongly associated with body height, but no such statement could be made for the body mass index (BMI) yet. The authors suggest the explanation that phenotypic trait of height has more to do with inheritance than BMI, which means that it is more likely that you inherit your height from your parents rather than the shape of your body (talking of mass, of course). Furthermore, although both height and BMI are traits not free from environmental influences (nutrition, eating and lifestyle habits), it looks like that non-genetic factors have a stronger impact on our BMI than our genetic make up; meaning the way we exercise (or not) basically influences our weight.

This result exemplifies that we are not completely bound to our DNA and that our actions (or in-actions) have a very strong influence on our health and well-being. Although we can say that persons with certain mutations in their genome are predisposed towards a condition, it is not certain that this risk will ever materialize. Genetic testing in combination with counseling together with a decisive lifestyle action can alleviate or prevent health care problems. Thus, instead of undergoing a gene therapy to reduce the BMI it might be more advisable to just visit the treadmill a few more times a week.

Another important finding of [23] is the illustration of the complexity of the genetic mechanisms of "complex" traits [22, 26]. Even though body height and BMI appear to be regulated by many SNPs, the contribution of a single SNP to the phenotypic trait quantified via the association measure is rather small. By looking at the quantified contribution and combinations of SNPs, Yang et al. argue that the *"genetic architecture"* [23] of both traits is inherently different.

What does *"genetic architecture"* according to Yang mean [23]? It was observed for both height and BMI that there are strong associations between a leading array of gene variants and the studied trait. On top of that array there exists another layer of SNPs which control the height trait in humans. These SNPs appear to exert control via either a jointly or conditionally depended functional relationship.

Those layered levels of genetic control are becoming an intensively studied topic. Due to the interesting findings of sequencing studies, it is becoming more and more apparent that the genetic mechanisms of trait expression (and disease pathogenesis) are more dynamic and complex than originally anticipated. This has led to the science of *"systems biology"* [27], which studies genetic interactions as a dynamic, cybernetic system. We will have more to say about this topic when we reach RNA-seq studies about gene expression and eQTL studies. For now it is sufficient to keep in mind that complex traits are regulated by more than one single SNP and that a complicated interaction network controls and governs the onset of traits and disease pathogenesis.

In [4], Manolio surveyed several GWAS publications and found several disease conditions that share genetic markers, despite the fact, that from an outside observer's standpoint, those disease might seem to be totally unrelated.

This point of view is contrasted by the facts reported by Manolio [4], namely that presumably independent and unrelated traits, share a common base of associated SNPs between them. This really sheds a new light on our understanding of those traits. Furthermore, it also opens up several new questions about the relation between those traits, questions that need to be answered by future studies.

For example, one intriguing observation by Manolio [4] is that type 2 diabetes seems to have shared genetic markers with several other disease conditions. The implications of these findings are slowly beginning to surface. About the consequences that type 2 diabetes has 2 common genetic markers with prostate cancer and 1 with height as well as coronary disease can only be speculated at this moment.

2.1.4 The road ahead: eQTL and beyond

Recently, another layer of complexity was introduced to genetic studies. Whereas the first DNA-sequencing studies concentrated on finding associations between genetic mutations and traits, according to several studies [5, 7, 28, 29] the phenotypic diversity that we see amongst organisms, especially amongst individuals of the same species, might not be the direct consequence of genetic variety in an organism's genetic code per se, but rather a multitude of gene expression patterns which allow for a much larger phenotype spectrum.

The technological developments that enable researchers to screen both the genome for genetic variations and the expressed genes for alterations in transcriptional activity are DNA-sequencing and RNA-sequencing [30]. With RNA-sequencing it becomes possible to capture the entire gene expression profile at once, also called the *transcriptome*, which includes expression values for both coding and non-coding regions [31]. A gene's expression value hints at the transcriptional activity of a gene, which means that there are more transcript copies of an active

or highly expressed gene in a cell whereas there are fewer RNA copies of a lowly expressed gene.

As was already mentioned in Section 2.1, evidence is mounting in favour of the hypothesis that DNA sequence variations alone cannot explain the vast and diversified manifestation of phenotypes and that rather the flexibility of the transcriptome is producing the magnitude of observable traits [5, 7].

As a consequence, the studies of genome wide association, which link genetic variants in the genome to phenotypic traits like disease, were extended to include an additional layer of information, namely the gene expression patterns obtained via a method like RNA-sequencing.

In contrast to SNP mutations, which in their simplest case consist of switching one of the 4 nucleotides and therefore are considered to be a discrete trait, the expression range of a gene's activity is measured on a continuous scale and is therefore referred to as a quantitative trait. For this reason, studies which aim to map a discrete trait like SNPs to a continuous trait like gene expression are called *quantitative trait locus* (QTL) studies.

A set of quantitative traits suitable for QTL could be for example height, blood pressure, cholesterol levels, or even the body mass index (BMI). To highlight the importance of association studies that try to link genetic variants to gene expression patterns an "e" is prefixed to QTL to distinguish those studies, thus obtaining eQTL for *expression quantitative trait loci* study.

According to Lathrop [5] and Pritchard [7] the benefit of conducting eQTL studies lies in the hope to uncover more knowledge about the underlying molecular mechanisms of disease and disease susceptibility by linking the SNPs previously identified in GWAS studies (but found to be located in non-coding regions with thus an unknown effect on the disease) with gene expressions whose activity could be involved in disease pathogenesis.

This elevates the importance of the previously thought to have no function "junk"-DNA (therefore the name "junk"). About the reason for the existence of these non-coding regions of the genome has been speculated a lot and recent developments support the hypothesis that non-coding parts of the genome like long non-coding RNAs actually have a major role in gene expression control.

This helps us reveal another level of functional complexity of the DNA, because links between SNPs, gene expressions, and phenotype (disease or trait), form a kind of interaction network.

Bartlett et al. [8] claim that an integrated analysis approach is required to fully reap the benefits of modern sequencing technologies for gaining a better understanding of disease and more importantly, find new methods to cure them. Especially, the areas of personalized medicine, better biomarkers for novel drugs and clinical tests, as well preventive screening are pointed out in [8].

A first step towards achieving the goal of integrated analysis, i.e. the simultaneous study of a patient's genetic and transcriptomic profile, a large enough number of datasets have to be collected in order to be able to establish links between the three entities phenotype, gene expression profile, and genetic code.

Before presenting some recent examples collected from the literature that clearly show the benefits of extending GWAS with eQTL and mapping SNPs to gene expressions, let us first state how a SNP associated with the transcription activity of a gene can "interact" with it.

One well understood mechanism described in [5] is SNPs that fall into the *transcription start site* (TSS) of a gene and thus dampen or enhance its transcriptional activity. Moreover, a study by Ogawa et al. [32] recently revealed by using next-generation sequencing, that mutations in cancer affect the splicing machinery of cancer related genes and alter their expression activity.

As a consequence, in order to be able to clarify the effects of SNPs on gene expression, SNPs found to be associated with certain genes are labeled as either *cis*-acting or *trans*-acting. It should be noted that the terminology regarding cis- and trans- effects in eQTL studies is different from other genetic studies, as was pointed out by Pritchard in [7].

Although the definition of cis- and trans- acting SNPs in eQTL studies is somewhat arbitrary and could be different in various studies, both Lathrop [5] and Pritchard [7] state that in general a cis-acting SNP is understood to be located in close proximity to the gene it is associated to. The distance on the DNA strand is measured in base pairs (or nucleotides). This means that the distance between cis-acting regulatory SNPs and the gene being regulated is short. In contrast, trans acting SNPs have a large distance from the gene they are influencing. Therefore, in eQTL studies *cis* and *trans* are terms indicating the distance of a gene associated SNP in genomic coordinates on the DNA.

Simply speaking a cis-acting SNP is "near" or "close" to the associated gene whereas a trans-acting SNP is "far" or "distant" from the gene it influences. The term trans-acting SNP even includes the possibility of the SNP being located on a totally different chromosome.

According to [5, 7], a cis-acting SNP is most of the time classified to be within range of 100kb upstream or downstream of the gene that it is associated to. All other SNP effects are categorized as trans-acting SNPs.

Despite the fact that the effect of cis-acting SNPs on the actual gene expression is stronger than that of trans-acting SNPs (shown in several studies), trans-acting SNPs seem to be on the other hand more numerous. Because trans-acting SNPs are equipped with the ability to influence the transcriptional activity of many genes, they have become known in the literature as "master regulators" [5, 7].

The merit of emerging eQTL studies is twofold; One, novel insights into suscep-

tibility to and pathogenesis of disease are promised to be revealed. Two, a better understanding and interpretation of the results of GWAS can be attained.

Let us first present some studies that shed new light on potential disease mechanisms utilizing eQTL.

A study conducted at the San Antonia Family Heart Hospital took a closer look at the genetic underpinnings of the various cholesterol levels of its patients and found that the expression level of a gene linked to cholesterol is influenced in a cis-acting manner by a SNP [10].

Another study [33] revealed by analyzing post-mortem brain tissue of Alzheimer patients that SNPs are associated with the genes MAPT and APOE, both of which are thought to have major involvements in disease pathogenesis of Alzheimer's. Thus, an important connection between genetic variants of a patient and the molecular mechanism of Alzheimer disease is drawn in [33].

To come back to the topic of how eQTL helps surface knowledge about molecular mechanisms of disease pathogenesis, it also shows us how disease are related with each other on a molecular level.

Although it is known that several genomic variants are simultaneously associated with several disease, as can be concluded from the overview tables provided by Manolio in [4], it was unknown in what kind of fashion those SNPs have the ability to induce disease pathogenesis. This changes with the advent of integrated genotype, gene expression, and disease status information analysis.

Manolio points out that Asthma shares many genetic variants with other disease, all of which have been uncovered by GWAS, but it was totally unclear how those SNPs interact in order to have an effect on disease until eQTL studies came along. [4].

Asthma has been studied by many researchers [4,5,9], and the studies concluded that there is strong evidence that the gene ORMDL3 has an important function regarding Asthma.

The first parts of a molecular mechanism that is involved in Asthma were uncovered by Moffatt et al. in [9], who showed that several cis acting SNPs appear to regulate the expression of the ORMDL3 gene in Asthma patients.

Integrated genome and transcriptome studies also have the potential to identify relationships between disease on a molecular level, as was already done for Asthma and Crohn's disease.

On the same standing regarding the importance of ORMDL3 to Asthma, the gene PTGER4 is believed to have a strong role in Crohn's disease [5].

A connection of molecular mechanisms between Asthma and Crohn's disease was discovered in [34, 35]. The researchers observed that there is a concurrent

correlation between SNPs and both the expression levels of PTGER4 in Crohn's disease and ORMDL3 in Asthma.

From these results it can be concluded that there seems to be shared or common biological mechanisms that underlie both disease. The notion of several disease sharing genomic interaction networks is not new and has already been reported several times, as for example by Manolio in his GWAS review in [4].

The novelty of eQTL is its ability to shine light on the molecular mechanisms of disease which manifest themselves via genetic variants exerting an effect on transcriptome variation whose regulatory interaction networks play a role in disease pathogenesis and susceptibility.

What lies beyond eQTL is the observation that other genetic variants and even lifestyle factors have and could have an influence on the function of the DNA. Except SNPs there are other genetic variants like CNVs and InDels that could alter and perturb the regulatory mechanisms of the DNA. On top of that, evidence of epigenetic mechanisms influencing the activity of genes is beginning to emerge with histone modifications and methylation patterns playing an active part in genome regulation. Last but not least, the effect of our lifestyle on our well-being should not be underestimated.

3 Applications of information theory in eQTL analysis

3.1 Information theoretic approaches to the gene mapping problem

Because of the importance of gene mapping to further our understanding of the genetic basis of traits and complex disease, a lot of work has went into the development of tools and algorithms to accurately extract associations between SNPs, gene expression, and phenotypic traits.

Especially in the field of biomedicine the results of association studies that either link genomic variations to phenotypes (GWAS) or discover connections between genetic variants and gene expressions (eQTL) have a big impact on our understanding of how phenotypes emerge and disease progress.

Discovery of novel biomarkers can help researchers devise non-invasive genetic tests for pre-screening patients for potential disease risks. Even though medications based on gene therapy have not reached market potential yet, an intrinsic understanding of the dynamic interactions between SNPs and gene expressions and how they influence disease progression is of vital importance for the creation of such therapies. Therefore, in this chapter we will discuss some selected applications of information theoretical analysis methods in the domain of gene mapping and DNA sequence information analysis.

At this time it is necessary to introduce some basic formalism and concepts of information theory in order to better apprehend the following work. We will explain the methodology that will be used to analyze the eQTL datasets, define the random variables that are necessary to work with the datasets, and provide a compact outline of information theoretic concepts in the context of DNA- and RNA-sequence analysis.

3.1.1 Mathematical representation of eQTL data

When performing gene mapping using DNA- and RNA-sequence data in case-control studies, we are usually presented with a dataset that contains the geno-

types, i.e. SNPs, of the sequenced population, gene expression profiles for a number of genes, and information about a phenotype, may it be a specific trait or a disease status.

The task of an analysis is to extract "relevant" or "meaningful" association from such a dataset that can be used to further our understanding of the genetic principles that govern the creation of a specific phenotype.

Let us start with the genotype information. Via DNA-sequencing we identified a total number of L SNPs at various locations in the genome. Since a person inherits at each genomic locus l one allele from the mother and the other allele from the father, a single nucleotide polymorphism (SNP) is defined to be any genomic locus l that experiences nucleotide variations.

It has been reported [18] that at each locus l usually no more than two major alleles are present, which means that only a combination of two bases out of the 4 letter DNA alphabet A, T, C, and G appears as a SNP.

Let us name the random variable for SNPs S and genetic variation in a dataset is represented as a chain of tuples that hold the discovered nucleotide combinations, which is illustrated in Table 3.1.

Subject	SNP 1	SNP 2	$\cdots$	SNP L
Patient 1	(A, A)	(C, G)	$\cdots$	(T, T)
$\ldots$	$\ldots$	$\ldots$	$\cdots$	$\ldots$
Patient N	(T, A)	(C, C)	$\cdots$	(T, T)

Table 3.1: Representation of genetic variation discovered in N patients for a total number of L SNPs.

Furthermore, current sequencing technologies make it difficult to distinguish the gametic phase which means that it is not clear from the SNP data which SNP was contributed from the father's side and which one was contributed from the mother's side.

As a consequence, we assume throughout this work, that at any locus l containing a SNP in the genome, we can only distinguish two homozygous cases and one heterozygous case.

To illustrate this concept with an example, let us assume that we sequenced a person's DNA and obtained the following SNP (A, G). For our analysis to work, a coding concept needs to be established that transforms this SNP into a numerical value. A coding scheme was suggested by Hagenauer et al. in [18, 36] which we will apply in this thesis to map SNPs to numerical values.

Since we can effectively only distinguish between 3 states, namely the two homozygous cases (A, A) and (G, G) and one heterozygous case which can either be (A, G) or (G, A), the SNP state (A, A) is mapped to 0, (A, G) and (G, A) to 1,

and (G, G) to 2. The SNP encoding mechanism of [18, 36] is illustrated in Table 3.2.

Genotype SNP Variation	Encoded Numerical Value
(A, A)	0
(A, G)	1
(G, A)	1
(G, G)	2

Table 3.2: The encoding procedure for bi-allelic SNPs according to Hagenauer et al. [18].

With the coding scheme of Hagenauer [18] it is possible to convert the SNPs into numerical values and make the data accessible to analysis.

Hence, the SNP dataset can be represented as a numerical vector $\mathbf{S} = [S_1, \cdots, S_l, \cdots, S_L]$ of length L with the SNP random variables S_l taking one of the values from the alphabet $\mathcal{A}_S = \{0, 1, 2\}$, i.e. $S_l \in \{0, 1, 2\} \quad l = 1, \cdots, L$.

In addition to gene variants, we must also introduce random variables for the gene expressions, which are part of eQTL data.

Via RNA-sequencing (RNA-seq) we obtained gene expression values for a total number of G genes. Their expression values are represented by the random variable E_g and since gene expressions are continuous $E_g \in \mathscr{R} \quad g = 1, \cdots, G$, with G being the total number of genes for which we have expression values in our dataset. We summarize the RNA-seq data in a vector $\mathbf{E} = [E_1, \cdots, E_g, \cdots, E_G]$.

Finally, we treat the phenotype trait random variable T of the eQTL dataset. Although the phenotype trait can either be a discrete random variable in the case of discrete traits like disease affection status of a patient or a continuous random variable if we are dealing with phenotypic traits like a person's height, in this thesis the focus is on disease affection status. Therefore, the phenotype trait random variable T is discrete in nature.

Under the simple assumption of a case-control study in a cohort consisting of N individuals, the phenotype trait random variable for each individual T_i is a binary random variable, i.e. $T \in \{0, 1\}$, which describes if an individual falls into the case-category 1 or control-category 0. Therefore, the affection status data is given as a binary sequence vector $\mathbf{T} = [T_1, \cdots, T_i, \cdots, T_N]$.

We finally arrive at the representation of an eQTL data set which is comprised of SNPs $\mathbf{S}$, gene expressions $\mathbf{E}$ and phenotype trait $\mathbf{T}$.

3.1.2 Claude Shannon's information theory and its relation to genetics

Claude Shannon, the founder of information theory [37], introduced several important concepts which prove to be very useful when trying to extract useful information from genomic datasets. Although Claude Shannon is mostly known for his fundamental contributions to the field of information technology, without which modern computers and algorithms would not be possible, the reader of this thesis might be surprised to discover that Claude Shannon actually wrote his PhD thesis about genetics [38].

Of practical importance to medical doctors and biomedical researchers is the concept of entropy that Shannon introduced in his groundbreaking work. When performing an association study, entropy can not only measure the amount of information that is contained in a phenotypic trait, but also gauge the amount of information that the analysis reveals about it.

The outline in this Section of how information theory can be applied to analyze genomic data obtained from next-generation sequencing machines follows the work of Hagenauer et al. who investigated many potential use cases of information theoretic applications to genetics in a series of papers [18,19]. Therefore, the main ideas regarding the connections between information theory and genetic analysis are due to Hagenauer [18].

An example from a case-control study that investigates SNP associations with a complex disease, shows that the trait T can either indicate that a patient is affected or not. Assuming that the same number of affected and unaffected patients were sequenced, the maximum amount of information that physicians can uncover about the disease trait is 1 bit. This stems from the fact that the entropy [39] of a random variable, the trait T in our case, is defined as

$$H(T) = -\sum_{t \in \mathcal{A}_T} P(t) \log_2 P(t), \tag{3.1}$$

where $P(t)$ is the probability that an individual in our dataset has the specified trait t and $\mathcal{A}_T$ is the range of values that the realization t of the random variable for traits T can take. In our example of a simple case-control study the trait T is the disease affection status with $\mathcal{A}_T = \{0, 1\}$ and $P(T = 0) = P(T = 1) = 0.5$, thus making the entropy $H(T) = 1$bit, since the logarithm is taken to base 2. Unless otherwise stated, in this thesis we agree upon that the logarithm is always taken to base 2 and delivers by definition results in bits.

Should one wish to measure the impact of SNPs on a trait in an association study, the mutual information needs to be calculated. This can be easily achieved by introducing some more concepts from information theory (for a detailed treaty

about information theory refer to the book by Cover and Thomas [39] or MacKay [40] whose book is also freely available online: [41]).

The joint entropy $H(S,T)$ measures the combined information in a SNP-phenotype trait pair (S,T), which can be obtained with the formula

$$H(S,T) = -\sum_{s\in\mathcal{A}_S}\sum_{t\in\mathcal{A}_T} P(s,t)\log_2 P(s,t). \tag{3.2}$$

In this case $P(s,t)$ is the joint probability of the SNP and the trait occurring together. The probabilities are obtained by counting the occurrences, i.e. the number of times the SNP and the trait appear together, and then dividing the number by the sample size N.

Although with the two equations above we have the basic ingredients to define the mutual information, we will show also another type of entropy, namely the conditional entropy, because it will ease the interpretation of results and make the findings of information theory based association studies more comprehensible to the general audience.

In the case we know the information about a SNP at a certain locus in the genome, we can ask ourselves how much uncertainty remains in our estimate about the trait T. The remaining information regarding the phenotypic trait T is calculated via the conditional entropy

$$H(T|S) = -\sum_{s\in\mathcal{A}_S}\sum_{t\in\mathcal{A}_T} P(s,t)\log_2 P(t|s) = \sum_{s\in\mathcal{A}_S} P(s)H(T|S=s) \tag{3.3}$$

leading to the *chain rule* of entropy

$$H(S,T) = H(S) + H(T|S) = H(T) + H(S|T) \tag{3.4}$$

which means that the combined joint information in a SNP-trait tuple (S,T) is the information content of the SNP plus the remaining information content of the trait. Of course, a corollary is that given a certain trait, the joint entropy $H(S,T)$ can also be expressed as the information content of the trait plus the remaining uncertainty of the SNP.

This leads us finally to the definition of mutual information, also abbreviated as MI throughout this thesis:

$$I(S;T) = \sum_{s\in\mathcal{A}_S}\sum_{t\in\mathcal{A}_T} P(s,t)\log_2\frac{P(s,t)}{P(s)P(t)}; \tag{3.5}$$

an association measure with many important properties regarding gene mapping.

First of all, in contrast to measures like correlation, mutual information catches any kind of statistical relationship between SNPs, gene expressions, and phenotypic traits [19, 39, 40], thus enhancing any genomic analysis because it enables the analysis to go beyond detecting only simple linear associations.

Another interesting property of mutual information is that it is 0 iff (if and only if) there is no statistical association present.

Mutual information is also referred to as *"shared information"* [40], because it tells us how much information a trait and a certain SNP share. An interesting corollary of that observation is, that a person's genotype information tells us as much about its phenotype (trait) as does the phenotype (trait) tell us about its genome.

3.2 Related work

3.2.1 Mutual information relevance networks

One of the very early attempts of applying Shannon's concept of mutual information to the analysis of gene expression data was the pioneering work of Butte et al. [42]. The motivation behind his work was to understand how and if genes can influence the behaviour of each other by means of either suppressing or up-regulating their activity.

In order to discover those relationships, Butte et al. [42] measured the expression activity of several genes under various conditions and used mutual information as a dependency measure. Since the gene expressions were continuous random variables, a quantization, i.e. digitalization, had to be performed for making the analysis computational accessible. Although a simple quantization based on histograms was used, the initial study yielded satisfactory results [42].

Butte faced the problem of an exploding search space when considering joint and multiple associations between gene expressions. In his work [42] he introduced a solution to this dilemma, a concept known as *relevance chains*.

Let us say that we are looking within a pool of candidate genes for associations with a target gene expression E_{target}. In a first step, the relevance chains algorithm calculates the ordinary mutual information between our target E_{target} and all the other gene expressions $\mathbf{E} = [E_1, \cdots, E_G]$ in the data set, i.e $I = (E_{\text{target}}, E_g)$ for $g = 1, \cdots G$. That is the process for obtaining a list of genes which are most associated with the target gene according to the amount of mutual information. If one is only interested in single associations the relevance chain algorithm can be terminated here.

On the other hand, if an analyst wants to discover joint or multiple associations, this can be accomplished by continuing with the relevance chain algorithm.

The first link of a relevance chain is the gene with the highest mutual information that was discovered in step #1. Let us denote this gene expression as $E_{\#1}$. In order to look for joint associations, the detected gene expression $E_{\#1}$ is taken as the first input to be appended to the tuple of joint gene expressions which are to be tested for association with the target gene expression E_{target}. Together with the remaining gene expressions in the dataset $\mathbf{E}$ the tuple $(E_{\#1}, E_g)$ is tested for association with the target E_{target} by calculating the mutual information $I(E_{\text{target}}, (E_{\#1}, E_g))$.

The gene expression yielding the highest mutual information in combination with $E_{\#1}$ in step #2 is denoted as $E_{\#2}$ and forms the next link in the relevance chain. Then again, the process of step #2 is repeated by appending $E_{\#2}$ to the list, i.e. the list becomes $(E_{\#1}, E_{\#2})$, and checking the mutual information for the remaining gene expressions with the target by forming the combination $(E_{\#1}, E_{\#2}, E_g)$ and calculating the mutual information $I = (E_{\text{target}}, (E_{\#1}, E_{\#2}, E_g))$.

By concatenating the gene expressions with the highest mutual information value in each step, the relevance chain is extended until a terminating condition is met.

3.2.2 Gene mapping with Shannon's mutual information

Hagenauer et al. refined the work of Butte et al. in [19] and also contributed a formula that allows the calculation of mutual information between discrete and continuous random variables.

In contrast to Butte's work [42], Hagenauer and Dawy et al. [18, 19] focus on extracting SNP-trait associations in case-control studies using Shannon's mutual information as a dependency measure. Some of the several advantages of using mutual information are mentioned in their paper [19].

Apart from the benefits of using mutual information as a dependency measure, which have been mentioned in Section 3.1.2, it is argued in [19] that mutual information gives quantitative results that allow the physician or researcher who is performing the analysis to evaluate the importance of SNP-trait associations based on the amount of information that the SNP contributes to the trait.

Furthermore, a link between mutual information and the $p-value$ based on the χ^2-statistic is presented in [43] that allows attachment of $p-values$ to mutual information values, if the analyst deems this necessary.

When Dawy et. al. applied their method to the autoimmune disease dataset of Ueada et al. [44], their information theoretic method delivered the same results as the logistic regression method that was used in the original paper for analysis plus one potential novel marker that was not reported in [44] by the authors but brought to light in [19].

For finding associations between SNPs and the affection status of patients regarding the autoimmune disease dataset of Ueada et al. [44], Dawy et al. [19] calculated the mutual information as outlined in Section 3.1, and reported those SNPs as significant which have a higher mutual information value than a certain threshold.

Another important result of [19] is the derivation of a formula for calculating the mutual information between a discrete SNP random variable and a quantitative trait like gene expressions.

We will use this mathematical approach later in combination with kernel density estimates for analyzing eQTL data and finding associations between SNPs and gene expressions.

The method for calculating the mutual information for discrete S and continuous T in this case according to [19] is

$$I(S;T) = \sum_{s \in \mathscr{A}_S} \int_{\mathscr{A}_T} f(s,t) \log_2 \frac{f(s,t)}{P(s)f(t)} \tag{3.6}$$

where $f(\cdot)$ are the continuous probability density functions which have to be estimated from the available sample via kernel density estimators. The complete derivation of the above formula is given in [19].

Mutual information has also been applied in other successful works that use information theoretic methods to analyze genomic data [42, 45–48].

Practical implementation for calculating the mutual information in a mixed environment of discrete and continuous random variables

We use a slightly different approach for calculating the mutual information because it simplifies the implementation in Python and the computation of the kernel density estimate (KDE) of $f(\cdot)$.

Cover [39] already introduced the concept of *differential entropy* in order to measure the information content (or amount of uncertainty) [39, 40] of a continuous random variable. Thus, using Cover's definition of *differential entropy* h, we obtain the information content $h(T)$ of quantitative traits T as

$$h(T) = -\int_{\mathscr{A}_T} f(t) \log_2 f(t) dt \tag{3.7}$$

with $\mathscr{A}_T$ being the realm of realizations or support set of the random variable T and $f(t)$ the probability density function of T.

By utilizing the *chain rule* property of entropy, we first show how to calculate the conditional entropy between SNPs S and phenotypic traits T before proceeding to present our calculation formula for mutual information.

The motivation behind our approach is that it is easier to compute the kernel density estimate if the continuous random variable T is conditioned on the discrete random variable S. Experiments where the mutual information between SNPs and gene expression was calculated in eQTL datasets showed that our computation approach seems to be more stable in contrast to the method described in [19].

Therefore, instead of directly calculating the mutual information, we bypass it using a two-step approach that uses the conditional entropy, which is estimated according to

$$h(T|S) = -\sum_{s \in \mathscr{A}_S} P(s) \int f(t|S=s) \log_2 f(t|S=s) dt. \tag{3.8}$$

By combining the computation results of the differential entropy of the trait plus the conditional entropy of the trait given the SNP, our implementation computes the mutual information according to

$$I(S;T) = h(T) - h(T|S) \tag{3.9}$$

in order to obtain the degree of association between SNP S and trait T.

3.2.3 ARACNE: Reconstruction of gene regulatory networks

Of particular relevance to our work is the ARACNE algorithm by Margolin et al. [48]. Margolin et al. developed an interesting algorithm to reconstruct gene regulatory networks from gene expression datasets. The authors show that their algorithm outperforms other gene network estimation algorithms that either use *relevance chains* mentioned in Section 3.2.1 or Bayesian networks.

When dealing with MI-values that have been estimated from a limited sample size, either via frequency counts in discrete cases or kernel density estimates in continuous cases, the quality of the estimate of $P(\cdot)$ and $f(\cdot)$ greatly influences the quantitative value of mutual information, as outlined in [48]. Furthermore, Margolin et al. [48] point out that there is yet no way to automate this process, but an important result is given by the authors in their paper regarding mutual information ranked lists.

Even though it might not be possible to accurately calculate the absolute amount of mutual information, the impact on ranked lists of MI-values is small according to [48]. This means that if we calculate the MI between all pairs of gene expression values in a dataset and rank those MI-values, then, if we change the parameters for obtaining the kernel density estimates of the underlying probability density functions necessary for the computation of mutual information, the relative rank of a gene expression pair's MI-value remains approximately the same. This theorem was empirically proven in [48].

Thus, ARACNE proceeds constructing a gene network by first calculating the pairwise mutual information between all combinations of gene expressions and drawing an edge between gene pairs that share information with each other. After the first step has been completed, an interesting application of the *data processing theorem* [39, 40] of information theory reduces the amount of false positive (or weak connections) in the dataset and only edges between gene pairs with "strong" MI-values are kept whereas weak edges are discarded.

3.2.4 Maximal information coefficient

A recent promising development in finding useful associations between SNPs and traits, consequently extracting "useful" information from eQTL datasets, is the introduction of the *maximal information coefficient* (MIC) by Reshef et al. [49].

The paper of Reshef et al. tackles the problem of obtaining a good estimate of mutual information between a pair of random variables by trying to maximize the amount of MI via a dynamic grid partitioning method. According to the authors, by maximizing the mutual information value in this way it is possible to perform data mining on any kind of dataset and extract novel associations from the data [49].

Before presenting the claimed benefits of the MIC method, we will briefly outline the MIC calculation methodology when applied in the context of eQTL analysis.

Let us assume we have an eQTL dataset comprised of N patients in the form of Section 3.1 and we are interested in obtaining the MIC of a SNP-trait pair, where the trait in this case are the gene expressions. Consequently, the MIC computation method of Reshef et al. [49] proceeds as follows.

The vectors $\mathbf{S}$ and $\mathbf{T}$ span a two-dimensional area with each scattered point being a tuple (S_i, T_i) for $i = 1, \cdots, N$. In order to find the mutual information, a grid is laid on the area and the obtained cells induce a probability mass function from which a mutual information value can be calculated. The grid structure is dynamically morphed and for each grid configuration the obtained MI-value is recorded. The multitude of grid structures form a surface plot of MI-values, with each MI-value representing a different grid configuration. Hence, the maximal

information coefficient is defined as the maximum MI-value that can be found on the generated surface plot.

Among the beneficial properties of MIC the authors of [49] claim:

- Generality: ability to capture a wide range of functional relationships.
- Equitability: similar strength associations get similar scores regardless of functional relationship.

Like mutual information, MIC is in principle sensitive to any kind of functional relationship between pairs of random variables. Therefore, as long as some sort of statistical dependency exists, MIC should detect that association.

Among the alleged advantages of information theoretic dependency measures like MI or MIC, the one reported in the literature quite often is the ability to identify associations in a dataset that exhibit non-linear functional relationships, in contrast to classical measures like Pearson or Spearman correlation which are only able to cope with linear functionality. For a detailed comparison the reader may take a closer look at Reshef's paper [49].

The summary of this chapter is that with respect to genomic analyses, frameworks that are based on information theory have an edge over other classical and state-of-the-art methods. Although there are some computational challenges when trying to calculate the mutual information in a mixed environmental setting consisting of both discrete and continuous random variables, the works of other authors clearly show the advantages of applying information theory to genomic analysis.

4 Introducing the MDL-principle to eQTL analysis

4.1 Analyzing genomic data with the MDL-principle

The main contribution of this thesis is the analysis tool qMAP which enables biologists and physicians to accurately extract important associations between SNPs and gene expressions in eQTL datasets. It is implemented as a Python software module and utilizes ideas and concepts of Rissanen's *minimum description length principle* (MDL) [50] in order to identify SNP-trait/gene expression associations in eQTL data.

Rissanen's approach to statistical modeling of data is grounded in the belief that a good model delivers accurate predictions which can be used to encode, i.e. to compress, the data [50]. According to Rissanen, the optimal statistical model is determined by measuring the amount of bits necessary to encode the explanatory model of the data plus the number of bits resulting from encoding the data using the explanatory model.

The mathematical tools and concepts surrounding the MDL principle were established and refined in a series of published articles [51–56].

Several authors delivered significant contributions to the MDL principle by either presenting novel theoretical concepts [57–59] or interesting applications to practical problems [60, 61].

The entire mathematical theory around the MDL concept has a long history of development and accumulated a large corpus of publications around the topic with many contributing authors. Because this thesis concentrates on genomic analysis with the MDL-principle, a few selected examples in the context of genome research will be briefly mentioned which only constitute a tiny fraction of the spectrum of MDL related research activities.

The notion of the MDL-principle has been applied to infer relations in transcriptome activity between genes in [62] by Tabus and Astola. In a branch of systems biology [27] that deals with inferring gene regulatory networks from gene expression data sets, Dougherty et al. [63] as well as Zhao et al. [64] created inference

approaches utilizing MDL. Furthermore, applications of MDL have also undertaken inroads into the clustering of gene expressions, exemplified by the article of Jörnsten and Yu [65].

Before introducing our algorithm, we will explain the basics behind the MDL-principle, give a small overview of related work, and outline the motivation behind the development of our tool.

4.1.1 MDL eQTL analysis

When looking for associations between SNPs and gene expressions in eQTL data, the effect of individual SNPs on a gene's expression value is usually very weak (see Section 2.1). Furthermore, the functional relationship between SNPs and gene expressions is unknown.

The purpose of qMAP is to alleviate some of the shortcomings current tools like PLINK [66] have.

If we wish to know more about the mechanisms of the genomic function of the DNA, we have to go beyond single SNP analysis, but start analyzing SNPs as well as gene expressions in a joint context.

Although it can be possible that there is a master regulator SNP that single-handedly controls a gene's activity, literature has shown [5, 7, 8] that several SNPs jointly influence the gene activity. On top of that, the notion emerges that networks of interacting SNPs govern transcriptome activity.

Moreover, the mRNA products of various genes have an effect on each other resulting in gene-gene interaction networks where one gene is being regulated by the expression of another gene.

From the above explanations it becomes obvious that gaining a complete understanding of all the involved genomic mechanisms that lead to a specific phenotypic trait or are responsible for disease pathogenesis is very challenging.

Our focus for the eQTL data analysis algorithm will be in identifying SNP-gene transcript associations and draw up the interaction network.

Given the above statements, a suitable analysis algorithm should have at least the following properties:

- Natively handle a mixture of discrete and continuous random variables.
- Be sensitive to and identify a broad spectrum of functional relationships.
- With increasing sample size, converge to the correct solution.

Current information theoretic approaches which have been introduced in Section 3.2 are promising candidates for an algorithm to extract useful information from an eQTL dataset.

In this thesis we introduce another approach based on the minimum description length principle and benchmark it against current methods; including the recently developed maximal information coefficient (MIC) [49], mutual information measures between discrete and continuous random variables using kernel density estimators (MI-KDE) (extended from the works of [19,48]), and the genome analysis toolkit PLINK [66].

Discovering important associations in eQTL data is in essence a learning task and several authors like Rissanen [51,57] and Grünwald [60] have shown that the MDL principle is well suited for such tasks.

As Hagenauer promoted the use of information theory in genomic analysis [18], Rissanen puts forward the idea that any underlying structure of a dataset can be leveraged to find a more compact representation of that dataset [51].

Using regularities in a dataset is usually how a compression algorithm works [39,40] in order to shrink the size of the dataset. By establishing a link between data compression and machine learning, Rissanen showed for the MDL principle that an efficient learning algorithm is also an efficient compression algorithm [50–54].

Putting this into the context of eQTL analysis, we can say that the more underlying regularities of the interaction network between SNPs and gene expressions we can discover, the more we learn about the data and the better we can find a compact description of the data, i.e. a compressed version of the eQTL dataset that requires less disk space than the original file.

As a consequence, this implies that our understanding of disease susceptibility increases as well, since we identify the associated interaction networks that make up the molecular machinery of disease. Thus, by finding all correct associations between SNPs and gene expressions, we can get an optimal compressed representation of the eQTL dataset and hence a good understanding (descriptive statistical models) of underlying disease pathogenesis mechanisms. A corollary of this statement is of course, that if an algorithm is able to automatically discover the shortest description length of an eQTL dataset, then we can say that we have learned all relevant interactions between SNPs and gene expressions in that dataset. Based on Rissanen's MDL principle we will introduce the qMAP tool that thrives towards achieving that goal.

4.1.2 Basics of MDL

Given our eQTL dataset $\mathbf{Q} = [\mathbf{S}, \mathbf{E}]$ consisting of the data vectors for both SNPs and gene expressions, let us denote the description length Λ of $\mathbf{Q}$ for the case when no code is used by Λ_{eQTL}, if another encoding scheme is used by Λ_{Code}, and the description length via MDL coding by Λ_{MDL}. Then, the essence of the MDL principle is to find a code of the eQTL data that satisfies:

$$\Lambda_{MDL} \leq \Lambda_{Code} \leq \Lambda_{eQTL}. \tag{4.1}$$

In order to find this code, we have to introduce an optimization criterion. In the MDL framework, this optimization criterion is called the *stochastic complexity* [53, 54] of the data, which is the joint description of the mathematical model that was used to encode the data plus the encoded data itself. The joint description is usually denoted in bits, which is the amount of space the mathematical formulation of the model and the encoded data would require on a storage medium.

The description length of the model used for encoding the data is also referred to as the *parametric complexity* [60]. One can imagine that mathematical models that have more parameters which can be tuned, consequently have more degrees of freedom which can be used to capture and offer explanations to a wider range of phenomena.

Speaking in mathematical terms this means that statistical models with a larger number of freely tunable parameters show the behaviour of being able to fit a great variety of data.

On the other hand, the encoded data is also referred to as the *"likelihood of the data given the parameters and the model"* by MacKay [40]. This means that if our model of the data is correct, most predictions that we make about the data will be correct. Therefore in coding terms, the description length Λ of the data will be short because the probability of our predictions to fit the observed data are high. Generally speaking, elements of a dataset that occur with a high probability need less bits during the encoding process whereas in contrast, improbable or incorrectly modeled elements require much more bits during the encoding procedure.

Combining the two above statements it is said that the stochastic complexity (SC) [60] of a dataset is:

$$\text{Stochastic Complexity} = \text{Encoded Data} + \text{Parametric Complexity}. \tag{4.2}$$

Intensive research [56, 60, 61] has led to the conclusion that the mathematical definition of stochastic complexity should be based on Shtarkov's *normalized maximum likelihood distribution* (NML) [67].

Although the NML code by Shtarkov has a phalanx of properties which are of utmost importance for proper statistical modeling of data, acknowledged and reviewed in Grünwald's book [60], for our work it is sufficient to mention only one property of the NML code which is of relevance to the eQTL analysis algorithm.

Like the case for mutual information, the feature of the NML code most interesting to us is its property to find a natural balance between model complexity and prediction accuracy. Hence, an MDL model based on the NML code of a dataset has the benefit of simultaneously giving good encoding results for the present data as well as being able to offer good coding performance on future, unseen data.

Putting the above explanation in simple terms results in the statement that the MDL principle using the NML code protects the algorithm from overfitting (see e.g. [57, 60]).

Let us build the basic NML code for our eQTL dataset $\mathbf{Q}$ by starting with the first building block, namely the *model class* M. The model class is the spectrum of statistical models which we consider for modeling the data. Since we are dealing with a great variety of potential manifestations of data distributions in eQTL data, we require a model class that is able to cope with this situation.

In the case of the random variable for SNPs S using the coding scheme described in Section 3.1, S has 3 realizations while the random variable for traits T only has two realizations. Both distributions $p(s)$ and $p(t)$ can be elegantly described by the multinomial distribution [60, 61].

Because the multinomial distribution can easily be expanded to accommodate cases where the random variables S and T have more realizations and due to its flexibility as well as the ability to reasonably approximate other distributions, we use multinomial distributions as the primary model class M in the qMAP algorithm.

The flexibility of the multinomial model class M can be seen in the size of the parameter space Θ which holds the parameters θ that describe the realization probabilities of the various random variables in our dataset:

$$M = \{P(\cdot|\theta) : \theta \in \Theta\}. \tag{4.3}$$

When employing one of the available models in the model class M to encode the dataset $\mathbf{Q}$, the likelihood of the parameter(s) θ have to be estimated from the available data by means of the *maximum likelihood* (ML) estimator $\hat{\theta}(\mathbf{Q}, M)$. A good introduction to probabilities and inference procedures is given in MacKay [40]. The ML-estimator is the estimator that finds the parameter θ which maximizes the likelihood:

$$\hat{\theta}(\mathbf{Q}, M) = \text{argmax}_{\theta \in \Theta} P(\mathbf{Q}|\theta). \tag{4.4}$$

As the name already implies, the normalized maximum likelihood distribution [60, 61, 67] is the ordinary likelihood distribution divided by a normalizing term.

The normalizing term $R(M)$ is defined according to Shtarkov [67] as the sum of all potential likelihood distributions that can be generated with the current model class M for all possible realizations $\tilde{\mathbf{Q}}$ of the original data $\mathbf{Q}$:

$$R(M) = \sum_{\tilde{\mathbf{Q}}} P(\tilde{\mathbf{Q}}|\hat{\theta}(\tilde{\mathbf{Q}}, M)). \tag{4.5}$$

Calculating the above sum gives us the number of bits necessary to mathematically describe the model class M. Obviously, calculating this sum in practice is more than challenging.

Luckily due to Kontkanen and Myllymäki [60, 61, 68, 69], an elegant and efficient computation method for exactly calculating the parametric complexity $R(M)$ for the multinomial model class M exists. As an integral part of qMAP, Kontkanen and Myllymäki's algorithm will be explained in Section 4.1.3.

After having confirmed that the normalization term $R(M)$ can be computed in practice, we give Shtarkov's definition of the NML-distribution as

$$P_{NML}(\mathbf{Q}|M) = \frac{P(\mathbf{Q}|\hat{\theta}(\mathbf{Q}, M))}{R(M)}. \tag{4.6}$$

The stochastic complexity SC of the data, which is more precisely the NML-code, which in theory produces the shortest description length of the eQTL data $\mathbf{Q}$, is simply the negative logarithm of the NML-distribution:

$$\begin{aligned} SC(\mathbf{Q}|M) &= -\log\, P_{NML}(\mathbf{Q}|M) \\ &= -\log\, P(\mathbf{Q}|\hat{\theta}(\mathbf{Q}, M)) + \log\, R(M), \end{aligned} \tag{4.7}$$

4.1.3 Kontkanen and Myllymäki (KM) calculation method for NML codes

Our approach for estimating the shortest description length of a SNP-gene expression pair is going to be based on finding an NML-code for the eQTL data.

Akin to the method used by MIC to find the maximum information coefficient between two pairs of random variables by optimizing a grid partitioning that induces a joint distribution from which the mutual information can be derived, our approach utilizes Kontkanen and Myllymäki's algorithm [70] in order to obtain the normalized maximum likelihood code of a SNP-gene pair association via dynamic grid optimization using the stochastic complexity from the MDL-principle as the optimality criterion.

The methodology for finding such an optimal MDL-grid has been invented by Kontkanen and Myllymäki [70] in the context of obtaining MDL-optimal histogram densities from continuous probability distributions. They demonstrated in their paper that depending on the available sample size, the MDL-histogram density estimator automatically selects the total number of histogram bins and a variable bin size for each bin that optimizes the stochastic complexity of the data [70].

Due to the importance of the dynamic programming algorithm that was presented in their paper for solving the problem of efficiently calculating the NML code, we will outline Kontkanen and Myllymäki's grid optimization algorithm for the $1D$-case, i.e. obtaining the NML code for the distribution $p(e)$ of one gene expression random variable E. To honor Kontkanen and Myllymäki's contribution we will refer to their optimization method as the KM-method. Kameya in his time-series quantization paper [71], which builds upon Kontkanen and Myllymäki's work, also refers to their algorithm as the KM-method.

When we want to obtain the MDL-optimal code for one gene expression value, the KM-method for the one dimensional case works as follows.

We have gathered expression values for one gene from N patients in our eQTL dataset $\mathbf{Q}$. They are recorded with finite precision ε, let us say 6 decimal points after the comma. Therefore, our data vector for which we would like to obtain the MDL-optimal shortest description is $\mathbf{E} = \mathbf{e}^N = [e_1, \cdots, e_N]$, containing each patient's measured gene expression values e.

When the grid is laid on the data, the grid's boundaries have to be set. This can be done by looking at the maximum and minimum expression values e_{max} and e_{min} in our dataset $\mathbf{E}$. Furthermore, the fact that the data are stored with finite precision ε is used by the KM-method as a technical assistance construct to not only derive the boundaries of the overlay grid but also the incision points for the grid optimization procedure.

If we assume that the data have been recorded with precision ε then all gene expression values e in our dataset $\mathbf{e}^N$ will be draws from the set $\mathscr{E}$ defined as

$$\mathscr{E} = e_{min} + \delta \cdot \varepsilon \tag{4.8}$$

for δ in range from 0 to $\frac{e_{max}-e_{min}}{\varepsilon}$. Hence, the boundaries of the grid are defined to be $e_{min} - \frac{\varepsilon}{2}$ as the lower and $e_{max} + \frac{\varepsilon}{2}$ as the upper grid boundary.

Since the KM-method will be used to select both the optimal grid granularity and the location of grid points, the set of putative grid points has to be defined.

Note that we use a slightly different terminology than in the original KM-paper [70], where the *grid points* are referred to as *cut points*. The set of putative grid points Υ contains the upper and the lower boundary of the grid as well as all

points that fall in between neighbouring values of a sorted list of expression values e contained in $\mathbf{e}^N$. Thus, a grid point ξ separates two consecutive e-values in the sorted list, which leads to the mathematical definition of the grid point set Υ, i.e.

$$\Upsilon = e_{min} + \frac{\varepsilon}{2} + \delta \cdot \varepsilon \text{ for } \delta \in [0, \cdots, \frac{e_{max} - e_{min}}{\varepsilon} - 1]. \tag{4.9}$$

The grid points ξ_κ belonging to the set Υ give us a grid layout consisting of at most K grid intervals which are stored as an increasing sequence in the vector $\mathbf{\Xi} = [\xi_1, \cdots, \xi_{K-1}]$.

Through optimization of the NML-code the KM-method is going to select an MDL-optimal grid partitioning, i.e. that it delivers the MDL-optimal number of grid intervals K and also the grid points ξ_κ that span the grid Ξ.

Since the grid borders are static, only the grid points ξ_κ falling into the interval $[e_{min}, e_{max}]$ are considered in the optimization procedure, i.e. $\xi_1 = e_{min} + \frac{\varepsilon}{2}$ and $\xi_{K-1} = e_{max} - \frac{\varepsilon}{2}$.

Recall that the family of multinomial distributions is used as the model class M and that a grid Ξ consisting of K intervals can have parameters θ_κ to model the induced data distribution ϕ_Ξ of the grid Ξ. Those parameters θ_κ belong to the parameter set Θ satisfying:

$$\Theta = (\theta_1, \cdots, \theta_K) : \theta_\kappa \geq 0, \quad \sum_{\kappa=1}^{K} \theta_\kappa = 1. \tag{4.10}$$

A grid Ξ with the lower boundary $\xi_0 = e_{min} - \frac{\varepsilon}{2}$, the upper boundary $\xi_K = e_{max} + \frac{\varepsilon}{2}$ and the grid points $\mathbf{\Xi} = [\xi_1, \cdots, \xi_{K-1}]$ gives rise to the multinomial distribution ϕ_Ξ of gene expressions defined via

$$\phi_\Xi(e|\theta, \Xi) = \frac{\varepsilon\theta_\kappa}{\lambda_\kappa} \tag{4.11}$$

with $\lambda_\kappa = \xi_\kappa - \xi_{\kappa-1}$ giving us the lengths of the grid intervals $\kappa = 1, \cdots, K$. Via this basic definition we can see that the grid partitions the data, in our case the gene expression values, so that the grid induced density ϕ_Ξ gives us the probability of the data point e falling into the grid interval $]\xi_{\kappa-1}, \xi_\kappa]$ of length λ_κ.

For the entire data sample $\mathbf{e}^N$, the grid produces the probability distribution

$$\phi_\Xi(\mathbf{e}^N|\theta, \Xi) = \prod_{\kappa=1}^{K} \left(\frac{\varepsilon\theta_\kappa}{\lambda_\kappa}\right)^{h_\kappa}. \tag{4.12}$$

In accordance to Kontkanen and Myllymäki [70], the NML-code for the grid density ϕ_Ξ is derived in the following way. As the parameters θ_κ denote the

probability of a value e falling into the grid interval denoted by κ, the maximum likelihood estimate $\hat{\theta}_{ML}$ of those parameters is

$$\hat{\theta}_\kappa = \frac{h_\kappa}{N}, \tag{4.13}$$

which are just the relative frequencies. Replacing the parameters θ_κ in Equation 4.12 with relative frequency counts, we arrive at an expression for the maximum likelihood estimate of the grid density $\hat{\phi}_\Xi$ that can be computed straightforward from the available data:

$$\hat{\phi}_\Xi(\mathbf{e}^N|\theta,\Xi) = \prod_{\kappa=1}^{K}\left(\frac{\varepsilon h_\kappa}{\lambda_\kappa N}\right)^{h_\kappa}. \tag{4.14}$$

After having obtained an expression for the maximum likelihood estimate of the grid intensity, we proceed in calculating the normalizing constant $R(M)$ in order to get the NML-distribution induced by the grid.

For this the sum of Equation 4.5 has to be evaluated for every thinkable dataset realization $\tilde{\mathbf{Q}}$ of size N. Evaluating Equation 4.5 really is a herculean task, but since we are using multinomial distributions to model eQTL data, it is possible to apply Kontkanen and Myllymäki's recursive algorithm to obtain not only an exact solution to the computation of the parametric complexity, but on top of that achieve it in a speedy manner [61, 68–70].

The straightforward but computationally intractable approach for calculating the normalizing constant of the NML distribution for a grid Ξ with K intervals and a total of N data points

$$\begin{aligned} R(N,K,M) &= \sum_{\tilde{\mathbf{q}}^N\in\tilde{\mathscr{Q}}^N}\prod_{\kappa=1}^{K}\left(\frac{\varepsilon h_\kappa}{\lambda_\kappa N}\right)^{h_\kappa} \\ &= \sum_{h_1+\cdots+h_K=N}\frac{N!}{h_1!\cdots h_K!}\prod_{\kappa=1}^{K}\left(\frac{\lambda_\kappa}{\varepsilon}\right)^{h_\kappa}\prod_{\kappa=1}^{K}\left(\frac{\varepsilon\cdot h_\kappa}{\lambda_\kappa N}\right)^{h_\kappa} \\ &= \sum_{h_1+\cdots+h_K=N}\frac{N!}{h_1!\cdots h_K!}\prod_{\kappa=1}^{K}\left(\frac{h_\kappa}{N}\right)^{h_\kappa} \end{aligned} \tag{4.15}$$

is reduced to the recursive formulation [61, 68–70]

$$R(N,K,M) = R(N,K-1,M) + \frac{N}{K-2}R(N,K-2,M). \tag{4.16}$$

The initialization parameters for the recursion are set according to [61, 68–70] as

$$R(N, 1, M) = 1 \tag{4.17}$$

and

$$R(N, 2, M) = \sum_{h_1+h_2=N} \frac{N!}{h_1!h_2!} \left(\frac{h_1}{N}\right)^{h_1} \left(\frac{h_2}{N}\right)^{h_2}. \tag{4.18}$$

For grids Ξ that have $K > 2$ intervals, the parametric complexity is obtained by calculating the initial values of the recursion, plugging them into Equation 4.16 and then running the recursion step $K - 2$ times to arrive at the final solution for the parametric complexity $R(N, K, M)$.

With both a maximum likelihood expression for the grid induced density and an exact solution for the normalization constant, the NML-code can be computed. Thus, the MDL grid optimization criterion given by the stochastic complexity SC is formulated as:

$$\begin{aligned} SC(\mathbf{Q}|M) &= -\log \frac{\prod_{\kappa=1}^{K} \left(\frac{\varepsilon h_\kappa}{\lambda_\kappa N}\right)^{h_\kappa}}{R(N, K, M)} \\ &= \sum_{\kappa=1}^{K} -h_\kappa \left(\log\left(\varepsilon h_\kappa\right) - \log\left(\lambda_\kappa \cdot N\right)\right) + \log R(N, K, M). \end{aligned} \tag{4.19}$$

4.1.4 The KM-dynamic programming algorithm for optimizing the stochastic complexity of a grid Ξ

Solving the MDL grid partitioning problem by minimizing the stochastic complexity of the data requires a sophisticated algorithmic approach for making the calculation of an exact solution to the NML-code of Equation 4.19 computationally feasible.

An elegant solution for this problem has been described by Kontkanen and Myllymäki in [70]. In the case of estimating a binned histogram version of a one dimensional probability density function, Kontkanen and Myllymäki presented an algorithm in [70] which automatically selects both the number of bins and their elastic boundaries in a data depended fashion using the MDL stochastic complexity as the optimization criterion.

Their algorithm can be adapted to also calculate stochastic complexities of the data for various grid configurations. Since exact solutions are attainable, it is possible to acquire the MDL-optimal grid configuration via minimization of the NML-code with Kontkanen and Myllymäki's algorithm.

The MDL-optimal solution gives us the grid partitioning that defines the stochastic complexity of the data and that will allow us later to discover associations between SNP and gene expressions by declaring those associations as relevant which yield short description lengths of the eQTL data. As has already been stated in Section 4.1.1, in the MDL framework, good data compression can be interpreted as the same thing as finding good explanations of the eQTL data, i.e. SNPs which "explain" the expression profile of a gene quite well.

SNPs that achieve this are utilized by the MDL-principle to further reduce the stochastic complexity of the eQTL data by yielding shorter NML-code lengths, therefore revealing to us the underlying structure of eQTL data in the form of interaction networks between SNPs and gene expressions.

Each grid configuration yields different scores for the stochastic complexity of the data, which are expressed through the code length in bits via the eQTL data's NML code of Equation 4.19. Since the software implementation adhered to the notation of Kontkanen and Myllymäki's paper [70], the same symbol for the grid optimization score will be used, namely B. According to the KM-method, the score which has to be optimized in order to obtain the NML code of the data is given by

$$\begin{aligned} B(\mathbf{e}^N|Z,K,\Xi) &= SC(\mathbf{Q}|\Xi) + \log\ \left(\frac{Z}{K-1}\right) \\ &= \sum_{\kappa=1}^{K} -h_\kappa \left(\log\ (\varepsilon h_\kappa) - \log\ (\lambda_\kappa N)\right) \\ &\quad + \log\ R(N,K,M) + \log\ \left(\frac{Z}{K-1}\right). \end{aligned} \tag{4.20}$$

Let us give some clarifying explanations regarding the above formula. The parameter K specifies the number of grid intervals. The chosen partitioning points ξ_κ are contained in the vector $\boldsymbol{\Xi}$. Because the optimization goal is to find the optimal placement of the grid points ξ_κ and their optimal number K, we have to compare all grid configurations that can possibly arise via combinations of putative candidate grid points from the manifold set indicated by Kontkanen and Myllymäki as

$$\hat{\Upsilon} = \left(\{e_\tau - \frac{\varepsilon}{2} : e_\tau \in \mathbf{e}^N\} \cup \{e_\tau + \frac{\varepsilon}{2} : e_\tau \in \mathbf{e}^N\}\right) \ni \{e_{min} - \frac{\varepsilon}{2}, e_{max} + \frac{\varepsilon}{2}\} \tag{4.21}$$

with $e_{min} - \frac{\varepsilon}{2}$ and $e_{max} + \frac{\varepsilon}{2}$ being the grid's borders ξ_0 and ξ_K, which are excluded from the candidate grid point set because they are present in all potential grid configurations and need not be selected for.

Thus, the optimization procedure has to retrieve a vector $\boldsymbol{\Xi} \in \hat{\Upsilon}$ that minimizes the NML code of the data in the form of the MDL-score $B(\mathbf{e}^N|Z, K, \Xi)$ with Z being the size of the candidate grid point set $\hat{\Upsilon}$, i.e. $Z = |\hat{\Upsilon}|$.

The term log $(\frac{Z}{K-1})$ encodes the grid point set and assists in forming the MDL-score because it is argued in [60] by Grünwald and [70] by Kontkanen and Myllymäki, that *"In practical model selection tasks, however, the stochastic complexity criterion itself may not be sufficient"* [70]. Therefore, several authors have concluded that the model index should also be somehow encoded [60, 70]. The authors Kontkanen and Myllymäki show in [70] that for the grid optimization problem this encoding reduces to the term log $(\frac{Z}{K-1})$.

The elegant dynamic programming solution of [70] for calculating the MDL-scores B for various grid configurations starts by first sorting the candidate grid point set $\hat{\Upsilon}$ in ascending order:

$$\hat{\Upsilon} = \{\hat{\xi}_1, \cdots \hat{\xi}_Z\} \quad \text{with} \quad \hat{\xi}_1 < \hat{\xi}_2 < \cdots < \hat{\xi}_Z. \tag{4.22}$$

The gird point $\hat{\xi}_{Z+1}$ is defined as the upper boundary of the grid, namely $\hat{\xi}_{Z+1} = \xi_K = e_{max} + \frac{\varepsilon}{2}$. In oder to calculate a final solution for the score $B(\mathbf{e}^N|Z, K, \Xi)$, the KM-method makes an ingenious step by restricting the data set $\mathbf{e}^N$ to a smaller subset $\mathbf{e}^{N_\zeta}$ and calculating the score for that subset first. Then, the calculated score for that smaller subset is used as a basis to recursively calculate MDL-scores for increasing data ranges.

The reduced data set $\mathbf{e}^{N_\zeta}$ consists of the gene expression values e_τ falling into the interval outlined by $[e_{min}, \hat{\xi}_\zeta]$. For an increasing sequence of ζ from 1 to $Z+1$ the data subset is represented by the vector $\mathbf{e}^{N_\zeta} = [e_1, \cdots, e_{N_\zeta}]$ and the MDL-score which optimizes the stochastic complexity for that data subset is denoted by

$$\hat{B}_{K,\zeta} = \min_{\boldsymbol{\Xi} \in \hat{\Upsilon}} B(\mathbf{e}^{N_\zeta}|Z, K, \Xi) \tag{4.23}$$

The KM-method's elegance is expressed in the observation [70] that the final MDL-score of a grid configuration for the data can be obtained in a successive manner. Namely, if one obtained the MDL-optimal score $\hat{B}_{K,\zeta}$ for a grid configuration whose number of intervals is fixed to K, then the final solution will be the MDL-score $\hat{B}_{K,Z+1}$ because it encompasses the entire data range of $[e_{min}, \hat{\xi}_{Z+1}]$ of gene expression values in $\mathbf{e}^N$.

Therefore, the final solution can be sequentially build up from previous partial solutions. This means that the MDL-score for a grid configuration Ξ consisting of K intervals can be obtained in a successive manner by making use of the previously MDL-score for a grid consisting of $K-1$ intervals.

A grid configuration Ξ is defined by its number of intervals K and the separating grid points $\hat{\xi}_\zeta$ resulting in the characteristic grid point vector $\mathbf{\Xi} = (\hat{\xi}_{\zeta_1}, \cdots \hat{\xi}_{\zeta_{K-1}})$. This grid configuration imposes a limit on the data set given by $[e_{min}, \hat{\xi}_{\zeta_K}]$.

The main point of the KM-method is that the MDL-score $B(\mathbf{e}^{N_{\zeta_K}}|Z, K, \Xi)$ for $\hat{\xi}_{\zeta_K} > \hat{\xi}_{\zeta_{K-1}}$ is obtained by building upon the $K-1$ MDL-score $B(\mathbf{e}^{N_{\zeta_{K-1}}}|Z, K-1, \Xi')$ with the grid points $\mathbf{\Xi}' = (\hat{\xi}'_{\zeta_1}, \cdots \hat{\xi}'_{\zeta_{K-2}})$ via the relationship:

$$\begin{aligned} B(\mathbf{e}^{N_{\zeta_K}}|Z, K, \Xi) = B(\mathbf{e}^{N_{\zeta_{K-1}}}|Z, K-1, \Xi') \\ - \left(N_{\zeta_K} - N_{\zeta_{K-1}}\right) \cdot \left(\log\ \left(\varepsilon\left(N_{\zeta_K} - N_{\zeta_{K-1}}\right)\right)\right) \\ - \log\ \left((\hat{\xi}_{\zeta_K} - \hat{\xi}_{\zeta_{K-1}})N\right) \\ + \log \frac{R_{h_K}^{N_{\zeta_K}}}{R_{h_{K-1}}^{N_{\zeta_{K-1}}}} \\ + \log \frac{Z-K+2}{K-1} \end{aligned} \tag{4.24}$$

Consequently, the KM dynamic programming recursion equation for calculating the MDL-optimal NML-code of the data for the currently used grid configuration becomes

$$\begin{aligned} \hat{B}_{K,\zeta} = \min_{\zeta'}\{\hat{B}_{K-1,\zeta'} - (N_\zeta - N_{\zeta'}) \cdot \log\ (\varepsilon(N_\zeta - N_{\zeta'})) \\ - \log\ \left((\hat{\xi}_\zeta - \hat{\xi}_{\zeta'}) \cdot N\right) \\ + \log \frac{R_{h_K}^{N_\zeta}}{R_{h_{K-1}}^{N_{\zeta'}}} \\ + \log \frac{Z-K+2}{K-1}\} \end{aligned} \tag{4.25}$$

with the index ζ' ranging from $K-1$ to ζ.

The recursion's starting conditions are initialized with the following procedure. First, the MDL-score is calculated in a ζ dependent fashion for a grid whose boundaries are defined by ξ_{ζ_0} and ξ_{ζ_K}. Data points of the subset of gene expressions $\mathbf{e}^{N_\zeta}$ fall in ascending order into the above defined grid interval leading to the MDL-scores

$$\hat{B}_{1,\zeta} = -N_\zeta \cdot \left(\log\ (\varepsilon \cdot N_\zeta) - \log\ \left(\left(\hat{\xi}_\zeta - (e_{min} - \frac{\epsilon}{2})\right) \cdot N\right)\right). \tag{4.26}$$

By increasing the index ζ to the maximum range $Z+1$ that includes all data points of the dataset, the first partial solutions of the dynamic programming solution

are obtained. The implementation details will be explained in subsequent Section 5 and this will make understanding of the calculation procedure much easier. It is basically a double for-loop that is applied to solve the KM-method's dynamic programming approach. The outer for-loop takes care of the index ζ running from K to $Z+1$ while the inner for-loop increases the index ζ' in the range from $K-1$ to $\zeta-1$.

The only parameter the user has to set is the maximum number of allowed grid intervals K_{max} to provide the recursion with a break point if taking too long to determine the MDL-optimal grid granularity. Each recursion step records the MDL-scores $\hat{B}_K$ for different grid configurations Ξ in a table and the minimum score among them is chosen to be the solution of the NML-code.

Among all potential grid configurations the one with the shortest description length, i.e. best MDL-score $\hat{B}$, is selected as the final grid Ξ upon which the NML-code for the eQTL dataset is determined.

Through a technique called *back-tracking* [70], the grid shape, i.e. the placement of intersections ξ between consecutive grid intervals, can also be conveniently recovered from the above dynamic programming recursion. The details will be explained in the implementation Section 5.

4.2 Extending the KM-method for eQTL applicability

As we have seen in the previous Sections 4.1.3 and 4.1.4, the KM-algorithm is able to obtain the NML-code for one dimension of an eQTL dataset, namely for the data $\mathbf{e}^N$ of continuous gene expression random variables E.

In order for the KM-algorithm to work in a heterogeneous setting consisting of both discrete random variables for SNP genotypes S and continuous gene expression random variables E, which are the basic building blocks of any eQTL dataset, we have to extend the current KM-algorithm to make it applicable to those kind of data. Based on our extensions, which will be presented in this section, we will later proceed to create an algorithm that identifies important associations between SNPs and gene expressions via minimization of the stochastic complexity of the eQTL data.

In the area of time series analysis, Kameya [71] created a dynamic two-dimensional time series discretization approach based on the expansion of the KM-method into two dimensions for continuous data.

The achievement of Kameya's approach was to show that by simultaneously optimizing both axis of a $2D$-grid, the grid was able to dynamically adapt to the time series data and give a simplified, compartmentalized representation of the time

series, even in the presence of noise, that could be used to highlight important characteristics of the data set [71].

Compared to Kameya's algorithm of [71], our goal is different in the sense that we want to extract useful information from eQTL data by discovering associations between SNP-gene pairs. In contrast to Kameya's method, our algorithm needs to function in a mixed environment that consists of both discrete and continuous random variables. As a consequence, we cannot directly apply the KM-algorithm but have to extend it appropriately.

We utilize the knowledge that for discrete SNP data the grid configuration is determined by Hagenauer's haplotype encoding procedure [18] and statically set accordingly. This property has to be incorporated into the KM-method to make it work in a mixed setting.

Furthermore, we also present an efficient approach for calculating the NML-distribution after the SNP encoding process has been combined with the KM-algorithm. This modification enables a speed-up in calculation time when compared to [71] because only the grid axis responsible for compartmentalizing gene expression data needs to be optimized.

A side note on this topic is, that by determining the grid configuration for SNPs in the haplotype coding process, our extension of the KM-algorithm can be augmented to deal not only with 3-dimensional grids but also N-dimensional grids that would allow the detection of epistatic and epigenetic effects by testing joint ensembles of SNPs for important associations with gene expressions. Through this approach it becomes possible to detect interaction contexts of SNPs and their influence on genes. Our current implementation's main focus is the discovery of association pairs between single SNPs and gene expressions. Detection of epigenetic and multi-factorial SNP effects remains future work.

4.2.1 The mKM-algorithm for associating quantitative traits with discrete genotypes

Given our eQTL dataset $\mathbf{Q}$ that contains G gene expression vectors $\mathbf{e_g^N}$ and SNPs $\mathbf{s_l^N}$ with L genotyped genomic locations for N patients, our goal is to obtain the NML-code for a SNP-gene pair from which the association strength can be inferred.

The first step is to lay a $2D$-grid over the data. Then, our extended version of the KM-algorithm is applied to determine the optimal grid configuration for the gene expressions under the assumption that the grid configuration for the SNP axis is given via the haplotype coding step outlined in Section 3.1.1.

We name the method mKM-algorithm, with the letter m implying that it is an extension of the KM-algorithm for the mixed setting of continuous gene expression

and discrete SNP random variable pairs.

As is the case for the $1D$-KM-algorithm, the mKM-algorithm defines the grid boundaries for the gene expression axis in the same way as outlined in Section 4.1.3, with the grid point set being $\hat{\Upsilon}$.

What changes compared to the $1D$-KM is the definition of the induced grid distribution $\dot{\phi}_{\Xi}$. Let us assume that the haplotype coding mapped the genotyped SNP to 3 symbols. Therefore, the grid axis for the SNP vector has 3 intervals $\tilde{K}$. With the grid configuration $\tilde{K}$ set for the SNP axis, the mKM-algorithm only needs to optimize the remaining axis for the gene expression values.

For both dimensions we use the multinomial distribution of Section 4.1.2 as the model class M. The parameter $\theta_{\kappa,\tilde{\kappa}}$ is now defined to be the probability of a SNP-transcript value pair (s_ν, e_τ) to fall in a grid area outlined by its border $]\xi_{\kappa-1}, \xi_\kappa]$ on the gene expression axis ξ_E and $]\tilde{\xi}_{\tilde{\kappa}-1}, \tilde{\xi}_{\tilde{\kappa}}]$ on the SNP axis $\tilde{\xi}_S$.

Assuming the SNP encoding scheme of Section 3.1, with a mapping into the alphabet $\mathcal{A}_S = \{0, 1, 2\}$, the 3 intervals $\tilde{\kappa} = 0, 1, 2$ are defined by the borders

$$\begin{aligned} \tilde{\xi}_0 &=]-0.5, 0.5] \\ \tilde{\xi}_1 &=]0.5, 1.5] \\ \tilde{\xi}_2 &=]1.5, 2.5]. \end{aligned} \tag{4.27}$$

The new parameter set $\dot{\Theta}$ induced by the two dimensional grid becomes

$$\dot{\Theta} = (\theta_{1,0}, \cdots, \theta_{K,\tilde{K}}) : \theta_{\kappa,\tilde{\kappa}} \geq 0, \quad \sum_{\kappa=1}^{K} \sum_{\tilde{\kappa}=0}^{2} \theta_{\kappa,\tilde{\kappa}} = 1. \tag{4.28}$$

Let the data for which we wish to determine the association strength be the two vectors containing the encoded SNP values and gene expressions $\mathbf{q}^N = [\mathbf{e}^N, \mathbf{s}^N]$. Then, the distribution $\dot{\phi}_{\Xi}$ induced by the grid for this $2D$-dataset is given by

$$\dot{\phi}_{\Xi}\left([\mathbf{e}^N, \mathbf{s}^N] | \dot{\theta}, \Xi\right) = \prod_{\kappa=1}^{K} \prod_{\tilde{\kappa}=0}^{2} \left(\frac{\varepsilon \dot{\theta}_{\kappa\tilde{\kappa}}}{\lambda_\kappa}\right)^{h_{\kappa\tilde{\kappa}}} \tag{4.29}$$

with λ_κ being the length of the grid interval on the axis for the gene expressions ξ_E and $h_{\kappa\tilde{\kappa}}$ the number of data points (s_ν, e_τ) falling into the lot denoted by the index $\tilde{\kappa}$ for SNP data points and the interval $]\xi_{\kappa-1}, \xi_\kappa]$ for gene expression data points.

Because the interval lengths $\lambda_{\tilde{\kappa}}$ are set to 1 by definition for the grid axis $\tilde{\xi}_S$ for the SNP data, the term does not appear in Equation 4.29. From the above considerations follows that in order to obtain the maximum likelihood estimate

$\dot{\hat{\theta}}_{ML}$ for the $2D$-grid parameters $\dot{\theta}_{\kappa\tilde{\kappa}}$ we simply have to divide the number of data points falling into the grid location indicated via coordinates $(\kappa, \tilde{\kappa})$ by the available sample size N, i.e.

$$\dot{\hat{\theta}}_{\kappa\tilde{\kappa}} = \frac{h_{\kappa\tilde{\kappa}}}{N}. \tag{4.30}$$

This leads us to the maximum likelihood equation for the $2D$ grid distribution

$$\dot{\phi_{\Xi}}\left([\mathbf{e}^N, \mathbf{s}^N]|\dot{\theta}, \Xi\right) = \prod_{\kappa=1}^{K}\prod_{\tilde{\kappa}=0}^{2}\left(\frac{\varepsilon h_{\kappa\tilde{\kappa}}}{\lambda_\kappa N}\right)^{h_{\kappa\tilde{\kappa}}}. \tag{4.31}$$

The greatest difference in the definition of the stochastic complexity for a 1-dimensional and a 2-dimensional grid is the normalizing term R, since in the $2D$-case it has to account for a mixture of multinomial models.

In [61] on pp. $342 - 344$ Kontkanen et al. derived a recursive formula for calculating the parametric complexity for multinomial mixtures that yields an exact solution without having to resort to approximations or time consuming brute-force computation approaches.

Similar in concept to the formulation for the one dimensional multinomial model class given in Equation 4.16, the normalization term R for a $2D$-grid with intervals numbering K and $\tilde{K}$, i.e. the parametric complexity $R(N, K, \tilde{K}, M)$, is given as

$$R(N, K, \tilde{K}, M_2) = \sum_{r_1+r_2=N}\frac{N!}{r_1!r_2!}\left(\frac{r_1}{N}\right)^{r_1}\left(\frac{r_2}{N}\right)^{r_2}\cdot R(r_1, \tilde{K}, M_2)\cdot R(r_2, K-\tilde{K}, M_2) \tag{4.32}$$

which combines the easily obtainable results for the parametric complexity for the $1D$-case to form the result for the $2D$-case. M_2 indicates the case where a mixture of 2 multinomial distributions is used in order to simultaneously describe a SNP-gene pair. In contrast, M indicates the multinomial model class known from Section 4.1.3 for the one dimensional case which is used e.g. to describe a vector of gene expressions in the eQTL dataset.

Since the encoding procedure for the SNP values already gives us the number of intervals $\tilde{K} = 3$ for the SNP-axis $\tilde{\xi}_S$ of the grid as well as the multinomial model which will be used to obtain the NML-code for the SNP-data, in the case that the expression-axis ξ_E of the grid contains $K = 1$ intervals, the normalizing parameter $R(N, 1, 3, M_2)$ is reduced to $R(N, 3, M)$. By applying the calculation procedure for the 1-dimensional case of the parametric complexity from Section 4.1.3, $R(N, 3, M)$ is obtained in a straightforward manner.

In case where no data points are available, $R(N, K, \tilde{K}, M_2)$ is defined according to [61] to be:

$$R(0, K, \tilde{K}, M_2) = 1. \tag{4.33}$$

The above conditions can be used to initialize the recursion and obtain the parametric complexity for any $2D$-grid configuration Ξ defined via K. Thus, we are able to calculate the NML-code for a grid that captures and describes the statistical relationship between genotypes and gene expressions via its stochastic complexity by

$$\begin{aligned} SC\left([\mathbf{e}^N, \mathbf{s}^N]|M_2\right) &= -\log \frac{\dot{\phi}_{\Xi}([\mathbf{e}^N, \mathbf{s}^N]|\dot{\theta}, \Xi)}{R(N, K, \tilde{K}, M_2)} \\ &= -\log \prod_{\kappa=1}^{K} \prod_{\tilde{\kappa}=0}^{2} \left(\frac{\varepsilon h_{\kappa\tilde{\kappa}}}{\lambda_\kappa N}\right)^{h_{\kappa\tilde{\kappa}}} \\ &\quad + R(N, K, \tilde{K}, M_2). \end{aligned} \tag{4.34}$$

The score that needs to be optimized for the 2-dimensional grid is an extended version of Section 4.1.3 which includes the grid labels of the SNP-axis, that were already obtained by the SNP encoding procedure. The MDL-optimal partitioning of the expression-axis dependent on the SNP-axis is acquired by

$$\begin{aligned} B\left([\mathbf{e}^N, \mathbf{s}^N]|Z, K, \tilde{K}, \Xi\right) &= SC\left([\mathbf{e}^N, \mathbf{s}^N]|M_2\right) + \log \left(\frac{Z}{K-1}\right) \\ &= -\log \frac{\dot{\phi}_{\Xi}([\mathbf{e}^N, \mathbf{s}^N]|\dot{\theta}, \Xi)}{R(N, K, \tilde{K}, M_2)} + \log \left(\frac{Z}{K-1}\right) \\ &= -\log \prod_{\kappa=1}^{K} \prod_{\tilde{\kappa}=0}^{2} \left(\frac{\varepsilon h_{\kappa\tilde{\kappa}}}{\lambda_\kappa N}\right)^{h_{\kappa\tilde{\kappa}}} \\ &\quad + \log R(N, K, \tilde{K}, M_2) + \log \left(\frac{Z}{K-1}\right). \end{aligned} \tag{4.35}$$

What changes when compared to the $1D$-KM-algorithm is the way of counting the frequencies $h_{\kappa\tilde{\kappa}}$. With $\tilde{K} = 3$ each interval $\tilde{\kappa}$ of the grid's SNP-axis $\tilde{\xi}_S$ can also be interpreted as a label representing the corresponding encoded SNP. By dividing the expression axis ξ_E into K intervals, $h_{\kappa\tilde{\kappa}}$ is equal to the number of gene expression data points falling into $]\xi_{\kappa-1}, \xi_\kappa]$ whose corresponding SNP genotype has label $\tilde{\kappa}$.

To start the dynamic programming recursion for obtaining the MDL-score $\hat{B}$ we first calculate the score for the grid if the expression-axis ξ_E just has $K = 1$ intervals

$$\hat{B}_{1,\tilde{K},\zeta} = \sum_{\tilde{\kappa}=1}^{\tilde{K}} -N_{\zeta\tilde{\kappa}} \cdot \Big(\log\ (\varepsilon N_{\zeta\tilde{\kappa}}) - \log\ \Big(\Big(\hat{\xi}_{\zeta} - (e_{min} - \frac{\varepsilon}{2})\Big) \cdot N\Big)\Big), \qquad (4.36)$$

with $\tilde{K} = 3$ and $N_{\zeta\tilde{\kappa}}$ giving the frequency counts for each SNP label $\tilde{\kappa} \in \tilde{K}$ of data point pairs $[s, e] \in [\mathbf{e}^N, \mathbf{s}^N]$ for the data range restricted to ζ.

Because we only need to optimize one axis of the grid, namely the gene expression axis ξ_E, the optimization procedure for any κ number of intervals proceeds as

$$\hat{B}_{\kappa,\tilde{K},\zeta} = \hat{B}_{\kappa-1,\tilde{K},\zeta'} - \sum_{\tilde{\kappa}=1}^{\tilde{K}} h_{\zeta\tilde{\kappa}} \cdot \log\ \frac{\varepsilon h_{\zeta\tilde{\kappa}}}{\lambda_{\zeta\tilde{\kappa}} N} + \log\ \frac{R(N_\zeta, K, \tilde{K}, M_2)}{R(N_\zeta, K-1, \tilde{K}, M_2)} + \log\ \frac{Z-K+2}{K-1} \qquad (4.37)$$

with an enclosing double loop $\zeta = K \cdots Z+1$ and $\zeta' = K-1 \cdots \zeta-1$ which calculates the final solution in a consecutive way.

The mKM-algorithm enables us to compute the stochastic complexities for ensembles of SNP-gene pairs. With the mKM-algorithm in place, we proceed to explain our MDL-analysis algorithm for eQTL datasets.

4.2.2 MDL-score for assessing association strengths between genotype and transcriptome in eQTL data via the mKM-method

The mKM-algorithm is the core function for obtaining an MDL-optimal grid layout that minimizes the stochastic complexity of a SNP-gene expression data pair $[\mathbf{e}^N, \mathbf{s}^N]$ originating from the eQTL data $\mathbf{Q}$.

For finding important associations, we derive a score based on these three quantities:

- stochastic complexity score of the gene expression data vector $HE = B(\mathbf{e}^N)$
- stochastic complexity score of the SNP data vector $HG = B(\mathbf{s}^N)$
- stochastic complexity score of the joint gene expression-SNP data pair $HGE = B([\mathbf{e}^N, \mathbf{s}^N])$

A pragmatic approach for obtaining an estimate of the resulting stochastic complexity of an eQTL data pair, when a patient's genotype is used as an explanatory

model for a gene expression profile, is to compare the 3 scores and calculate the amount of "obtained knowledge" in terms of the resulting NML-code length.

Intuitively speaking, if we contrast the 3 description lengths, we obtain an MDL-score telling us how good a SNP model fits or explains the observed gene expression profile of a patient. Good, predictive SNP models should yield good predictions of the gene expression values, leading to a grid layout that better describes the data resulting in an NML-code using fewer bits.

If the NML-code indicates that we need HE bits to describe the gene expression independently, i.e. without any further knowledge about the patient's genotype, and HG bits to describe the patients genotype, then the amount of knowledge we gain by using the patient's genotype as a model for explaining the gene expression, is equivalent to the score

$$score_{MDL} = HE + HG - HGE \tag{4.38}$$

with HGE being the MDL-optimal joint description of the SNP-transcript pair which is obtained through grid optimization via the mKM-algorithm.

Therefore, the importance of SNP-gene expression associations is ranked according to the amount of "shared knowledge", i.e. the better a genotype profile can be utilized to further compress a gene expression profile and drive down the stochastic complexity manifested through the NML-code, the stronger and more important is the association.

5 Making of qMAP - The MDL-analysis software for eQTL

5.1 Implementing qMAP in python

The software qMAP for discovering associations between SNPs and gene expressions in eQTL data is implemented in Python. Because publicly available eQTL datasets are usually distributed in a PLINK [66] compatible format consisting of three files, qMAP also takes as input those three files. Each file contains different data with a separate file for the genotype data and the gene expression data, as well as a mapping file containing auxiliary information about each SNP.

5.1.1 PLINK input file format

The three files that make up the eQTL dataset are the **.pheno*-file containing the gene expression values, the **.ped*-file containing the genotypes, i.e. SNP-data, and the **.map*-file containing information about each SNP, e.g. its dbSNP record.

Each file adheres to the formatting that is explained on the website of the PLINK toolkit software found at [72]. For creating eQTL data files compatible with both PLINK and qMAP, please adhere to the formatting instructions of the PLINK website.

The **.map*-file is the simplest to create as each row contains annotation information about each SNP present in the **.ped*-file.

The **.ped*-file containing the genotype information is a tab- or space-delimited file with each row describing one sample, e.g. the genotype profile of one patient and each column containing the following information:

Family ID	Individual ID	Paternal ID	Maternal ID	Sex	Genotypes

where the "Genotypes" column is expanded accordingly to include the SNP sequence of length L, i.e.

Genotype 1 $\cdots$ Genotype L

Each genotype column corresponds to one entry in the **.map*-file via the mapping relation: Row ν of the **.map*-file is the annotation information of SNP ν in the **.ped*-file, i.e. the "Genotype ν" column. Some mapping examples would be Genotype 5 in the **.ped*-file and row 5 in the **.map*-file or Genotype L corresponding to row L.

Care should be taken when creating a **.ped*-file because the first 6 columns, which contain information about the patient, are mandatory whereas the genotype columns are flexible. Furthermore, the genotypes are stored using letters from the nucleotide alphabet G, A, T, C and each genotype is stored as a bi-allelic marker:

G C T T A T $\cdots$

The phenotype data, which is in the case of eQTL the gene expressions, are stored in the **.pheno*-file with each row representing a patient. The **.pheno*-file contains a header-row that includes the "Family ID" , "Individual ID", and the names for the phenotypes, usually the gene names or the name of the disease for the affection status. Thus, the file is formatted as

Family ID Individual ID Affection Status Phenotype 1 $\cdots$ Phenotype G

with Phenotype 1 to G representing the names of the G genes whose gene expression values were recorded. For a proper analysis it is mandatory that the Family ID and Individual ID of the same patient to be the same in both the **.pheno*- and **.ped*-file.

5.1.2 Flowchart of the Python program

The Python analysis software qMAP consists primarily of two modules for obtaining the NML-code of a SNP-gene pair and consequently the association strength via the MDL-score; namely the grid optimization modules for the 1-dimensional and 2-dimensional case.

The program flowchart in Figure 5.1 depicts the main parts of the analysis algorithm and highlights the steps necessary to obtain the MDL-score as a measure of association strength between SNPs and gene expressions in eQTL data.

5.1.3 Data initialization module

As input qMAP requires three files which have been formatted according to the rules outlined in Section 5.1.1. The genotype data in the **.ped*-file as well as the

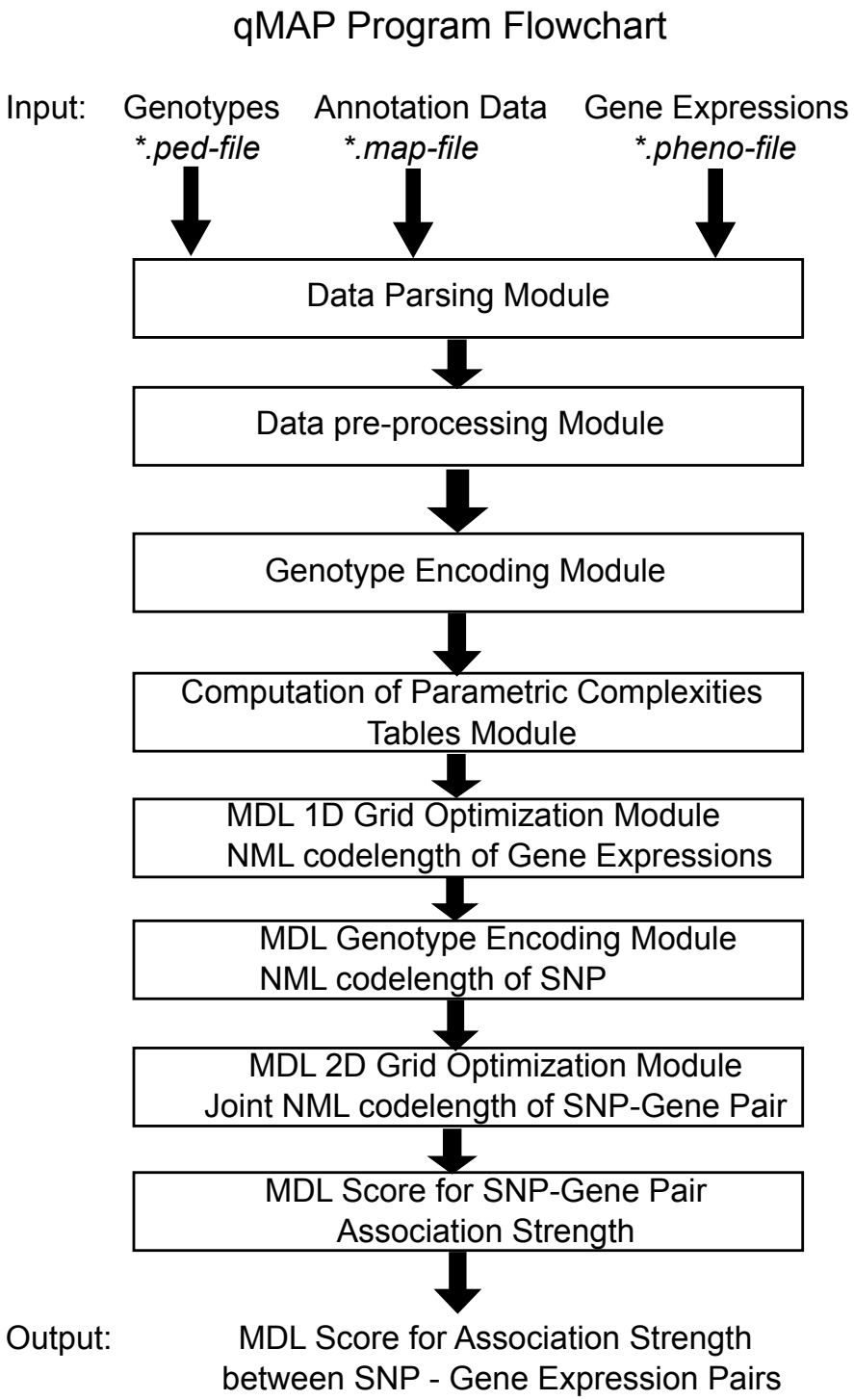


Figure 5.1: The data processing and analysis flowchart of qMAP.

gene expression data in the **.pheno*-file are parsed into Python NumPy-arrays by qMAP's parser-module. Contents of the **.map*-file are kept in a Python-dictionary that uniquely maps SNP IDs to their respective annotation information, which will be used later when the program outputs the obtained association test results for each SNP-gene pair.

As the input data are from real experiments, there is always the possibility of data contamination, failed experiments, incorrect read-outs of genotype and gene expression data, as well as another multitude of factors that could introduce artifacts into the data.

To cope with this situation, the raw data is pre-screened and filtered according to user defined quality assurance criteria. By default, qMAP incorporates the threshold criteria of [33, 73] for filtering the raw data.

The allele checker module scans through the genotype and confirms if each genotyped locus is bi-allelic. If more than 2 alleles are present at a locus, that position is masked and excluded from further analysis.

Depending on the user defined error rate tolerance, for each genotyped locus all samples are scanned for missing or incorrect entries in the eQTL data. If the number of either missing or incorrect values exceeds the allowable tolerance for the error rate, those data entries are also masked and excluded from further analysis.

Moreover, SNPs can also be filtered with respect to the minor allele frequency. In [33] it was suggested to drop SNPs from the analysis if their minor allele frequency is less than 1%. The data initiation module has the option to tag SNP loci with minor allele frequencies which do not pass the filtering criteria and exclude them from downstream analysis.

Gene expression data are also pre-processed in order to remove genes from the analysis where it was not possible to obtain robust transcript counts for all samples. This is achieved via a user defined threshold value. Suggestions about reasonable filtering values can be found for example in [33, 73].

Nonetheless, should a data entry be found to contain missing gene expression values, but the total number of incorrect entries does not surpass the user defined threshold, then in that case, the missing data entries are filled with placeholder values, namely the median value of the gene expressions estimated from the available samples.

In a first preprocessing step, each SNP in the **.ped*-file is tagged with an internal running SNP ID number and the genotype is converted to a numerical value via the encoding procedure outlined in Section 3.1. This coding step also determines the SNP axis grid interval $\tilde{\xi}_S$.

Then, the gene expression values are loaded and each gene obtains a unique gene ID number which will be used by qMAP for internal processing steps when

calculating association strengths between various pairs of SNPs and genes.

By scanning through the data we obtain the resolution parameter ε. Currently, the implemented procedure sorts the gene expression values and then calculates the difference between adjacent value pairs.

Algorithm 1 Get data resolution ε Module

```
1: Differences = []
2: Sort input data vector
3: for All adjacent value pairs do
4:     Calculate difference
5:     Append calculated value to Differences-vector
6: ε = min(Differences)
```

Since qMAP will evaluate many SNP-gene pairs, we can speed up the computation by pre-calculating the values for the parametric complexities R for the $1D$ and $2D$ grid optimization case and store all necessary results in tables, including intermediate results which will be accessed during the dynamic programming procedure.

Implementing the KM-algorithm's recursive formula for calculating parametric complexities for the class of multinomial models, for both the 1-dimensional and 2-dimensional case, qMAP builds the necessary tables $pc1$ and $pc2$ that store the parametric complexity values $R(N, K, M)$ and $R(N, K, \tilde{K}, M_2)$ for $1D$ multinomial model classes M and $2D$ multinomial model classes M_2 respectively and keeps them ready in memory.

Instead of having to recalculate the parametric complexity value R in each loop of the dynamic programming equation in 4.25, the values are calculated once and then just accessed on demand when optimizing the grid layout.

Although it is costly to pre-build the tables, this step saves computation cost if the MDL-score is calculated for many SNP-gene pairs because it replaces the time consuming recursive calculation of the parametric complexity in each step with a simple table query.

5.1.4 $1D$ grid optimization module for obtaining NML-codelengths of gene expression quantitative traits

The MDL-scoring function for obtaining the NML-codelength for the gene expressions implements the 1-dimensional KM-algorithm for the grid optimization of Section 4.1.3 and 4.1.4.

Although the module is implemented in a general way, i.e. it can determine the NML-code of any kind of continuous input data, e.g. continuous trait variables

like BMI or blood pressure and thus perform association studies between SNP and continuous traits, the main focus of this application is to perform association tests between SNPs and gene expressions in eQTL data and therefore, the explanations mainly focus on gene expression data.

The input vector **e** is sorted in rank ascending order and the set of putative grid points Υ is obtained.

Algorithm 2 Get grid points Υ Module

```
Grid points vector Υ = []
for All data points e in e do
    ξ_l = e − ε/2
    ξ_r = e + ε/2
    if ξ_l ∋ Υ then
        Append ξ_l to Υ
    if ξ_r ∋ Υ then
        Append ξ_r to Υ
Return Υ
```

To speed up computation, several modifications to the mathematical formulation of the original KM-algorithm have been made in this Python implementation.

Instead of counting the number of data points N_ζ of **e** that fall into the interval defined by the grid boundaries ξ_ζ in each iteration of the dynamic programming procedure, the counts N_ζ are obtained in an initial for-loop and stored in a table. Since the dynamic programming recursion re-uses previous results, by storing the counts for later retrieval in a table the program's execution time is accelerated.

In order to obtain both the MDL-optimal number of grid intervals K and the respective placing of the borders ξ_κ, the back-tracking procedure requires us to store for each iteration step the MDL-score $\hat{B}_{\kappa,\zeta}$ and the grid configuration $\Xi = \Upsilon_{\kappa,\zeta}$ which achieves the MDL-optimal score. Therefore, we declare two Python dictionaries that hold those values for each (κ, ζ)-key tuple. Via the unique key (κ, ζ), the back-tracking procedure is able to reconstruct the placement of grid points ξ_κ for the MDL-optimal grid interval number K, which defines the NML-code of the gene expression dataset **e**.

Another speed-optimization is to store all combinations of $\hat{B}_{\zeta,\zeta'}$-subscores in the $K = 2$ case so that those subscores do not have to be laboriously recalculated in each iteration step for higher K, i.e. $K > 2$.

The code example of the $1D$ grid optimization Python module 9 shows the implementation of the original KM-algorithm including all speed-up modifications.

Algorithm 9 shows how we obtain a grid layout Ξ for calculating the NML-codelength of one gene expression data vector **e**. The KM-method outputs a

Algorithm 3 $1D$-grid optimization Module for NML-coding (adapted from [70])

Input: data $\mathbf{e}$, ε, parametric complexities table R, $K = K_{max}$
$\Upsilon = getGridPoints\,(\mathbf{e})$ via Algorithm 2
N is number of data points in $\mathbf{e}$
Dictionary $nCounts = \{\}$ for data points N_ζ in the interval $[e_{min} - \frac{\varepsilon}{2}, \xi_\zeta]$
$Z + 1 = |\Upsilon|$
for ζ in range$(1, Z + 1)$ **do**
    $nCounts[\zeta] = N_\zeta$
Dictionary for MDL-scores $B_{\kappa,\zeta} = \{\}$
Dictionary for MDL-subscores $Bsc_{\zeta,\zeta'} = \{\}$
Dictionary for grid configurations $\Upsilon_{\kappa,\zeta} = \{\}$
for κ in range$(1,K + 1)$ **do**
    if $\kappa > 2$ **then**
        for ζ in range$(\kappa, Z + 1)$ **do**
            $B_{min} = []$
            $Z_{min} = []$
            for ζ' in range$(\kappa - 1, \zeta)$ **do**

$$score = B_{\kappa,\zeta}[(\kappa - 1, \zeta')] + Bsc_{\zeta,\zeta'}[(\zeta, \zeta')] + \log\ (Z + 1 - \kappa) - \log\ (\kappa - 1) + \log\ R(nCounts[\zeta], \kappa - 1) - \log\ R(nCounts[\zeta'], \kappa - 1 - 1)$$

                Append $score$ to B_{min}
                Append ζ' to Z_{min}
            $B_{\kappa,\zeta}[\kappa, \zeta] = min(B_{min})$
            $\Upsilon_{\kappa,\zeta}[\kappa, \zeta] = Z_{min}[argmin(B_{min})]$
    else if $\kappa == 1$ **then**
        for ζ in range$(1, Z + 1)$ **do**
            $B_{\kappa,\zeta}[(\kappa, \zeta)] = nCounts[\zeta] \cdot (\log\ (\varepsilon \cdot nCounts[\zeta]) - \log\ (|\xi_\zeta - \xi_0| \cdot N))$
    else if $\kappa == 2$ **then**
        for ζ in range$(\kappa, Z + 1)$ **do**
            $B_{min} = []$, $Z_{min} = []$
            for ζ' in range$(\kappa - 1, \zeta)$ **do**

$$Bsc_{\zeta,\zeta'}[\zeta, \zeta'] = -\,(nCounts[\zeta] - nCounts[\zeta']) \cdot (\log\ (\varepsilon \cdot (nCounts[\zeta] - nCounts[\zeta'])) - \log\ (|\xi_\zeta - \xi_{\zeta'}| \cdot N))$$

$$score = B_{\kappa,\zeta}[(\kappa - 1, \zeta')] + Bsc_{\zeta,\zeta'}[(\zeta, \zeta')] + \log\ (Z + 1 - \kappa) - \log\ (\kappa - 1) + \log\ R(nCounts[\zeta], \kappa - 1) - \log\ R(nCounts[\zeta'], \kappa - 1 - 1)$$

                Append $score$ to B_{min}, Append ζ' to Z_{min}
            $B_{\kappa,\zeta}[\kappa, \zeta] = min(B_{min})$
            $\Upsilon_{\kappa,\zeta}[\kappa, \zeta] = Z_{min}[argmin(B_{min})]$
Return $B_{\kappa,\zeta}$, $\Upsilon_{\kappa,\zeta}$

dictionary containing the best MDL-scores for all grid configurations defined by the number of grid intervals K and the corresponding borders of those intervals $\xi_\kappa \in \Upsilon$.

Before continuing to calculate the NML-codelength of $\mathbf{e}$ we first have to extract the MDL-optimal grid layout for the data. The achievable granularity of the grid layout not only depends on the complexity of the data, which we model with multinomial distributions from the model class M, but also on the available sample size N. In order to obtain the layout we apply the backtracking algorithm of the KM-method. Details about the backtracking algorithm are explained in Kontkanen and Myllymäki's original paper [70]. Nonetheless, the reader might be interested in another easy to follow mathematical formulation by Kameya in [71].

Although the backtracking procedure might be difficult to understand at first sight, the implementation is quite easy. The dictionary $B_{\kappa,\zeta}$, which is the output of the $1D$-grid optimization algorithm 9 utilizing the KM-method, is used together with the grid point set Υ as input for the backtracking algorithm. Since $Z+1$ encompasses all data points of $\mathbf{e}$, the MDL-score $B_{\kappa,\zeta}[\kappa, Z+1]$ is the grid optimization solution for each number of grid intervals $\kappa = 1, \cdots, K_{max}$. Therefore, for each κ the best score B_κ can be obtained from the dictionary $B_{\kappa,\zeta}$. Consequently, among all the scores B_κ we get for different grid configurations, the minimum score $min_\kappa(B_\kappa)$ defines the best κ, i.e. K_{opt}, and hence, the MDL-optimal grid configuration.

Algorithm 4 Back-tracking Module (adapted from [70])

1: Input: $B_{\kappa,\zeta}$, Υ, $K = K_{max}$
2: Optimal MDL-scores $B_{opt} = []$
3: **for** κ in range$(1, K)$ **do**
4: Append $B_{\kappa,\zeta}[(\kappa, Z+1)]$ to B_{opt}
5: MDL-optimal $K_{opt} = argmin(B_{opt}) + 1$
6: Return MDL-optimal number of grid intervals K_{opt}

With the MDL-optimal number of grid intervals K_{opt} and grid layout $\Upsilon_{K_{opt}}$ we finally obtain the NML-code of the gene expression data $\mathbf{e}$ from the dictionary $B_{\kappa,\zeta}$ via $B_{\kappa,\zeta}[(K_{opt}, Z+1)]$. This is the shortest description of the gene expression data according to the KM-algorithm and thus the stochastic complexity score HE becomes

$$HE = B_{\kappa,\zeta}[(K_{opt}, Z+1)]. \tag{5.1}$$

This concludes the implementation of the 1-dimensional grid optimization Python module and we move on to the $2D$-module necessary for obtaining the joint NML-code of a SNP-gene expression pair before being able to acquire the final MDL-score for the association strength.

5.1.5 $2D$ grid optimization module for obtaining the NML-code of the joint description of genotype and transcriptome

The $2D$-grid optimization module extends the structure of the $1D$-module. Of both the grid axes, one for the gene expression data and one for the SNP data, only the partitioning of the gene expression axis has to be optimized. By making use of the SNP encoding scheme, the partitioning of the SNP axis of the grid is already given. Yet, similar speed optimizations that were introduced in the $1D$-grid implementation also apply in the $2D$-case. What needs to be taken into account when implementing the $2D$-grid module is the preset SNP axis whose number of intervals $\tilde{K} = 3$ are also called labels. The use of those labels will make the dynamic programming computation more efficient through the use of efficient auxiliary data structures.

Like in the 1-dimensional case, the set of putative grid points Υ is obtained from the input gene expression data $\mathbf{e}$ via Algorithm 2.

The first difference between the $1D$-grid and $2D$-grid module is the way of counting the frequencies of data points falling into a grid interval. By using the labels ϱ for referring to the grid intervals $\tilde{K}$ of the SNP axis, the dictionary $nCounts$ is extended to hold for every ζ a vector $\mathbf{n}_{\zeta\varrho}$ containing the counts $N_{\zeta,\varrho}$ for $\varrho = 1, 2, 3$ of data points (s, e) falling into the grid area defined by the interval $[e_{min} - \frac{1}{2}, \xi_\zeta]$ on the gene expression axis and the label ϱ on the SNP axis; see Algorithm 5.

Algorithm 5 $2D$-grid frequency counter

```
Input: [e, s]
n_{ζϱ} = []
for ϱ in range(3) do
    n_{ζϱ} = N_{ζ,ϱ}
Return n_{ζϱ}
```

With the extended counting method, the $2D$-grid optimization method proceeds in a similar fashion like the $1D$-method for obtaining the MDL-optimal grid layout, but incorporates different parametric complexities because the $2D$ grid points are now modeled via a mixture of multinomial models.

Before proceeding in laying out the entire implementation of the mKM-method for optimizing a 2-dimensional grid, we will first define some helper functions that aid in the dynamic programming procedure.

The first function is the calculation of the initial score of the recursion, i.e. the case for $K == 1$. The second function that will be defined is the table construction function that records all MDL-subscores for ζ,ζ', and ϱ when $K == 2$.

When initializing the grid optimization function, since the SNP axis $\tilde{\xi}_S$ has already been compartmentalized according to the genotype encoding process of Section 3.1, the MDL-optimal placement of grid interval points ξ_κ and the MDL-optimal number of total grid intervals has to be obtained in a SNP dependent fashion only for the remaining gene expression axis ξ_E via the mKM-algorithm.

In contrast to the $1D$-method, the initialization of the dynamic programming recursion uses a slightly modified scoring function that takes the grid labels ϱ of the SNP axis $\tilde{\xi}_S$ into account.

Algorithm 6 $2D$ initial scores Module

Input: ζ, $nCounts$, ε, Υ, N
$\mathbf{n}_{\zeta\varrho} = nCounts[\zeta]$
$score = 0$
for ϱ in range(3) **do**
    $score \; += -\mathbf{n}_{\zeta\varrho}[\varrho] \cdot (\log \; (\varepsilon \cdot \mathbf{n}_{\zeta\varrho}[\varrho]) - \log \; (|\xi_\zeta - \xi_0| \cdot N))$
$score \; += R(\sum_{\varrho=1}^{3} \mathbf{n}_{\zeta\varrho}[\varrho], 3)$
Return $score$

Algorithm 7 $2D$ MDL-subscore Module

Input: ζ, ζ', $nCounts$, ε, Υ, N
$\mathbf{n}_{\zeta\varrho} = nCounts[\zeta]$
$\mathbf{n}_{\zeta'\varrho} = nCounts[\zeta']$
$subscore = 0$
for ϱ in range(3) **do**

$$subscore \; += -(\mathbf{n}_{\zeta\varrho}[\varrho] - \mathbf{n}_{\zeta'\varrho}[\varrho]) \cdot (\log \; (\varepsilon \cdot (\mathbf{n}_{\zeta\varrho}[\varrho] - \mathbf{n}_{\zeta'\varrho}[\varrho])) - \log \; (|\xi_\zeta - \xi_{\zeta'}| \cdot N))$$

Return $subscore$

The $2D$ MDL-subscore Module is applied during the second recursion step for $K == 2$ of the dynamic programming procedure to help construct the table $Bsc_{\zeta,\zeta'}$ that holds all MDL-subscores for various combinations of ζ and ζ'.

Since we run through all possible combinations of ζ and ζ' with the double for-loops ζ in range$(\kappa, Z+1)$ and ζ' in range$(\kappa - 1, \zeta)$, it is wise to store those intermediate scoring results in the table $Bsc_{\zeta,\zeta'}$ so that they do not have to be re-calculated in subsequent recursions.

Another technical assistance is the function that computes the MDL-score of each particular grid configuration. It sums up the information contained in the tables for MDL-subscores $Bsc_{\zeta,\zeta'}$ and parametric complexities for the 2-dimensional case R^{2D}.

Algorithm 8 $2D$ MDL-score Module

1: Input: ζ, ζ', $nCounts$, κ, $Bsc_{\zeta,\zeta'}$, $B_{\zeta,\zeta'}$, Υ, R^{2D}
2: $\mathbf{n}_{\zeta\varrho} = nCounts[\zeta]$
3: $\mathbf{n}_{\zeta'\varrho} = nCounts[\zeta']$
4: $N_{\zeta} = \sum_{\varrho=1}^{3} \mathbf{n}_{\zeta\varrho}[\varrho]$
5: $N_{\zeta'} = \sum_{\varrho=1}^{3} \mathbf{n}_{\zeta'\varrho}[\varrho]$
6: $score = B_{\zeta,\zeta'}[(\kappa-1,\zeta')] + Bsc_{\zeta,\zeta'}[(\zeta,\zeta')] + \log\,(Z+1-\kappa) - \log\,(\kappa-1) + R^{2D}(N_{\zeta},\kappa-1) - R^{2D}(N_{\zeta'},\kappa-2)$
7: Return $score$

With the initialization function Algorithm 6 and MDL-subscoring function Algorithm 7 in place, we can proceed with the implementation of the mKM-algorithm for obtaining the NML-code of a joint SNP-gene expression data pair via $2D$-grid optimization.

The final step for obtaining the MDL-optimal grid layout is to use the backtracking Algorithm 4 to obtain the number of grid intervals K_{opt} and the placement of interval points for the gene expression axis of the grid. Together with the SNP axis, the MDL-partitioning of the gene expression axis is used to calculate the NML-code for the joint description of the dataset $[\mathbf{e},\mathbf{s}]$. The resulting code length is the MDL-optimal description of a SNP-gene expression data pair when using multinomial distributions to model the statistical properties of the data.

Consequently, the NML-code of a SNP-gene pair delivers the description length HGE

$$HGE = B_{\kappa,\zeta}[(K_{opt}, Z+1)], \tag{5.2}$$

with $B_{\kappa,\zeta}$ holding the MDL-scores for the 2-dimensional grid configurations and K_{opt} being the number of grid intervals on the gene expression axis.

5.1.6 MDL association score module

For each tested SNP-gene pair a MDL association score is calculated based on the achieved compression by the NML coding procedure of the previous sections. The MDL-score reflects the codelengths necessary to describe the relation between the SNP and gene expression profile. High MDL-scores indicate a strong association whereas lower scores point towards weaker associations.

The final MDL-score outputted by this Python module is made up of three components, namely the NML-codes for describing the genotype SNP, the description of the gene expression, and the joint description of the SNP-gene pair.

Algorithm 9 mKM $2D$-grid optimization Module for NML-coding

Input: data $[\mathbf{e}, \mathbf{s}]$, ε, $K = K_{max}$, parametric complexities table R^{1D}, R^{2D}
$\Upsilon = getGridPoints\,(\mathbf{e})$ via Algorithm 2
N is number of data points in $[\mathbf{e}, \mathbf{s}]$
Dictionary $nCounts = \{\}$ for grid area frequency counts obtained with Algorithm 5
$Z + 1 = |\Upsilon|$
Dictionary for MDL-scores $B_{\kappa,\zeta} = \{\}$
Dictionary for MDL-subscores $Bsc_{\zeta,\zeta'} = \{\}$
Dictionary for grid configurations $\Upsilon_{\kappa,\zeta} = \{\}$
for κ in range$(1,K+1)$ **do**
    if $\kappa > 2$ **then**
        for ζ in range$(\kappa, Z+1)$ **do**
            $B_{min} = []$
            $Z_{min} = []$
            for ζ' in range$(\kappa - 1, \zeta)$ **do**
                score calculated with $2D$ MDL-score Algorithm 8
                Append *score* to B_{min}
                Append ζ' to Z_{min}
            $B_{\kappa,\zeta}[\kappa, \zeta] = min(B_{min})$
            $\Upsilon_{\kappa,\zeta}[\kappa, \zeta] = Z_{min}[argmin(B_{min})]$
    else if $\kappa == 1$ **then**
        for ζ in range$(1, Z+1)$ **do**
            $B_{\kappa,\zeta}[(\kappa, \zeta)]$ = initial scores via Algorithm 6
    else if $\kappa == 2$ **then**
        for ζ in range$(\kappa, Z+1)$ **do**
            $B_{min} = []$, $Z_{min} = []$
            for ζ' in range$(\kappa - 1, \zeta)$ **do**
                Create $Bsc_{\zeta,\zeta'}[\zeta, \zeta']$-table using MDL-subscore Algorithm 7
                Calculate *score* with $2D$ MDL-score Algorithm 8
                Append *score* to B_{min}, Append ζ' to Z_{min}
            $B_{\kappa,\zeta}[\kappa, \zeta] = min(B_{min})$
            $\Upsilon_{\kappa,\zeta}[\kappa, \zeta] = Z_{min}[argmin(B_{min})]$
Return $B_{\kappa,\zeta}$, $\Upsilon_{\kappa,\zeta}$

Each NML-code can be obtained by utilizing the $1D$ grid and $2D$ grid optimization modules contained in qMAP. In general, short NML-codes indicate inherent functional relationships between genotype and transcriptome. For obtaining a measure of importance of those functional relationships, the NML-code of the joint description of a SNP-gene pair has to be contrasted against the individual description lengths of the genotype and the gene expression, obtained by NML-coding respectively.

Consequently, the MDL-score which shows the strength and importance of an association between a SNP and a gene expression pattern is defined as:

$$score_{MDL} = HE + HG - HGE \tag{5.3}$$

with HE being the NML-codelength provided by the qMAP Python module of Section 5.1.4 for describing the gene expression, HG being the NML-codelength of the genotype, and HGE the NML-codelength of the SNP-gene pair.

In order to obtain all the MDL-scores between a range of SNP loci and a quantitative gene expression trait in an eQTL dataset, the following loop is applied:

Algorithm 10 MDL-score calculation module

```
Input: SNP loci, gene expression
Calculate NML-code HE for the gene expression
for each SNP do
    Calculate NML-code HG for the genotype SNP
    Calculate NML-code HGE for SNP-gene pair
    Obtain MDL-score via score_MDL = HE + HG − HGE
```

After obtaining the MDL-scores for all SNP loci, the list of scores is ranked from the highest MDL-score at the top to the lowest MDL-score at the bottom.

6 Performance evaluation results

6.1 Evaluating qMAP using a synthetic simulated eQTL dataset

For the theoretical evaluation of the analysis approaches we are going to use the synthetic eQTL dataset that was created by Bartlett and Ray [8]. Although testing analysis algorithms on real datasets is paramount, with the help of a simulated dataset that tries to capture as many characteristics of a real eQTL dataset it is possible to study the performance of our qMAP algorithm and compare the results to the other state-of-the-art eQTL analysis methods PLINK [66], MI-KDE [19, 48], and MIC [49] under a fair and controlled environment.

Bartlett and Ray argue that eQTL analysis is an upcoming field of study in molecular biology and medicine because technological progress in the area of sequencing machines make it relatively cost effective to obtain both a genotype and transcriptome, i.e. gene expression activity, readout from a patient [8] and hence use those data to improve the diagnostic and treatment capabilities.

Adhering to ethical and data security standards it becomes possible to bundle data about genotypes, gene expressions, and disease status information in large databases. As has already been mentioned in the introduction, the main goal of an analysis should be to draw clear-cut associations between those three entities [8].

Because eQTL data are quite novel, analysis algorithms that can deal with the complexity of those data have to be designed. According to Bartlett and Ray, eQTL is such an *"information dense"* domain [8], that it is both challenging and very rewarding to create an analysis approach that can extract "useful" information out of eQTL-data.

When we directly apply an analysis algorithm to a real dataset without prior testing on a simulated dataset, the problem is how to discern "useful" information from noise or other unrelated interactions in which we are not particularly interested in, if we do not a priori know what is in the data (see further explanations in [8]). Therefore, any analysis approach should first be evaluated using a good simulated dataset in which the associations between SNPs, gene expression, and potentially a disease phenotype are known so that we can test if an analysis tool is able to reveal those basic interaction networks.

As has been argued by Bartlett and Ray in [8], creating a simulated eQTL dataset is quite difficult because one must be careful not to oversimplify the complexities of biological nature. One the one hand, the synthetic dataset should mirror the characteristic properties of real eQTL data. On the other hand, the synthetic dataset should contain a known set of associations between its entities forming the "interaction network" which should be detectable by capable analysis algorithms.

The authors Bartlett and Ray, by using an sophisticated simulation technique which will briefly be described in Section 6.1.1, created a synthetic eQTL dataset which balances the opposing poles of natural complexity vs. necessary simplification and thus has the potential to serve as a kind of "gold standard" for benchmarking eQTL analysis algorithms [8].

The eQTL dataset of Bartlett and Ray provides an excellent basis for benchmarking our MDL qMAP tool against the established gene analysis toolkit PLINK [66], our implementation of Shannon's mutual information dependency measure from information theory based on the paper by Dawy et al. [19] using kernel density estimates for acquiring the underlying statistical distributions in the data (MI-KDE), and the Maximal Information Coefficient (MIC), a novel dependency measure for pairs of random variables that can be used in exploratory data analysis settings [49].

All four analysis approaches are evaluated first on the synthetic eQTL data of Bartlett and Ray [8].

6.1.1 Simulating eQTL data

In order to preserve as much as possible from the complexity of real biological eQTL data, Bartlett and Ray developed a method that enabled them to mix several real eQTL dataset into a large synthetic one and simultaneously introduce an artificial network of associations between SNP, gene expressions, and disease status [8]. They refer to their methodology for aggregating various real eQTL datasets into a single dataset as *"data shuffling technique"* [8] and the ability to introduce artificial relationships into the synthetic eQTL dataset while at the same time preserving much of its real biological complexity as *"spiking in"* [8].

The artificial associations in the simulated eQTL dataset are simply referred to as *"the interaction network"* by Bartlett and Ray [8] and we adopt their terminology here.

The real biological eQTL dataset upon which Bartlett and Ray's simulated eQTL data is based upon are Liu's data from a human brain eQTL association study in [74] and Myers' investigation into human cortical gene expression [33].

Myers' data consists of 193 patients whose gene expression profile and genotype was determined after their death by following strict ethical rules [33]. It consists

of gene expression values for the database of all known genes at that time and a total of 500.000 SNPs for each of the 193 patients.

Liu's data contains 164 brain samples from both healthy and ill patients. The definition of ill patients in Liu's study [74] were persons who were diagnosed to have one of the following psychological disorders: Schizophrenia, Bipolar Disorder, Major Depression. On the opposite, the definition of healthy patients were those persons who were diagnosed not to have such a disease and therefore labeled as normal samples for the case-control study [74]. The brain samples for the eQTL dataset were obtained after each patient's death.

As has been pointed out in [8], a ready-to-use simulated dataset needs to undergo some quality assurance procedures. In contrast to real datasets, where the analyst has to preprocess the data in order to filter out for example failed experiments or missing data points, e.g. when it was not possible to obtain a patients genotype at a specific locus, a simulated dataset need not have missing or ambiguous data points in it. Therefore, the real datasets of Myers [33] and Liu [74] were filtered by Bartlett and Ray before using them as a basis to create a simulated eQTL data set [8].

Data shuffling technique

In this Section we briefly explain the *"data shuffling technique"* of Bartlett and Ray [8].

For simulation purposes, several real datasets need to be integrated into a unified data repository. It is not easy to mix data from several eQTL studies, but it is possible to obtain a general idea about the correspondence of genotypes with gene expression patterns in each patient according to [8].

These basic correlations are utilized by the *"data shuffling"* technique to integrate the information of several real eQTL datasets and generate a simulated one which keeps the biological complexity of real data [8].

By learning and establishing the various correlations between a genotype and an expression pattern, the data shuffling technique first proceeds by generating a random genotype for a simulated person.

Essentially, each simulated SNP is just a random draw from a database of existing variations. Furthermore, haplotype information is also included by specifying which allele comes from the maternal and which from the paternal chromosome. The database which Bartlett and Ray used to act as a pool of SNP variants are the two real eQTL datasets from Myers [33] and Liu [74] plus phased data from the 1000 genomes project [75, 76].

While compiling such a virtual genome for each patient in the simulated dataset, one has to make sure that the virtual genotype produces a gene expression pattern

which could be expected from a real genome. The data shuffling technique ensures that the correspondence between virtual genotypes and virtual gene expression patterns matches the observations in real datasets. This is achieved by applying a filter which checks if the correlations between SNPs and gene expressions in real data, i.e. the underlying statistical structure or model of the eQTL data, is similar to the obtained patterns between virtual genotype and simulated gene expression data [8].

After a virtual genome, which passed all the filtering criteria, has been assigned to our virtual patient, a virtual gene expression pattern that corresponds to the virtual genome has to be created.

This is achieved by utilizing statistical relationships which arise from mapping genotypes to gene expression patterns in real eQTL datasets. The obtained distributions were used by Bartlett and Ray to separate each SNP's effect on the gene expression via linear regression.

Consequently, the basic building blocks for constructing a virtual gene expression profile from a virtual genome were obtained; namely the statistical relationship between a SNP and its induced transcript expression of a gene. Now it becomes possible to simulate the gene expression pattern of a combination of many SNPs originating from a virtual genome. By comparing the simulated gene expression patterns against the patterns found in real eQTL data, Bartlett and Ray's filter rules only allow consistent virtual patterns to pass and henceforth be used as a virtual gene expression pattern in the synthetic eQTL data [8].

The data shuffling enables the creation of virtual patients having a virtual genotype and a virtual transcriptome. Thus, an arbitrary number of virtual patients can be created which make up the sample size of the synthetic eQTL data set.

The data shuffling technique delivers the ground work for the eQTL simulation by sampling real datasets and conveying as much as possible of the biological complexity into the simulated data. As a consequence, statistical relationships between virtual genomes and simulated gene expression patterns mimic the behaviour of real data. This establishes the first part of Bartlett and Ray's eQTL simulation procedure.

Spike-In of artificial interaction network

In a next step, some artificial associations between SNPs and gene expressions have to be incorporated into the simulated data. The purpose is to create a *"ground truth"* [8] which should be recoverable by analysis algorithms.

By overlaying a network of interactions between various SNPs and gene expressions on top of the data produced by the data shuffling technique, according to Bartlett and Ray [8], real biological data get mixed with a user defined model

of known associations so that the resulting simulated dataset both preserves the biological complexity inherent in real eQTL data and maintains the ability to benchmark algorithms against it because it uses a known data generating model.

The basis of Bartlett and Ray's artificial interaction network are 15 genes from the cadherin protein superfamily. Since it has been reported in the literature [8] that there is evidence for those genes interacting with each other in an ensemble, they are a good candidate to hide an overlaid artificial interaction network in a set of already existing biological interactions.

Using the data shuffling technique, virtual genotypes and gene expression patterns were created for those 15 genes.

Although the SNPs and their identifiers from the dbSNP database [20] originate from real datasets, it should be noted here that their biological interpretation is lost in the simulated eQTL dataset. The same is true for the gene expressions. While the SNPs and genes retain their known names from the literature, the detectable interactions in the simulated dataset only resemble and mimic the complexity of associations in real eQTL data. Therefore, analysis results of the simulated eQTL data cannot be interpreted as having any real biological meaning. In other words, the associations detected between SNPs and gene expression patterns by an analysis algorithm in the simulated eQTL dataset cannot be interpreted as real biological associations. The interpretations and conclusions resulting from such an analysis cannot be transferred to mean that such an association between SNP and gene expression actually exists in the original real eQTL dataset. After all, this is a simulated dataset devoid of true biological interpretations and its main purpose is to test the efficiency of eQTL analysis algorithms.

To highlight the differences between real and simulated eQTL data, gene names and dbSNP identifiers will be used when reporting the results for real datasets. As for the results regarding simulated eQTL data, internal gene ID and SNP ID numbers will be used.

Given the set of both virtual genomes and virtual gene expressions, Bartlett and Ray defined that the expression of each of the 15 genes is controlled and regulated by only one primary SNP in the virtual genome [8].

Modeling this interaction was accomplished by first calculating the correlations between virtual genomes and virtual gene expressions as well as between the virtual expressions of all 15 genes. The obtained correlations were stored in matrices.

A statistical model for the interaction network was created which models the influence of the SNP random variable with its 3 possible genotype outcomes on the gene expression [8].

The correlation matrices are altered accordingly in order to incorporate the sta-

tistical model for the interaction network.

After having defined the underlying statistical structure of the spike-in network, Bartlett and Ray proceed by creating the simulated eQTL dataset with their R-implementation of the simulation algorithm which is explained in detail in [8].

The resulting gene expression patterns were z-transformed to facilitate smoother analysis, which means that the distribution of the gene expression patterns are normalized to 0 mean and variance 1.

According to Bartlett and Ray [8], the above procedure has the effect that real biological signals are interwoven with artificially spiked-in signals. This ensures a *"balance between real biological complexity and simulation specificity"* [8] which enables the assessment of various eQTL analysis frameworks.

Simulated disease status

The eQTL dataset of Bartlett and Ray [8] comes with another interesting feature. It further contains a simple disease model that defines a disease status for all virtual patients. This feature enables to check analysis frameworks regarding their ability to not only detect associations between SNPs and gene expressions, but to also detect associations between disease status, genotype, and gene expression patterns as well, which is one of the main purposes of eQTL studies as outlined in Section 2.1.

Initially, the simulated eQTL dataset was created for a contest [8] in order to find analysis frameworks which are well suited to deal with eQTL data, i.e. assist physicians in gaining a better understanding of the data by extracting valuable information and provide interaction models that explain the onset and pathogenesis of disease.

To facilitate the above mentioned goal, an analysis approach should first identify all the SNPs and gene expressions which convey important information about the disease. Physicians are then able to gain a better understanding of the disease's molecular mechanisms. Based on the acquired knowledge, diagnostic methodology and/or treatment approaches might be improved.

The contest initiators and authors of [8] put great hopes into the analysis of eQTL, because they believe that personalized medicine might benefit a lot from eQTL by potentially optimizing the treatment strategy depending on personal genetic information.

Although the disease model used in this simulated dataset is simple in nature, it points to future directions of what analysis approaches ought to extract from eQTL data and what kind of results they should deliver to physicians.

For creating a virtual disease which affects the virtual patients of the synthetic eQTL dataset, Bartlett and Ray utilized Wright's liability-threshold model [77]

to assign a disease status to each patient depending on their gene expressions.

Out of the available 15 genes, 8 genes were chosen to determine the outbreak of the virtual disease. Especially the numerical values of the gene expressions, i.e. the transcriptomic activity of those 8 genes, determine in a probabilistic way if a virtual patient gets the disease or not.

Wright's liability-threshold model [77] requires two parameters to work, namely a threshold value and an outbreak probability. Given a user defined threshold ψ and the outbreak probability, in this case 80%, a virtual patient contracts the disease with a probability of 80% if and only if the sum of the 8 gene expressions is above the threshold ψ and always 0 if it is below [8].

Therefore, given the user defined threshold ψ and a vector $\mathbf{e} = [e_1, \cdots, e_8]$ containing the 8 gene expression values, the mathematical model employed by Bartlett and Ray for assigning a virtual disease status to the virtual patients in the simulated eQTL dataset can be expressed as

$$\begin{aligned} Pr\{\text{Disease affected}\} = 0.8 \quad \text{if} \quad \sum_{i=1}^{8} e_i \geq \psi \\ Pr\{\text{Disease unaffected}\} = 0 \quad \text{if} \quad \sum_{i=1}^{8} e_i < \psi. \end{aligned} \tag{6.1}$$

The 8 genes, which are used in Wright's liability-threshold model to determine the affection status of a virtual patient as a function of their gene expression values, together with the 8 primary SNPs, which regulate the gene activity, make up the *"interaction network"* [8] that characterizes the virtual disease of the synthetic eQTL dataset.

Despite the fact that there are 15 genes in the eQTL dataset, only 8 are actually related to the virtual disease whereas the remaining 7 genes are not. Nonetheless, for all 15 genes exactly one SNP was chosen by Bartlett and Ray to control the gene expression. These relationships between SNPs, gene expression, and disease are summarized for the synthetic eQTL data in Table 6.1. Since the synthetic data is based on real eQTL datasets, the table contains both the real gene names and the dbSNP identifiers for the SNPs, which were obtained from the **.map*-file, alongside their internal IDs, which were assigned to them during our analysis.

It should be noted that the entire synthetic eQTL dataset consists of 500 virtual patients, with 193 subjects being affected with the above defined virtual disease and 307 unaffected. Furthermore, we have genotype information of 7555 SNPs for each patient together with gene expression values for 15 genes.

Moreover, a visualization of the entities which make up the interaction network in the simulated eQTL dataset is given in Figure 6.1.

Gene Name	Internal Gene ID	Associated SNP dbSNP ID	Associated SNP Internal ID	Virtual Disease Status
CDH1	Gene 1	rs12920590	278	Associated
CDH10	Gene 2	rs13188622	406	Associated
CDH11	Gene 3	rs1345863	1243	Associated
CDH19	Gene 4	rs12955865	2475	Associated
PCDH1	Gene 5	rs713079	2745	Associated
PCDH10	Gene 6	rs28401388	3125	Associated
PCDH17	Gene 7	rs12583519	3177	Associated
PCDH19	Gene 8	rs7060516	3639	Associated
PCDH8	Gene 9	rs4456399	3786	Unassociated
CDH2	Gene 10	rs11083166	3904	Unassociated
CDH22	Gene 11	rs2425729	5049	Unassociated
CDH5	Gene 12	rs35143	5555	Unassociated
CDH6	Gene 13	rs34510977	5757	Unassociated
CDH7	Gene 14	rs11662394	6832	Unassociated
CDH9	Gene 15	rs1007588	7134	Unassociated

Table 6.1:

The entities are arranged in concentric layers with the virtual disease being in the center surrounded by 8 genes, whose gene expressions determine the disease status via the functional relationship described in Equation 6.1, and with the SNPs, which regulate the gene activity, arranged in the outer layer.

We chose different geometrical shapes to represent the various entities of the interaction network:

- ellipsis (disease status)
- square (gene expression)
- diamond (SNP)

Associations in the interaction network of the simulated eQTL dataset are depicted with edges.

6.1.2 Evaluation approach

For assessing the performance of our MDL-analysis program qMAP, we made a benchmark study using the simulated eQTL dataset as a testing ground. The benchmark includes state-of-the-art information theoretic analysis frameworks MI-KDE [19] and MIC [49] as well as the popular tool for genetic analysis called PLINK [66].

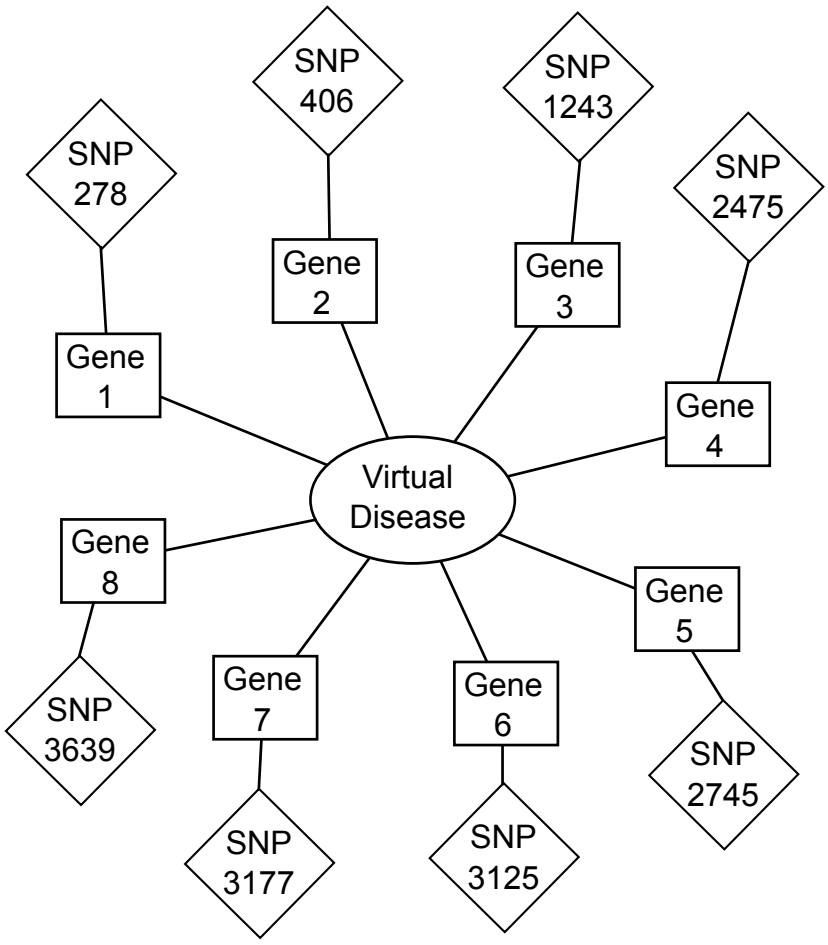


Figure 6.1: The interaction network characterizing the virtual disease in the synthetic eQTL dataset, which consists of 8 genes together with their primary SNP regulators.

The two analysis frameworks which are based on information theory are: The dependency measure of mutual information using a custom implementation of the formalism of Dawy et al. [19] together with kernel density estimators for obtaining the probability density functions necessary to calculate the mutual information. The novel framework MIC for *"explorative data analysis"* of Reshef et al. [49] which tries to maximize the mutual information between two pairs of variables through a grid optimization approach.

Since the mutual information is calculated based on kernel density estimates, we will abbreviate this method as MI-KDE. To summarize the abbreviations we use for the various programs that are evaluated in this study we present them in Table 6.2.

Program Name	Program Description
qMAP	MDL-based eQTL analysis framework minimizing the NML-code
MI-KDE	Uses kernel density estimates to calculate the mutual information
MIC	Calculates the maximal information coefficient
PLINK	Used by geneticists for case-control studies

Table 6.2: The 4 evaluated programs in this study which were benchmarked using the simulated eQTL dataset.

For all four analysis approaches we want to assess some basic performance measures. First of all the detection rate of the algorithms, which is their ability to detect the correct associations between SNPs and gene expressions in the simulated eQTL dataset. Moreover, we were also interested in evaluating the performance of each method for various sample sizes. This could help medical practitioners assess their confidence in the obtained analysis results.

If the sample size is small and a method shows a low detection rate for that sample size, then the obtained results should be interpreted with care, because they might not be very reliable. On the other hand, it is expected of a good analysis framework to approach the correct solution with increasing sample size. Naturally we would expect that the more data we feed into our analysis program the more precise the results about detected gene-SNP associations should become. With increasing sample size a good method should deliver more robust and reliable results.

That is the reason why we performed sub-sampling of the synthetic eQTL dataset, creating a sequence of 38 simulated eQTL datasets with increasing sample sizes from $N = 10$ to $N = 380$ with 20 subsamples for each N, resulting in a total of 760 synthetic datasets available for benchmarking.

Although the distribution of cases and controls regarding the virtual disease was a little bit skewed towards the direction of unaffected patients in the synthetic eQTL dataset originally created by Bartlett and Ray [8], the subsamples were compiled to contain an equal number of both affected and unaffected virtual patients from the original simulation eQTL dataset of [8].

The range of sample sizes, i.e. the virtual patients making up the simulated eQTL dataset, reaches from a total of 10 subjects to a total of 380 virtual patients. The step size between sample sizes was chosen to be 10 so that the sequence of sample sizes N becomes

$$N = [10, 20, 30, \cdots, 360, 370, 380]. \tag{6.2}$$

For each sub-sample of size N, we created a multitude of 20 simulated eQTL datasets by random sampling the original simulated eQTL data of Bartlett and Ray [8].

To ensure the balance of affected and unaffected cases, $\frac{N}{2}$ patients were randomly chosen from the pool of 307 unaffected cases and $\frac{N}{2}$ patients were randomly chosen from the pool of 193 affected cases, with the result that each sub-sample eQTL study consists of a total of N virtual patients.

During the sub-sampling process, where virtual patients were drawn from the pool of available candidates, the sampling of patients was done without replacement. This means that in each subsample no two identical patients exist.

Thus, the 20 mini-eQTL studies for each sample size N will help us in determining the detection rate, with the putative interpretation of accuracy or sensitivity and the assigned median ranks to the correct results in a ranked list, with the putative interpretation of specificity, for each analysis program. Furthermore, the simulated data also enable us to study the convergence behaviour of each algorithm, i.e. the question if an algorithm is able to detect the correct associations in eQTL data with increasing sample size and how fast it approaches that goal.

6.1.3 Evaluated analysis tools

We evaluated the analytical performance of the 4 frameworks by applying them to the simulated datasets. For each sample size, the algorithms had to extract SNP-gene transcript associations from the eQTL data for all randomly sampled sub-studies and report the strength of a detected association in a ranked list.

The program PLINK was run with default parameters in "quantitative association"-mode. In this mode PLINK outputs for each quantitative phenotype in the *.pheno*-file, in our case the expression patterns of the 15 genes, the association with every SNP found in the *.ped*-file and attaches this information to the annotation *.map*-file.

Each association test is accompanied by a p-value. The test results were ranked according to their p-value with the lowest p-value at the top of the ranked list indicating the best association detected between a SNP and a gene expression. For all 760 experiments ranked lists of p-values were obtained for every of the 15 genes with the 7555 SNPs.

The program MIC is relatively simple to run if the user adheres to the formatting procedure outlined on the program's website [78].

In order to make MIC accessible to eQTL analysis, a Python transformation script first converts the PLINK formatted *.ped*-genotype and *.pheno*-gene expression data into a suitable representation for the MIC software. To mimic the behaviour of PLINK, the gene expression patterns of each gene were combined in a matrix with the 7555 SNPs and written in a comma separated value list (*.csv*-file) to disk. Then it was indicated to the MIC software to compare the first column (containing the gene expression pattern) of the *.csv*-file to all other remaining

columns (containing the SNPs) and report the association strength in terms of the *maximal information coefficient*.

If the MIC-score is below a certain threshold, those SNP-gene pairs are not reported in the final output. Since the MIC-score indicates the strength of a detected association between SNPs and gene expression in the eQTL data, we ranked the MIC-scores, with the highest MIC-score indicating the best association test result between a SNP and a gene expression in the ranked list. Lower MIC-scores indicate weaker associations.

The program MI-KDE, which is our implementation of [19], accepts the same input files as PLINK. In contrast to PLINK and MIC, which can start analyzing the data immediately without further user input, the MI-KDE-program requires some manual adjustments before being able to process the data. Those adjustments are explained in the following paragraphs.

By inspecting the raw input data of the gene expression patterns, integration boundaries for the integral in Equation 3.6 have to be determined. Moreover, a sensible sampling rate for the continuous gene expression profile needs to be set manually before being able to obtain the probability density functions via kernel density estimates.

In our implementation Gaussian kernels are used, which are provided by the scientific Python library SciPy [79–81]. The SciPy kernel density estimator is able to automatically determine a proper kernel width for the estimation procedure according to the algorithms of [82–84]. Although the obtained probability density functions vary with different kernel widths and hence the obtained absolute mutual information values vary too, it was argued by Margolin et al. in [48] that the influence of this variation on ranked lists of mutual information values is not so large. For this reason, the standard implementation of kernel density estimates in SciPy seems to be sufficient for our analysis. While the absolute values of mutual information for SNP-gene pairs might change, the ranked ordering of mutual information scores remains for the most part unaffected according to [48]. Although not as extensive in scope as the experiments of Margolin et al. for their ARACNE software [48], initial experiments with ranked mutual information lists for eQTL data confirmed the observations made by Margolin et al. in [48].

Our implementation calculates the mutual information value in a different manner to the formula reported in [19] and instead uses the mathematical formula outlined in Equation 3.8 in Section 3.2. Nonetheless, it is still necessary to evaluate the integral in order to obtain the mutual information value for a SNP-gene pair. This is achieved using the trapz-algorithm of SciPy [79–81].

The above mentioned pre-analysis steps highlight some of the drawbacks when using mutual information as a measure of association strength in a mixed heterogeneous setting consisting of discrete random variables for SNPs and continuous random variables for gene expressions. These difficulties of a straightforward

application of mutual information to eQTL analysis were also a motivating factor in developing the MDL-based framework for extracting SNP-gene expression associations from eQTL data.

Despite the pre-processing drawbacks of MI-KDE, mutual information in itself is still a very powerful measure of association as will become clear in the subsequent results Sections 6.1.4 and 6.1.5.

For all the 15 genes in the dataset, we created ranked lists of mutual information using MI-KDE. In contrast to the p-values of PLINK and MIC-scores outputted by the MIC algorithm, ordinary mutual information values can be interpreted in an information theoretic way, which means that we can quantify the amount of information shared between each SNP and the respective gene expression pattern as well as evaluate in a quantitative manner how much information each SNP contributes to the gene expression pattern. This enables users of the analysis program MI-KDE to gauge how much they learn about the disease when looking at a patient's genotype and gene expression patterns.

The program qMAP is able to immediately analyze eQTL data in the same fashion as PLINK and MIC. Because qMAP was mainly designed for extracting useful information out of eQTL datasets it naturally parses PLINK formatted eQTL files. Like MIC, no further user input is required, making qMAP an efficient, easy-to-use analysis tool.

Adhering to the MDL-philosophy (see e.g. [50, 60]), all necessary parameters required for the eQTL analysis are obtained from the dataset automatically.

The obtained MDL-scores show the association strength of a SNP-gene transcript pair, measuring the influence a SNP has on the transcription activity of a gene. As is the case with MIC-scores, higher MDL-scores indicate stronger influences while smaller MDL-scores indicate weak influences.

MDL- and MIC-scores are not only obtained using different algorithms, but also their interpretation is different.

MIC-scores are based on the assumption that by maximizing the mutual information between random variables via a grid optimization approach, associations between entities in a dataset are revealed. According to the MIC way of thinking, the bigger the amount of shared information between entities is, the stronger the association between those entities should be.

In contrast to MIC-scores, MDL-scores are based on the NML-codelength of the data. Via a dynamic grid optimization procedure it is attempted to compress the gene expression patterns as much as possible on the supposition that the gene expression is regulated by a SNP. Consequently, the SNP showing the strongest regulatory influence to the gene expression should provide a good predictive model regarding the expression profile and thus yield short NML-codelength of the data.

Association between two entities in an eQTL dataset implies underlying functional relationship or structure. One of the selling arguments of the MDL-principle is that NML-coding exploits this underlying structure in order to find a more compressed representation of the data [50, 60].

MDL-score and NML-codelength are inversely proportional in qMAP; higher compression rates resulting in shorter NML-codelengths yield higher MDL-scores. From this line of thinking it is concluded that SNP-gene pairs yielding high MDL-scores are implied to have a strong association and an underlying functional relationship between each other.

The four programs qMAP, MI-KDE, MIC, and PLINK were fed with the simulated eQTL data as input and tasked to output ranked lists showing the association strength for each SNP-gene pair.

6.1.4 Measurement of detection rates for correct SNP-gene transcript associations

Let us first take a look at the overall detection performance of the algorithms. Here, the detection performance is the ability of each algorithm to identify correct associations between SNPs and gene expression in the synthetic eQTL dataset for various sample sizes.

Since the output of each algorithm is a ranked list of detected associations, we define that a *true positive* association has been detected if and only if the association appears on top of the list, which means that it is the first "search result" a potential user sees on the list.

The *rank* of such an association result is said to be 1, i.e. the first result when counting from the top of the ranked list. Intuitively, like Internet search engines, good analysis algorithms should output relevant results of an eQTL association study on the first search page the user sees. To put the previous sentence in perspective, ideally, correct associations in eQTL data, i.e. SNPs that definitely have an influence and/or statistical relationship with the transcription activity of a gene, should be assigned a score that ranks them close to the top of the ranked list.

Since we defined a *true positive* or a *hit* as correctly identified associations ranked as the first top result, we obtain the *detection rate* (also known as *hit rate*) of each algorithm as the number of correctly identified associations in the sub-samples divided by the total number of sub-samples.

We used random sub-sampling with $N = 20$ for this purpose. The *detection rate* (or *hit rate*) measure can be used to measure the accuracy of the evaluated programs, with higher detection rates implying higher accuracies of the algorithm.

We evaluated the detection rate of each eQTL analysis algorithm, i.e. its ability to extract correct associations between SNPs and gene expressions out of eQTL data.

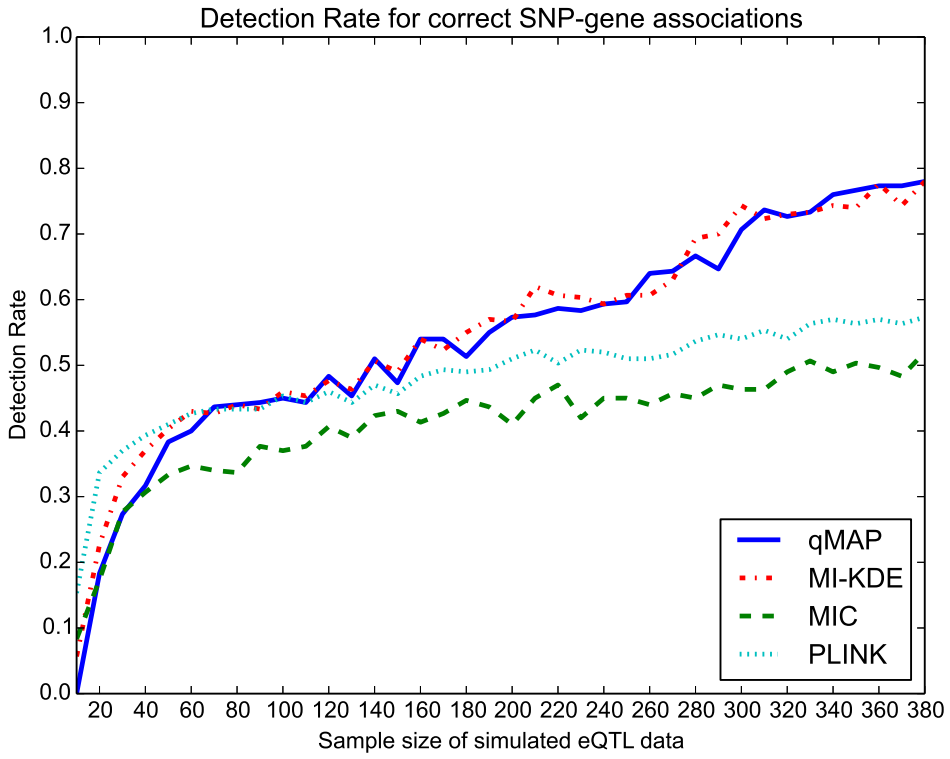


Figure 6.2: The overall detection rate of each algorithm for correct SNP-gene transcript associations in the simulated eQTL dataset depending on the sample size.

As can be seen in Figure 6.2, which depicts the detection rate of each tested algorithm for varying sample sizes, both the qMAP and MI-KDE algorithm outperform PLINK and MIC for increasing sample sizes. MI-KDE and qMAP almost deliver comparable detection rates. Both MI-KDE and MDL-qMAP more or less assign equal ranks to the associations detected in the eQTL data and consequently display a similar performance.

In contrast, PLINK seems to be geared to better deal with smaller sample sizes, because up to a sample size of 70 it shows, despite being low, the best detection rate among the tested algorithms. Yet, for sample sizes in that range the detection accuracy for all algorithms is less than 50%.

For eQTL studies containing around 70 to 110 patients, PLINK, MI-KDE and MDL-qMAP show a similar performance in terms of detection rates. In that

operating range the ability to detect correct associations in eQTL data is almost equal regardless which of the three algorithms is used.

The better performance of information theoretic analysis methods begins to surface for eQTL studies consisting of at least 110 samples according to our synthetic dataset. This is the point were a pure information theoretic measure like mutual information and our MDL-association measure begin to diverge from classical but popular approaches like PLINK and state-of-the-art association measures like MIC.

An interesting observation in Figure 6.2 is the similar performance of qMAP and MIC for small sample sizes, especially for eQTL studies with only up to 40 patients. The low detection rate and the slope in the increasing detection rate might be attributed to the grid optimization procedure both approaches employ in obtaining their final association strength scores, the MIC-coefficient for the MIC-algorithm and the MDL-score based on the NML-codelength for the qMAP program.

Although both algorithms use a completely different approach for optimizing the grid partitioning and a different optimization criterion, the resulting performance regarding the achievable precision is comparable. A hypothesis for this behaviour is presented in the following paragraph.

Regarding the dynamic programming algorithm of the KM-method for obtaining the MDL-optimal grid layout, an explanation for the low performance of qMAP can be given.

For small sample sizes only a low grid resolution is attainable via the MDL-principle in general and the KM-algorithm in particular. Because finer grid resolutions yield more detailed probability densities, in turn described by more complex statistical models of the multinomial model class, the MDL-principle adjusts the attainable model complexity depending on the available sample size. As a consequence, smaller sample sizes do not allow instantiations of complex models and hence the KM-grid optimization algorithm cannot distinguish the various gene expression patterns of the eQTL dataset. Consequently, lower grid resolutions as dictated by the MDL-principle fail to distinguish gene expression patterns and genetic profiles. Therefore, a lot of SNP-gene expressions are assigned equal or similar MDL-scores for association strength and thus, the correct SNP-gene expression is lost within the noise of the low grid resolution.

The MDL-grid resolution can be imagined as the resolution of a digital image. A correct SNP-gene expression association in an eQTL dataset can be thought of as a detail of the image we wish to see. If the resolution of the image is low, then the area containing the interesting information we wish to know is blurred and consequently, we cannot uncover that interesting part of the image.

In contrast, if we see a high-resolution image, it is possible to discern even the tiniest details. The same is true for the KM-grid optimization algorithm and the

NML-codelength of SNP-gene associations. With larger sample sizes, we obtain better grid resolutions that allow us to distinguish more gene expression and genotype patterns from each other, thus leading to the correct identification and recovery of the associations making up the interaction network in the eQTL data.

This observation directly leads us to another property of a good analysis algorithm. With increasing sample size, it is desirable that an analysis algorithm converges to the correct result, which in our case are the correct SNP-gene associations of the interaction network in the eQTL data. Judging from Figure 6.2, both the MI-KDE and the MDL-qMAP algorithm show signs of such a behaviour. With increasing sample sizes their detection rates get better and better whereas PLINK and MIC seem to plateau and then increase their detection rates only slowly.

Another interpretation of the detection rate curves in Figure 6.2 is the learning rate. Not only do we wish an algorithm to converge to the correct solution with increasing sample sizes, but we also want to arrive at that solution fast, using as less samples as possible.

Current experiments using the maximum available sample size in our synthetic eQTL dataset cannot forecast if PLINK and MIC will attain improved detection rates with much larger sample sizes. Projections from Figure 6.2 suggest that this might not be the case.

What is clear from Figure 6.2 is that MI-KDE and qMAP have faster learning rates than PLINK and MIC. Their detection rate (accuracy) improves much faster with increasing sample sizes when compared to PLINK and MIC. Since the learning rate can be defined as the slope of the curves in Figure 6.2, we identify three learning stages.

For small sample sizes all algorithms show a huge improvement in detection rate (accuracy) when slightly increasing the sample size. Then follows a stabilization plateau, where adding more samples to the eQTL data set does not pay off in terms of achievable detection rate until we reach a point of divergence. At that point the algorithms MI-KDE and qMAP show a better learning behaviour than PLINK and MIC. Increasing the sample size of an eQTL study pays off when analyzing the data with either MI-KDE or qMAP, because the detection rate, i.e. the accuracy of the algorithms, increases. While PLINK and MIC learn and approach the correct solution slowly, MI-KDE and qMAP display better learning performance, i.e. have faster learning rates, and thus converge to the correct solution more quickly.

An interesting property of the simulated eQTL dataset is the separation of strong and weak SNP-gene interactions. For the 8 genes associated with the virtual disease, the effect of each SNP on the gene expression pattern was simulated to be rather subtle and weak, whereas the non-disease associated gene expressions have strong ties with their primary regulatory SNPs. This property enables us to

check the ability of the analysis algorithms to uncover subtle effects. Therefore, we split-up the measured detection rate into a group containing strong effects, which consists of the 7 gene expressions not associated with the disease and into a group of weak effects, which is comprised of the 8 gene expressions associated with disease.

Consequently, we obtain detection rate performance measures for the 4 tested algorithms for their ability to detect both subtle and weak effects as well as strong influences of SNPs on gene expression patterns.

The detection rate results for weak associations are given in Figure 6.3 whereas the results for strong associations are depicted in Figure 6.4.

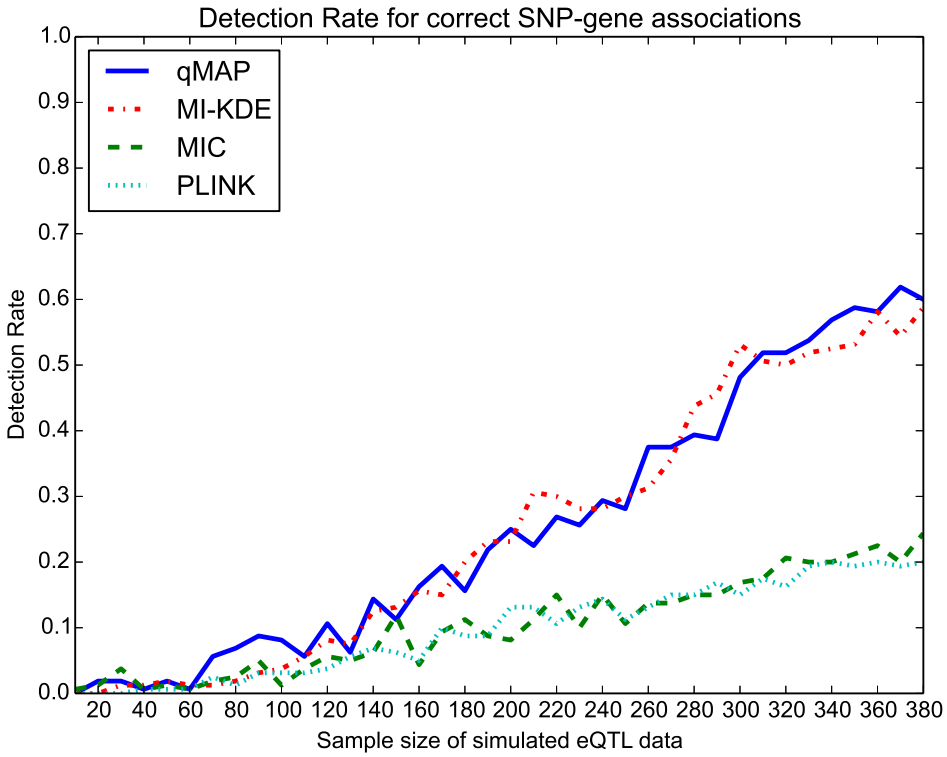


Figure 6.3: Each algorithm's detection rate of correct SNP-gene transcript associations depending on the sample size. Results for the 8 disease associated genes are shown whose primary regulatory SNPs exhibit weak control on the gene expression.

When comparing Figure 6.3 with Figure 6.4 an interesting effect emerges regarding the characteristic of each algorithm.

The first striking but expectable observation is that the overall achievable detec-

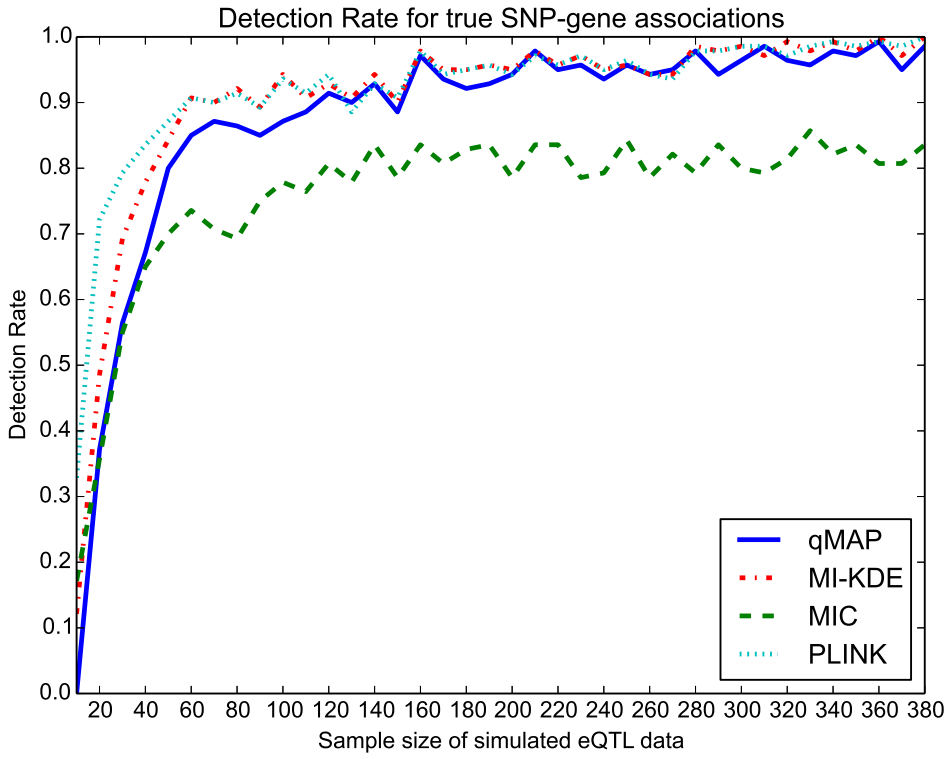


Figure 6.4: Each algorithm's detection rate of correct SNP-gene transcript associations depending on the sample size. Results for the 7 disease unassociated genes are shown whose primary regulatory SNPs exhibit strong control on the gene expression.

tion rate is much higher for strong associations between SNPs and gene expression patterns. Strong effects are easier to detect by any algorithm and thus all 4 algorithms can recover correct associations of the interaction network in eQTL data with a high detection rate.

For relatively small sample sizes from 10 to 60 samples, the detection rate of all algorithms rapidly increases. At an eQTL study with only 10 participants, PLINK has the highest detection rate with 32.9% followed by MIC with 17.1%, MI-KDE with 12.1% and 0% for qMAP. Nonetheless, for a sample size of 60 the achieved detection rate values are from best to worst; 90.7% for MI-KDE and PLINK, 85.0% for qMAP, and 73.6% for MIC.

Afterwards, the detection rate values stabilize and only improve slightly but steadily with qMAP, MI-KDE and PLINK having faster learning rates than MIC. Thus, all algorithms can correctly identify the strong SNP-gene interactions with increasing sample size reaching a final detection rate of 98.6% (qMAP), 100.0%

(MI-KDE), 100.0% (PLINK), and 83.6% (MIC) for a sample size $N = 380$. As can be seen in Figure 6.4 the achieved detection rates of the three algorithms qMAP, MI-KDE and PLINK are almost similar, which shows that the performance of the two information theoretic analysis approaches and the classical statistical approach of PLINK is comparable.

In contrast to the result for strong associations is the performance for much weaker effects. Not only is the detection rate lower than for strong effects, but also the behaviour of each algorithm is different.

When comparing the performance for large sample sizes (e.g. 380 samples), strong effects have detection rates of around 99% (cf. Figure 6.4) whereas weak effects can only be detected at a much lower rate of around 59% (cf. Figure 6.3).

It should also be noted that in order to achieve modest detection rates for weak SNP-gene expression associations large sample sizes are needed. Even a study that is comprised of 260 samples detects the correct associations in the simulation 37.5% (qMAP), 31.3% (MI-KDE), 13.8% (MIC), 13.1% (PLINK) of the time, while for strong associations the detection rates are 94.3% for qMAP, MI-KDE, PLINK and 78.6% for MIC.

While the detection rate for weak effects remains below 40% for all four algorithms until sample size 280, qMAP and MI-KDE show a stark increase in detection rate. On the other hand, although the detection rates improve for both PLINK and MIC as well, they remain well below those of qMAP and MI-KDE. PLINK and MIC achieve detection rates of 20.0% and 24.4% respectively for $N = 380$. qMAP and MI-KDE have detection rates of 60.0% and 58.8% for that sample size, more than triple the detection capability of PLINK.

For sample sizes 60 to 110, MDL qMAP is the algorithm that shows the highest ability for detecting putatively weak effects in eQTL data and for larger sample sizes outperforms PLINK and MIC, and furthermore displays an improved detection rate over MI-KDE.

This result shows that qMAP has a performance comparable to those of PLINK and MI-KDE for strong associations in the eQTL data and an improved ability over the other algorithms to identify correct associations in the data even when the influence of the SNP on the gene expression is subtle.

As a consequence, qMAP can help physicians to uncover more information from eQTL data. Compared to PLINK, MI-KDE and MIC, the reconstructed interaction network that depicts the associations between SNPs and gene expressions in the eQTL data is more accurate and includes also subtle effects that would have been missed by other algorithms.

To summarize the above results about the detection rate performance of the 4 tested algorithms, it can be stated that the two information theoretic approaches,

namely MI-KDE and qMAP outperform PLINK and MIC with respect to detecting the correct SNP-gene associations hidden in the simulated eQTL dataset.

Moreover, they show a faster learning rate than PLINK and MIC, which means that MI-KDE and qMAP need on average fewer samples in order to identify the correct associations.

Let us summarize the detection rate performance of the 4 tested algorithms by presenting the detection rates in percent at select sample sizes in Table 6.3.

Sample Size	qMAP	MI-KDE	MIC	PLINK
30	27.3	33.0	27.7	37.0
60	40.0	43.0	34.7	42.7
120	48.3	47.7	40.7	46.0
200	57.3	56.7	41.0	51.0
250	59.7	60.7	45.0	51.0
300	70.7	74.3	46.3	54.0
350	76.7	74.0	50.3	56.3
380	78.0	78.0	52.0	57.3

Table 6.3: Detection rates of correct SNP-gene transcript associations in the simulated eQTL dataset depending on the sample size. (Displayed values are in percent [%]).

The advantage of qMAP over MI-KDE is that not only are all necessary parameters automatically obtained from the given data, so that no further user input is required, but qMAP has better detection rate results for weak SNP-gene associations than MI-KDE. For strong associations between SNPs and gene expression MI-KDE and MDL qMAP have equivalent detection rates.

The tables containing the numerical values of the detection rates for each experiment can be found in Appendix A.

For the overall detection rate regarding the entire interaction network consisting of 15 genes please refer to Table A.1. Results for the 8 disease associated genes with weak effects between SNP-transcript associations are depicted in Table A.2. Detection values for SNPs exerting a strong influence on the gene expression activity, which is the case for the 7 disease unassociated genes, are shown in Table A.3.

Moreover, the detection rates are plotted separately for each of the 15 genes in Appendix A.

6.1.5 Measurements of assigned ranks for correct SNP-gene transcript associations

In this section we will evaluate another behaviour of the eQTL analysis algorithms, namely how they rank correct associations between SNPs and gene expressions in the eQTL dataset.

Even if an algorithm does not output the correct association on the top of the list, the assigned rank can be used to assess how well each algorithm can distinguish true positive associations from false ones or its ability to separate the signal containing the correct SNP-gene pair from the rest of the data, which could be referred to as noise.

Thus, the assigned rank of the correct SNP-gene pair assesses the specificity of each algorithm.

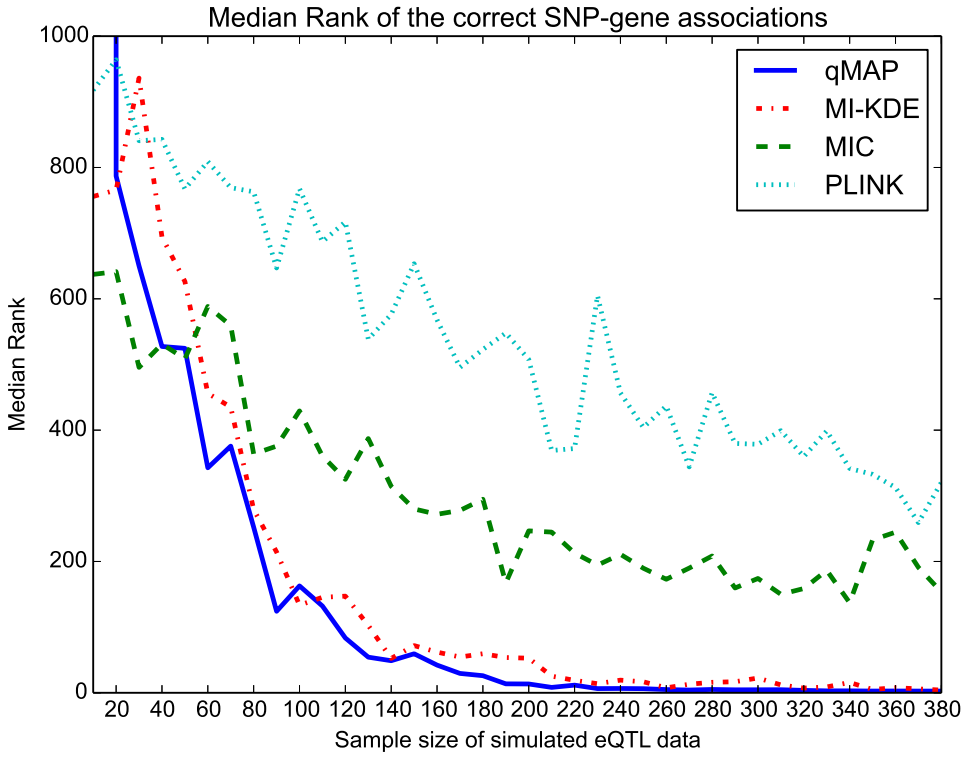


Figure 6.5: The median assigned ranks as a specificity measure obtained via the ranking of correct SNP-gene transcript associations in the simulated eQTL dataset depending on the sample size.

At each sample size, the ranks of the correct SNP-gene transcript pair for all 15

genes in the random sub-samples of the simulated eQTL dataset were recorded in a list and hence the median rank was calculated and plotted in Figure 6.5.

It can be observed in Figure 6.5 that for all 4 algorithms the assigned median rank, i.e. the definition of the algorithm's specificity in this context, increases with increasing sample size.

Yet, the algorithms qMAP and MI-KDE show a much higher affinity to rank the correct SNP-gene associations towards the top of the list than PLINK and MIC. For a sample size $N = 380$, the median rank of the true association is 3 for qMAP, 4 for MI-KDE. 154 for MIC, and 320 for PLINK. At lower sample sizes, e.g. $N = 140$, the median ranks are 49 for qMAP, 52 for MI-KDE. 314 for MIC, and 576 for PLINK.

This result shows, that although PLINK shows higher detection rates for smaller sample sizes than qMAP and MI-KDE, its assigned median rank to correct interactions is always lower than that of the information theoretic algorithms. Even for comparable detection rate values that are obtained in the range of 60 to 110 samples, qMAP and MI-KDE far outperform PLINK with regard to the median ranking of correct results, i.e. the ability to separate correct SNP-transcript signals from noise and accordingly assign a high rank to them in a ranked list of putative associations. This behaviour can be verified by comparing Figure 6.2 with Figure 6.5.

As has already been mentioned in Section 6.1.4, at very low sample sizes the grid optimization technique of the KM-method has a too low resolution so that it is not possible to obtain meaningful results and therefore the median rank at sample size 10 cannot be displayed for qMAP in Figure 6.5. From $N = 20$ onwards median ranks of qMAP are shown in the figures.

The faster learning rates of qMAP and MI-KDE over PLINK and MIC leads to higher specificity values at lower sample sizes. This characteristic of the information theoretic approaches can be observed in Figure 6.5 with the curves of qMAP and MI-KDE approaching faster the ideal rank of 1 for correct SNP-gene expression associations.

Since the performance gap regarding the median ranking is quite huge between qMAP and MI-KDE on the one side and PLINK and MIC on the other side, a zoomed version of Figure 6.5 highlights the better median ranking results of qMAP over MI-KDE.

The median ranks for sample sizes ranging from $N = 150$ to $N = 380$ are displayed in Figure 6.6, but only the results for qMAP and MI-KDE can be seen.

Not only assigns qMAP lower median ranks to the correct SNP-gene pair association results, but it does so much faster than MI-KDE. Therefore, qMAP displays a higher specificity than its competitor MI-KDE. The grid optimization based on the KM-method in conjunction with NML-coding shows higher aptitude in

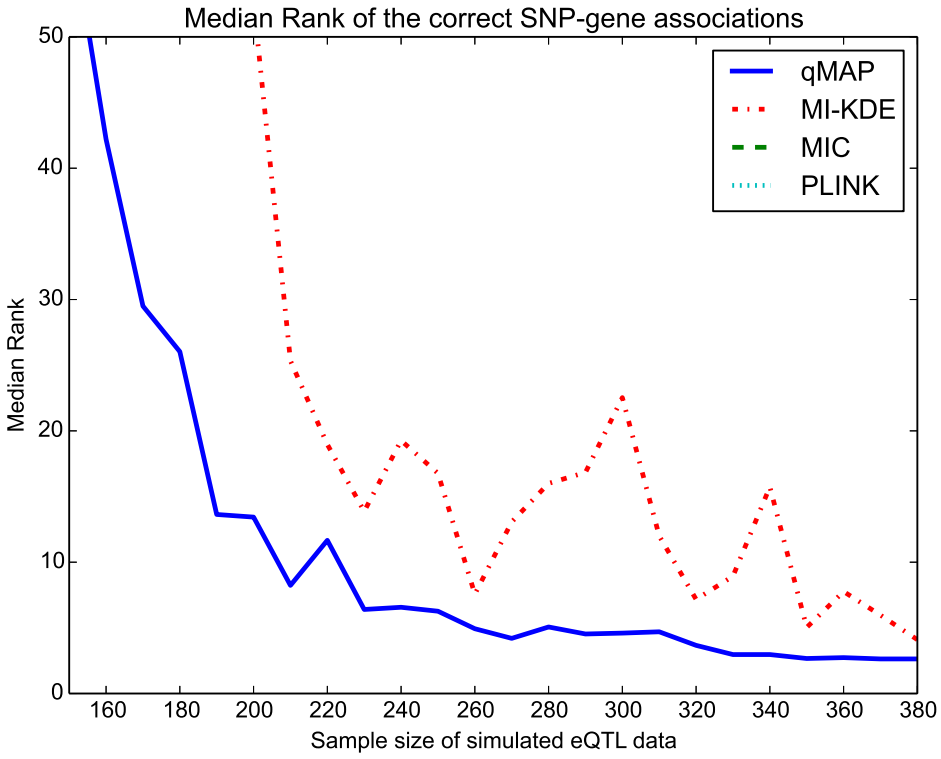


Figure 6.6: The median assigned rank of the correct SNP-gene expression pair in the simulated eQTL data depending on the sample size. Zoomed-in for larger sample sizes in order to show the performance difference between qMAP and MI-KDE.

discerning true signals, i.e. correct SNP-gene transcript associations, from background noise in eQTL data, yielding the best overall specificity for all tested algorithms.

Another observation derived from Figure 6.6 is that the assigned median ranks of the qMAP program do not fluctuate when compared to MI-KDE. An explanation for this behaviour might be, that the NML-coding via KM-optimization is more robust than other methods. This robustness of qMAP is an important asset.

For all the 15 genes of the simulated eQTL dataset the median ranking of the correct associations between SNP and gene transcript in the ranked list as outputted by the analysis program for selected sample sizes is presented in Table 6.4.

We also separated the median rank results into the group containing the 8 genes with weak associations to their controlling SNPs and the group of 7 genes whose

Sample Size	qMAP	MI-KDE	MIC	PLINK
30	650	936	496	841
60	343	456	588	809
120	83	147	325	717
200	13	53	247	508
250	6	17	190	404
300	5	23	174	378
350	3	5	233	333
380	3	4	154	320

Table 6.4: Median ranks of correct SNP-gene transcript associations in the simulated eQTL dataset depending on the sample size.

gene expression patterns are controlled by a strong association with their respective SNPs. Again, like in the case for the detection rate performance measurements of the 4 tested algorithms, a stark gap in performance can be observed in Figures 6.7 and 6.8.

To begin with, the specificity performance of qMAP outperforms MI-KDE, PLINK, and MIC in both scenarios. This underlines the usefulness of qMAP as a competitive alternative for eQTL analysis.

Let us first study the behaviour of the 4 methods for weaker SNP-gene associations because it highlights the advantages of the two information theoretic methods qMAP and MI-KDE over the methods of PLINK and MIC.

For weakly associated SNP-gene pairs the median ranks interpreted as specificity performance for each of the 4 algorithms are shown in Figure 6.7.

Lower values indicate better rankings of the correct results in each algorithms output with the ideal ranking being at the top of the list which corresponds to a rank of 1. We can observe in Figure 6.7 that qMAP and MI-KDE achieve higher specificity results at lower sample sizes and attain very high specificity values for large samples. PLINK and MIC cannot reach the performance results of qMAP and MI-KDE. Although a trend is visible for both algorithms which shows that with increasing sample size the specificity improves, there is a huge performance discrepancy when compared to qMAP and MI-KDE. MIC has better specificity values than PLINK, but is still far away from the values achieved by qMAP and MI-KDE.

Moreover, the rankings for qMAP and MI-KDE are more consistent than those of PLINK and MIC for increasing sample sizes. This means that the variance of the sample size of an eQTL study does not have such a strong influence on the rankings of the correct SNP-gene associations for the algorithms qMAP and MI-KDE. Given an arbitrary sample size N and varying it by $\pm\varepsilon$ samples, the ranks assigned to the correct SNP-gene associations in the simulated eQTL dataset do

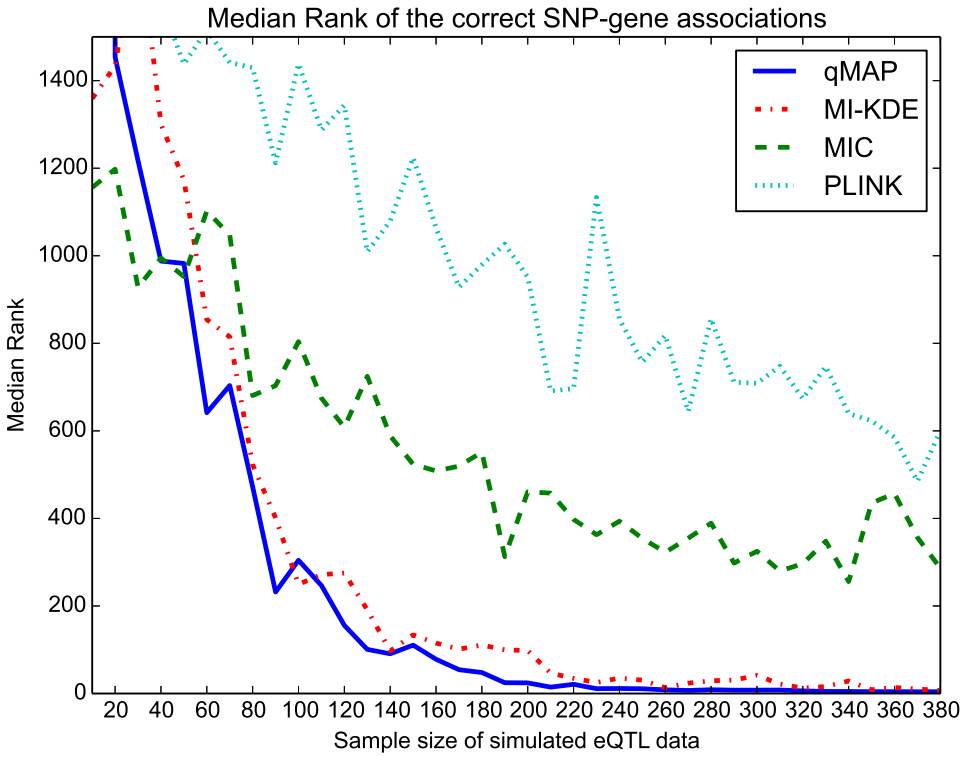


Figure 6.7: The median assigned rank of the correct SNP-gene expression pair in the simulated eQTL data depending on the sample size for weak associations between SNPs and gene expressions.

not vary much for qMAP and MI-KDE, whereas stronger fluctuations for PLINK and MIC can be seen in Figure 6.7.

Thus, the results of the methods qMAP and MI-KDE are less sensitive with regard to the total sample size and reveal a more consistent behaviour in ranking correct results than PLINK and MIC.

The higher specificity and more robust and consistent ranking of correct SNP-gene expression associations is an advantageous feature of the information theoretic algorithms qMAP and MI-KDE when compared to PLINK and MIC.

In particular, qMAP even displays better specificity than its direct competitor MI-KDE. This feature can only be visualized if we zoom into Figure 6.7 and concentrate at the median ranks for larger sample sizes. The zoomed-in area of Figure 6.7 is depicted in Figure 6.9.

The real advantage of the MDL qMAP method over all the other tested methods can best be seen in Figure 6.9, where it displays the best specificity performance

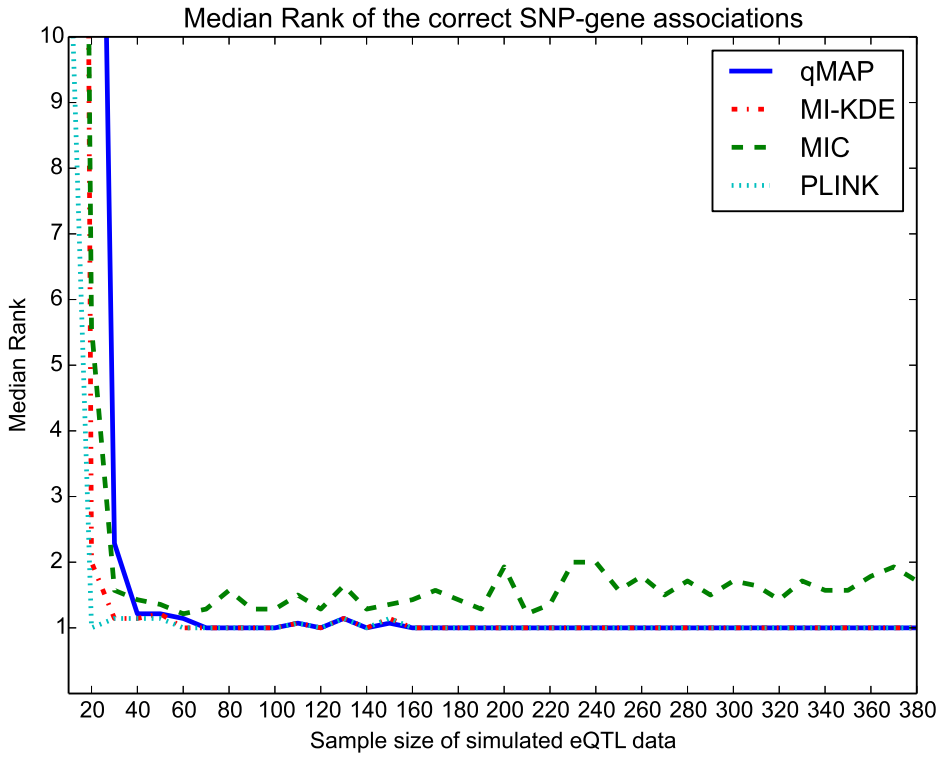


Figure 6.8: The median assigned rank of the correct SNP-gene expression pair in the simulated eQTL data depending on the sample size for strong associations between SNPs and gene expressions.

and continuously shows a robust and consistent ranking of the correct SNP-gene associations in the simulated eQTL dataset.

Despite the fact that qMAP shows outstanding specificity, robustness and consistency performance for weakly associated correct SNP-gene expression pairs, let us also compare the results for correct SNP-gene expression associations exhibiting a strong controlling influence of the SNP on the gene expression pattern. The median ranks for those associations are plotted in Figure 6.8.

In contrast to the weakly associated results of Figure 6.7 the median ranking of all algorithms are better for strong associations as can be seen in Figure 6.8.

All 4 algorithms start assigning high ranks to the correct SNP-gene associations quite quickly, with MIC performing much better in this case than for weakly associated results. For strong associations the specificity of MIC is comparable to those of MDL qMAP and MI-KDE, which implies that the MIC algorithm works better if an association between a SNP and a gene expression is strong. Also the

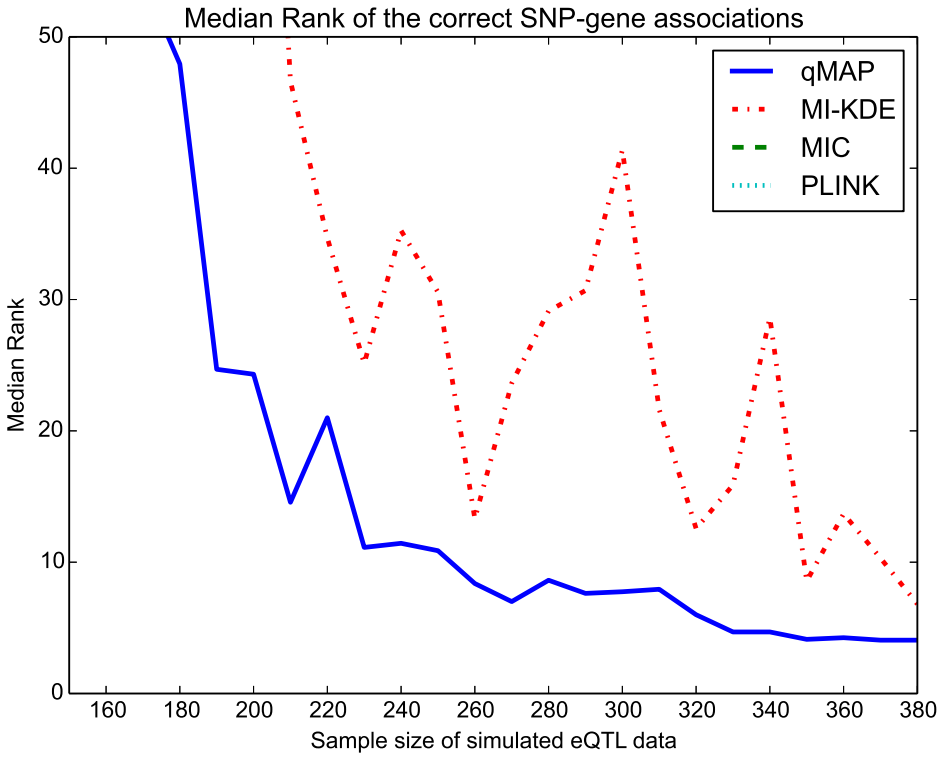


Figure 6.9: The median assigned rank of the correct SNP-gene expression pair in the simulated eQTL data depending on the sample size for weak associations between SNPs and gene expressions zoomed-in for larger sample sizes.

performance of PLINK is much better compared to weak associations. It should also be noted that all algorithms require much less samples in order to achieve good specificity values expressed as the ranking of correct SNP-gene associations. This result indicates that if the purpose is to look for strong associations only in eQTL datasets, one can achieve a certain specificity value with much less samples than would be needed, if one were to look for weak, subtle influences of SNPs on gene expressions.

If we compare the median ranking curve of each algorithm in Figure 6.8 with the detection rate curve in Figure 6.4, we can extract for each algorithm its performance characteristic and come to the conclusion that for strong effects in eQTL data, all algorithms have comparably well performance values for both detection rate and median ranking of correct associations.

For example consider an eQTL study containing 200 subjects and we want to extract SNPs from the data exerting a strong influence on gene expression patterns. Then, all algorithms have a detection rate of over 90% (except MIC) and

the correct result appears on top of the ranked list. The performance values are depicted in Table 6.5.

Program Name	Detection Rate (in [%])	Median Ranking
qMAP	94.3	1
MI-KDE	95.0	1
MIC	78.6	2
PLINK	94.3	1

Table 6.5: The detection rate and median ranking values for 4 algorithms applied on the simulated eQTL study with 200 samples.

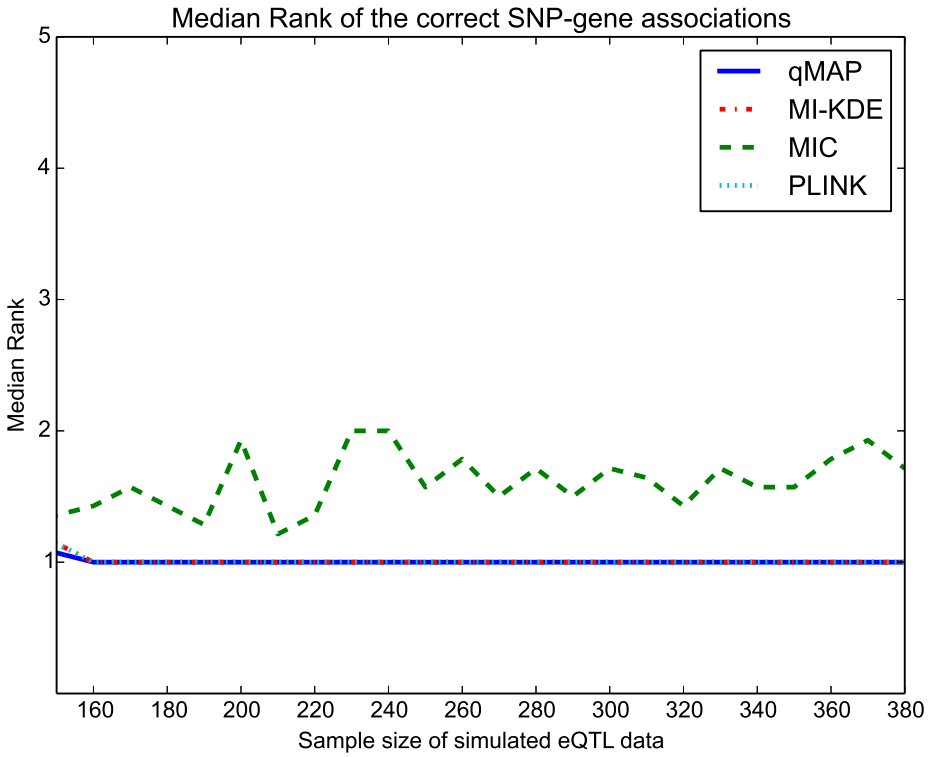


Figure 6.10: The median assigned rank of the correct SNP-gene expression pair in the simulated eQTL data depending on the sample size for weak associations between SNPs and gene expressions zoomed-in for larger sample sizes.

Tables depicting the measured median rank in each experiment are in Appendix A. For the entire dataset, the measured median ranks are shown in Table A.4. A separation between ranks assigned to weak associations for the disease associated

genes and ranks assigned to strong associations for the disease unassociated genes is done and the results are displayed separately in Table A.5 and Table A.6 respectively.

6.1.6 Recovered interaction networks

The ability to detect correct associations between SNPs and gene transcripts is directly related to the ability to reconstruct the interaction network.

In order to visualize the analysis results of the simulated eQTL dataset, the results of each algorithm are processed as follows. For the genes related to the simulated virtual disease, i.e. genes 1 to 8, an edge is drawn between the SNP and the gene if the algorithm identified that SNP to have the strongest association with the gene, i.e. the association strength of the SNP appears as the top scoring result.

It can be concluded from the detection rate analysis of Section 6.1.4 that the algorithms qMAP, MI-KDE, MIC, and PLINK will have completely different performances when it comes to the reconstruction of the interaction network.

Since the effects of the SNPs on the genes that are associated to the disease are very spurious and thus difficult to detect, we are going to report the reconstruction results when using the entire sample size of $N = 380$ patients. Even at that sample size, all algorithms are far from perfect from a comprehensive reconstruction of the network.

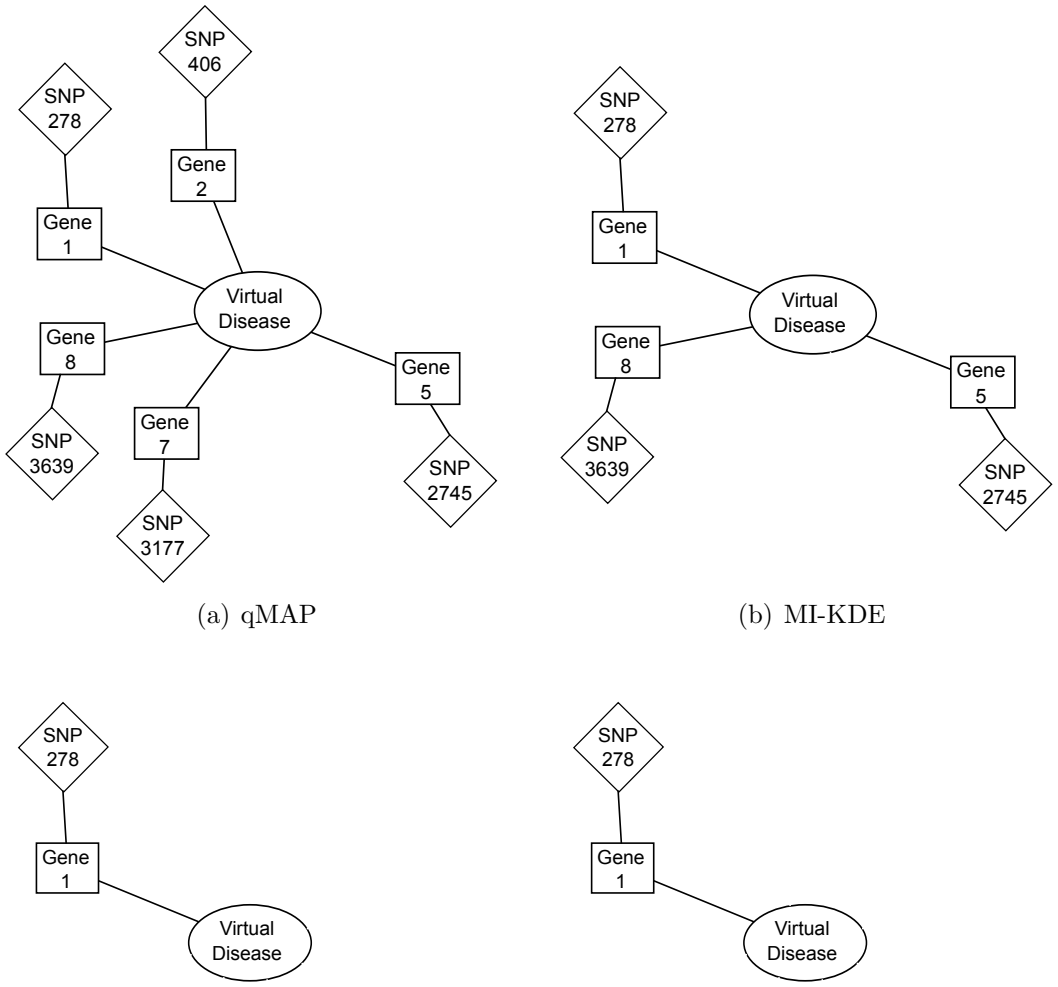
Correctly identified interactions will be reported as an edge between SNP and gene, whereas incorrectly identified interactions will be omitted in order to highlight the contrast between reconstructed and spiked-in interaction network.

Figures 6.11(a), 6.11(b), 6.11(c), 6.11(d) depict the visualizations of the reconstructed interaction networks for the algorithms qMAP, MI-KDE, MIC, and PLINK respectively. From the achieved detection rates of each algorithm, an edge was drawn between a SNP and a gene transcript if for that particular pair the detection rate was greater than the threshold of 70%.

The overall best performance delivers the qMAP algorithm because it achieves the most complete and accurate reconstruction of the interaction network that was spiked-in into the simulated eQTL data. At sample size $N = 380$ qMAP reliably identifies 5 of the 8 SNP-gene interactions which are associated with the virtual disease (detection rate yields of over 70%). This is the most comprehensive reconstruction amongst all the tested algorithm.

In second place comes the algorithm MI-KDE, detecting 3 of the 8 disease associated interactions.

PLINK and MIC have similar reconstruction ability but cannot achieve the performance of qMAP or MI-KDE. The reconstructed interaction network by MIC



(a) qMAP (b) MI-KDE

(c) MIC (d) PLINK

Figure 6.11: The reconstructed interaction network for $N = 380$ samples by the analysis program: (a) qMAP, (b) MI-KDE, (c) MIC, (d) PLINK

and PLINK can reliably identify the association between gene 1 and SNP 278, but for the other genes the detection rates are below the threshold of 70%.

Although showing a good detection rate for strong gene-SNP associations, PLINK's ability in reconstructing the interaction network that involves subtle functional impacts of the SNPs on the gene expression patterns that determine the pathogenesis of the virtual disease is quite limited. From the 8 genes whose expression

patterns decide if a patient gets the virtual disease or not, only the association between gene 1 and SNP 278 could be recovered while the other 7 were missed.

Especially, the connection between gene 3 and SNP 1243 is hard to detect for any of the four tested algorithms. The impact of the SNP on the transcription activity is too low to be confidently detected as it is missed by all algorithms.

The above results show that with the help of qMAP, the MDL-based analysis program for associating quantitative traits with discrete genotypes, it becomes possible to extract more useful information out of eQTL data. A more complete map depicting more correctly identified molecular mechanisms between SNPs and gene expression is more useful than an incomplete image of those interactions.

Even though the information theoretic MI-KDE approach is able to identify almost as much interactions as qMAP, the higher detection rates of qMAP make it possible to acquire a better understanding of the underlying interactions involved in disease by delivering a more complete and comprehensive map of the interaction network. This feature of qMAP is a real advantage when it comes to gaining a better understanding of disease by extracting more useful information out of the available data.

6.1.7 Analysis of a real human cortical eQTL dataset

Since qMAP has shown superior theoretical performance compared to existing algorithms, rivaled only by the information theoretic MI-KDE approach, it is time to demonstrate the utility of qMAP by applying it on a real dataset.

The website seeQTL [85] offers an online collection of links to resources hosting publicly available eQTL datasets. Via seeQTL the dataset of Myers et al. [33] was downloaded and re-analyzed using qMAP.

Before submitting the data to the analysis programs, genotypes and gene expression data were pre-processed according to the filtering rules that were employed in the original paper of Myers et al. [33].

SNPs with a minor allele frequency (MAF) of less than one percent ($MAF \leq 1\%$) were not further processed during the analysis. The classification of cis- and trans-effects of SNPs was done according to the definition that cis-acting SNPs are located at maximum $1Mb$ away from the $5'$ or $3'$ end of the gene or are directly located in the gene whose expression they influence. If the detected SNP does not adhere to the above definition of a cis-effect, it is classified as a trans-acting SNP (see Myers et al. [33]).

We focused on the SNPs that were reported in the study to be associated with gene expressions. In the study a total of 8 genes with 23 cis-acting SNPs were presented.

A particular focus was on the microtubule-associated protein tau (MAPT) gene in the study of Myers et al. [33], because there is mounting evidence that the gene expression pattern of MAPT is influenced by several SNPs residing in a haplotype block [86].

After obtaining the MDL-scores for each SNP-transcript pair, we confined our search to potential cis-acting SNPs around the MAPT gene by only reporting those MDL-scores whose SNP fall withing $1Mb$ from either the $5'$ or $3'$ end of the MAPT gene. All reported SNPs had on average an MDL-score of 154.6, among them the SNP $rs17571739$ of the original study with an association score of 154.4.

Since no fluctuations in the MDL-score for the reported SNPs could be observed, the hypothesis of Myers et al. [33] delivers an explanation for that phenomenon.

As was shown in [33], the SNP $rs17571739$ associated in cis with MAPT is in linkage disequilibrium with various other SNPs in an area of roughly $620kb$ in size and forms a haplotype block. Therefore, MDL-scores between SNPs located in that haplotype block and the target gene MAPT have similar values.

Although the MDL-scores of qMAP show subtle differences for the association strength of each SNP, the small variation in scores makes the reported SNPs virtually indistinguishable, which means that the cis-acting effect of that haplotype block on the MAPT gene expression cannot be traced to a single genetic variant yet.

Our re-analysis of the eQTL dataset with qMAP confirmed the validity of the original results initially reported by Myers et al. in [33]. Consequently, qMAP delivers consistent results with other state-of-the-art statistical analysis approaches.

For the remaining other genes, the concordance of results between qMAP and Myers' study was evaluated. It was confirmed that qMAP successfully detected the reported cis-acting associations. In general, if a detected cis-acting SNP lies in a haplotype block, qMAP assigns similar MDL-scores to SNP-transcript interactions originating from that haplotype block.

The analysis results for the cortical eQTL dataset are summarized in Table 6.6, showing a selection of the concordance between the results of Myers et al. [33] and qMAP.

Gene Name	Associated cis-acting SNP (reported by Myers et al. [33])	qMAP Detection	MDL-score
MAPT	rs17571739	positive	154.4
B3GTL	rs1005824	positive	164.6
SQSTM1	rs10277	positive	168.9
	rs1065154	positive	169.6
PTD004	rs10930638	positive	174.6
	rs10930654	positive	173.7
	rs11674895	positive	173.8
KIF1B	rs10492972	positive	163.0
	rs12120042	positive	162.5
	rs12120191	positive	162.5
HBS1L	rs1590975	positive	166.0
	rs2150681	positive	165.7
CHST7	rs760697	positive	147.6

Table 6.6: Comparison of the reported results of the cortical gene expression study by Myers et al. [33] and the SNP-gene transcript pairs detected by qMAP. SNPs residing in a haplotype block with similar MDL-scores were grouped together and represented by one SNP (for ease of comparison the representative SNP ID was chosen to be the same as in Myers et al. [33]).

7 Conclusion and Outlook

From ongoing research in genomics it is becoming apparent that phenotypic traits are not only the mere result of genomic variation which manifests itself via single nucleotide polymorphisms (SNPs) in a person's genome, but that a dynamic regulatory system that includes among other factors the extremely diverse landscape of the transcriptome, which is the realization space of gene expression activity.

By phenotypic trait we understand a multitude of an organisms's features, may it be the hair color, a disease, or even the cholesterol levels.

The relationship between SNPs and phenotypes have already been studied in depth by so called genome-wide association studies (GWAS).

Due to better sequencing technologies, the focus of scientific analysis is shifting towards studying the transcriptome, which promises to unlock more knowledge about disease and the molecular workings of the DNA. An approach which is trying to integrate both genotype and transcriptome data in order to improve our understanding of the underlying mechanisms of disease pathogenesis are expression quantitative trait locus (eQTL) studies.

Despite of the excellent work of GWAS, there remains an information gap in our understanding of how genomic variations actually influence or control the development of a specific phenotypic trait. This knowledge gap could be alleviated by studying the transcriptome because recent studies have revealed that there is a stronger link between a gene expression pattern and a phenotype than there is between a genotype and a phenotype.

This means that changes in the transcriptomic landscape have a more severe influence on a phenotype. Since GWAS found many phenotype associated SNPs that fall into non-coding regions of the genome, a hypothesis put forward by many geneticists is that those SNPs might actually regulate gene expressions which in turn alter the transcriptomic landscape leading to the onset of disease.

Therefore, it is important to discover those vital relationships between SNPs and gene expressions. Emergence of eQTL data offers a great opportunity to identify these interacting entities. SNPs which are found to be statistically associated with certain gene expressions and the transcriptomic landscape of a disease which consists of gene expression patterns associated with the disease, offers the possibility to increase our knowledge of the biological systems that underlie phenotype

development and disease pathogenesis.

For this purpose, we developed a novel information theoretic eQTL analysis tool in this thesis called qMAP (quantitative MDL Association Program) that identifies associations between SNPs and gene expressions based on the minimum description length.

Since it is important to acquire an accurate and complete landscape of the associations between SNPs and gene expression activity making up the interaction networks involved in disease susceptibility, qMAP is geared towards a high and robust detection rate of SNP-gene expression associations.

Compared to the analysis toolkit PLINK the detection rate of correct SNP-transcript pairs in the interaction network of a synthetic eQTL dataset improves by 20 percentage points to 78% when using qMAP. Consequently, qMAP delivers a more complete picture of the transcriptomic landscape and the associated regulatory SNPs that constitute the interaction network. It enables analysts to draw more knowledge out of eQTL data.

This dramatic improvement in detection ability is due to an analysis approach based on NML-coding that utilizes a dynamic grid optimization algorithm, originally invented by Kontkanen and Myllymäki, and referred to the KM-method here. In this thesis we further developed the excellent results of Kontkanen and Myllymäki by extending the KM-method to obtain a coding scheme for eQTL data, i.e. for a combination of discrete genotype random variables and continuous gene variation random variables.

Based on the NML-codelengths we obtained from the extended KM-method's dynamic grid optimization approach, we created an MDL-score which tells us the association strength between a SNP and a gene expression in the eQTL data.

The MDL-scores proposed in this thesis have an interpretation which is rooted in Rissanen's MDL-principle. The working hypothesis is that SNPs which regulate gene expressions could serve as a statistical model to predict the gene expression pattern. Therefore, according to the MDL-principle, those SNPs that yield short NML-codes of the gene expression, when used as a model, and have short NML-code self descriptions themselves, are prime candidates for regulating the gene expression and are said to be strongly associated with that gene.

To test the robustness and the performance of the novel algorithm, qMAP was employed to analyze a simulated eQTL dataset, created by Bartlett and Ray and gracefully made available to the author, that contains known statistical associations between genotype and transcriptome which make up the interaction network in the data. The obtained results were benchmarked against another array of potential analysis tools, including PLINK, the maximal information coefficient (MIC), and an implementation using kernel density estimates to calculate the mutual information between a discrete and continuous random variable (MI-KDE).

Vast improvement gains in detection rates could be observed when comparing qMAP to MIC and PLINK, yielding an average detection rate for all the entities making up the interaction network of the eQTL data of 78% for qMAP compared to 57.3% for PLINK and 52% for MIC when applied to the simulated eQTL dataset for sample size $N = 380$. Furthermore, the results between both information theoretic approaches MI-KDE and qMAP are consistent, yielding equivalent detection rates for $N = 380$.

In addition to the detection rates, the specificity of each algorithm was investigated and it was found that qMAP shows the highest specificity among all tested algorithms. It displayed a robust behaviour in continuously ranking the correct associations towards the top of ranked lists and proved to be a more sensitive method than MI-KDE.

When it came to reconstruct the interaction network of the virtual disease in the synthetic eQTL data, which consists of 8 genes and their primary regulatory SNPs, qMAP detected on average 5 SNP-gene pairs of the interaction network, outperforming MI-KDE with 3, MIC and PLINK with 1 detected interaction respectively in our simulation study.

We also applied qMAP to a real human cortical eQTL dataset that was created by Myers et al. and performed a re-analysis. Our results which were obtained by qMAP confirmed the findings of Myers. An interesting observation was that qMAP assigns similar MDL-scores to SNPs of a haplotype block.

The extension of qMAP to assign MDL-scores to joint interactions of SNPs on the transcriptome and the ability to include more data sources, like for example methylation patterns into the analysis, remains future work.

Nonetheless, in its current state, qMAP is able to extract equivalent or more information from eQTL data than other state-of-the-art approaches. The high detection rate, robustness, and sensitivity of qMAP make it a useful information theoretic tool for knowledge discovery in eQTL data.

A Figures & Tables

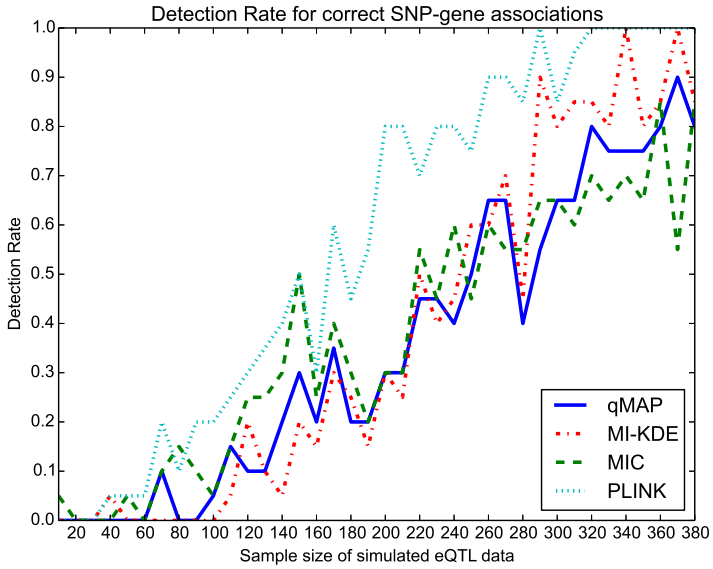


Figure A.1: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 1.

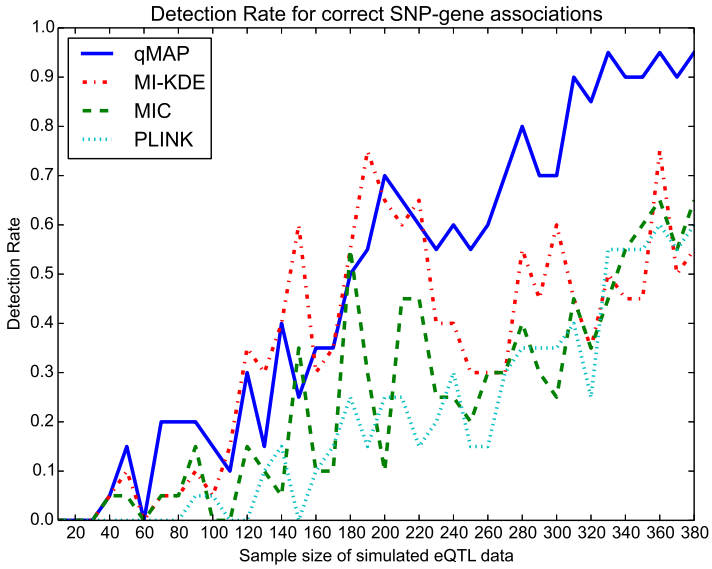


Figure A.2: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 2.

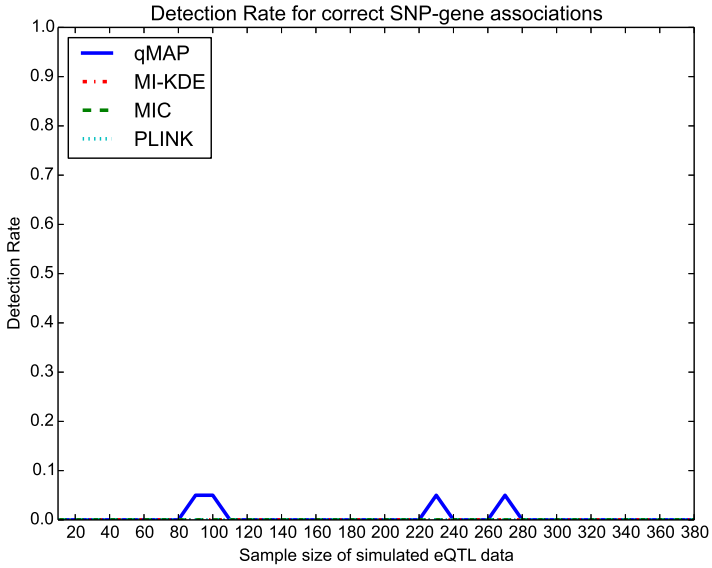


Figure A.3: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 3.

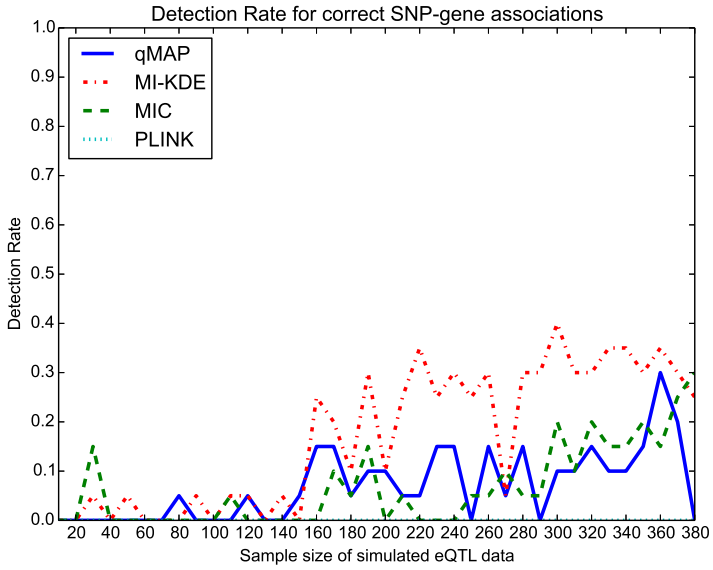


Figure A.4: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 4.

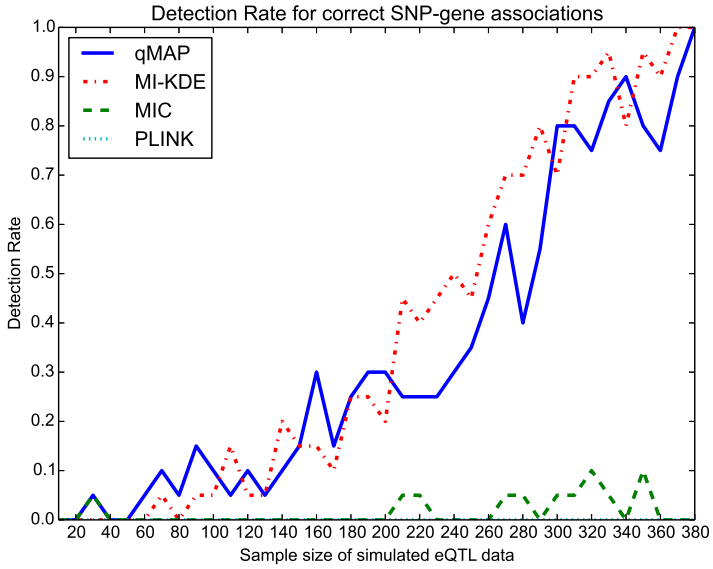


Figure A.5: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 5.

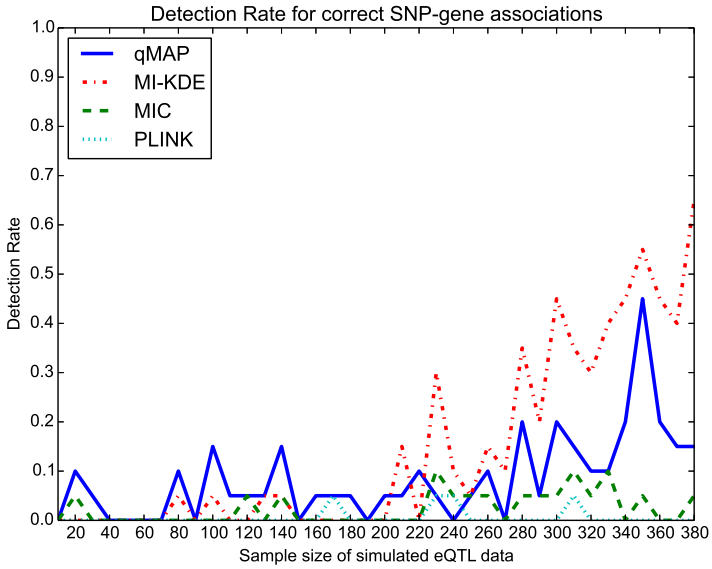


Figure A.6: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 6.

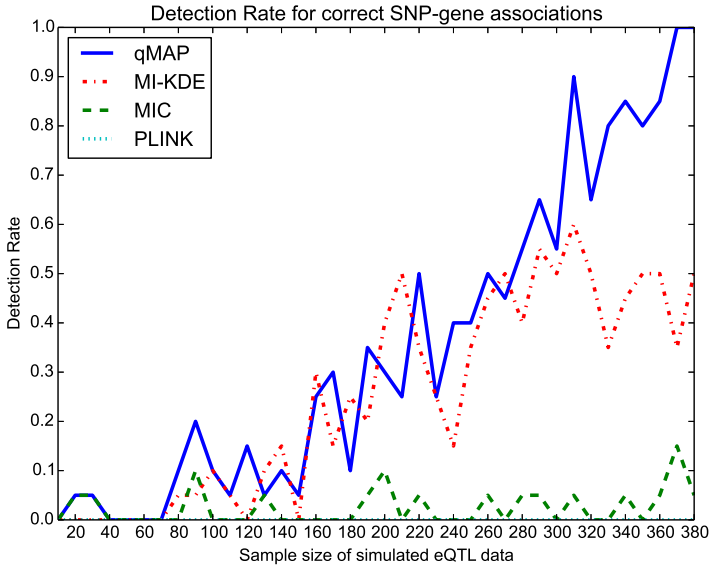


Figure A.7: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 7.

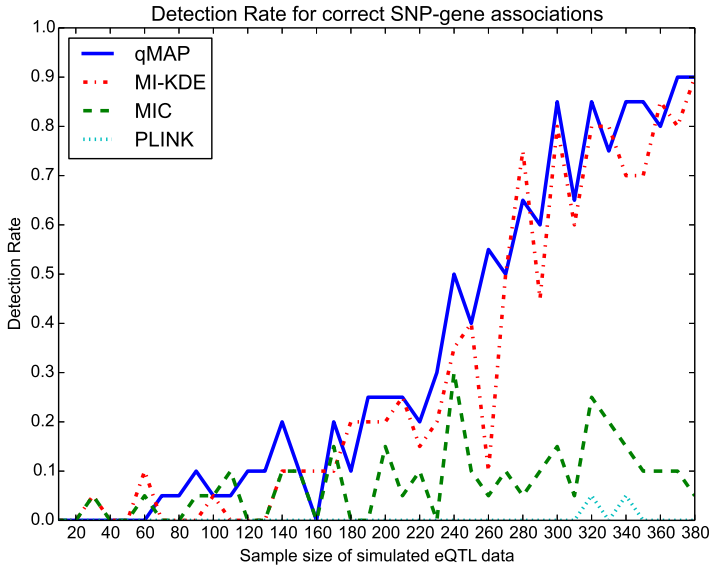


Figure A.8: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 8.

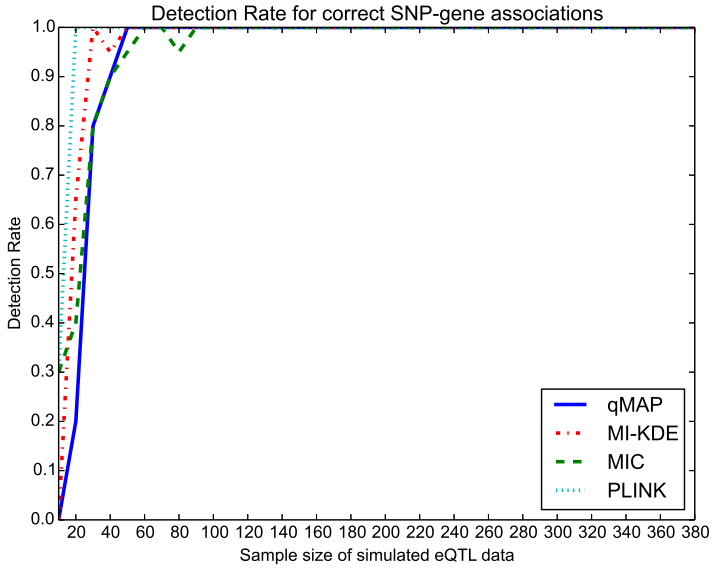


Figure A.9: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 9.

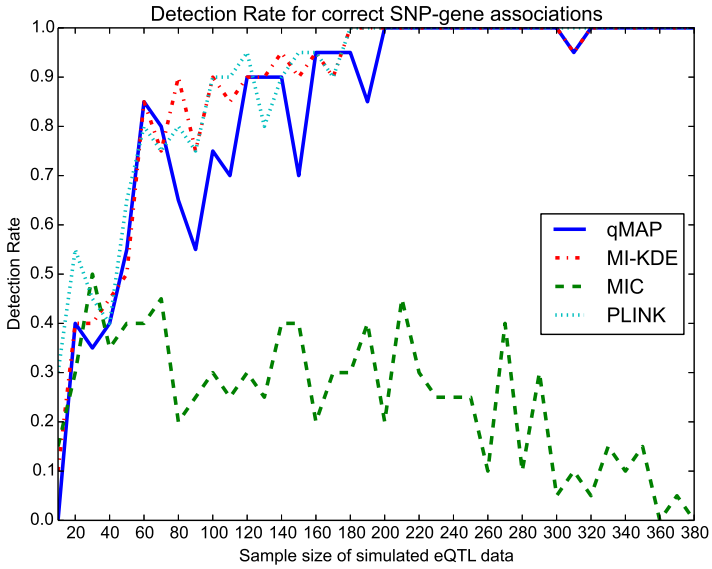


Figure A.10: The detection rate depending on the sample size of each algorithm correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 10.

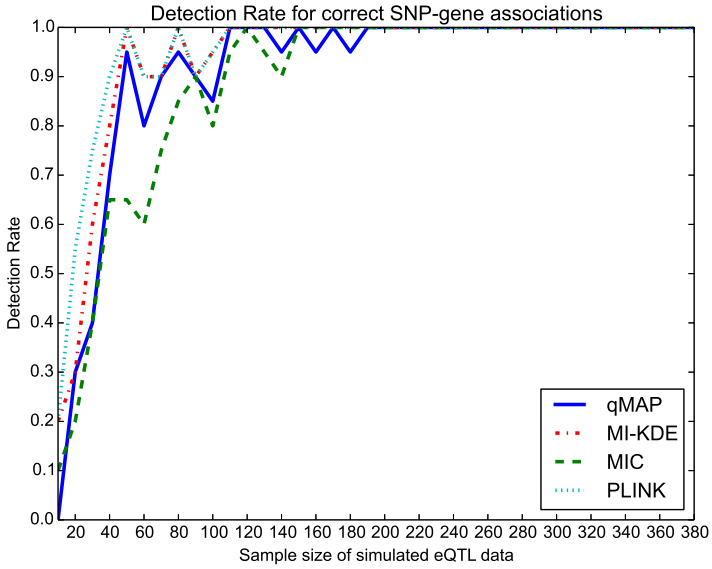


Figure A.11: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 11.

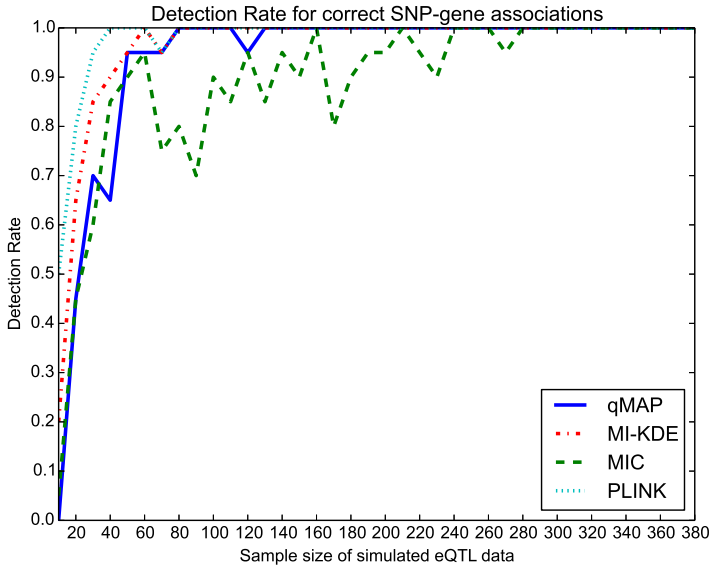


Figure A.12: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 12.

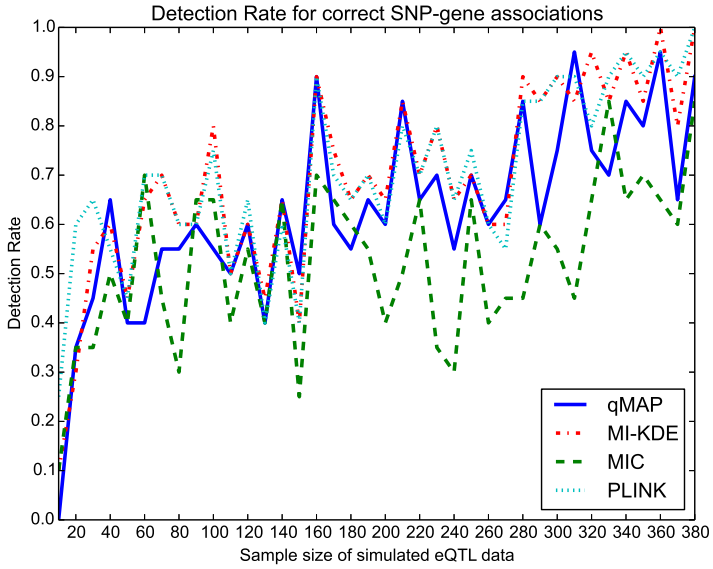


Figure A.13: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 13.

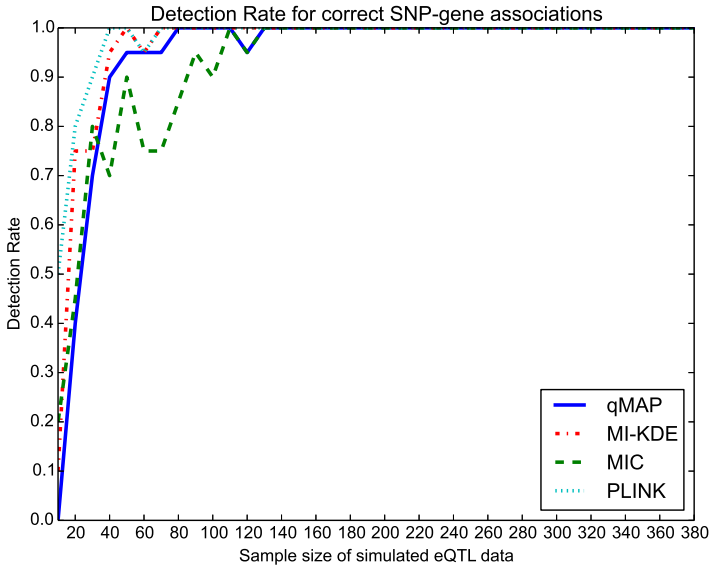


Figure A.14: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 14.

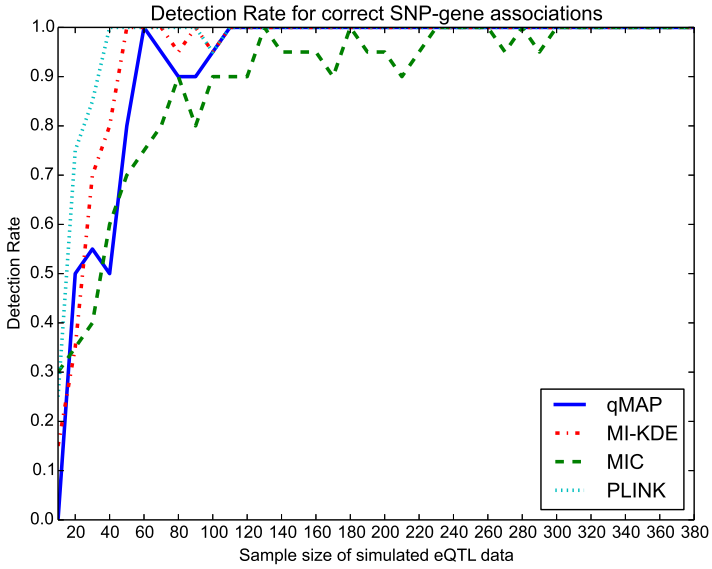


Figure A.15: Each algorithm's detection rate depending on the sample size for correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes. Displayed gene: Gene 15.

Sample Size	qMAP	MI-KDE	MIC	PLINK
10	0.0	5.7	8.3	15.3
20	18.3	22.7	17.3	33.7
30	27.3	33.0	27.7	37.0
40	31.7	37.0	30.7	39.3
50	38.3	40.3	33.3	41.0
60	40.0	43.0	34.7	42.7
70	43.7	42.7	34.0	43.3
80	44.0	44.0	33.7	43.3
90	44.3	43.3	37.7	43.3
100	45.0	46.0	37.0	45.3
110	44.3	45.3	37.7	44.3
120	48.3	47.7	40.7	46.0
130	45.3	46.3	39.0	44.3
140	51.0	50.7	42.3	47.0
150	47.3	49.0	43.0	45.7
160	54.0	54.0	41.3	48.3
170	54.0	52.3	42.7	49.3
180	51.3	55.0	44.7	49.0
190	55.0	57.0	43.7	49.3
200	57.3	56.7	41.0	51.0
210	57.7	62.0	45.0	52.3
220	58.7	60.7	47.0	50.3
230	58.3	60.3	42.0	52.3
240	59.3	59.3	45.0	52.0
250	59.7	60.7	45.0	51.0
260	64.0	60.7	44.0	51.0
270	64.3	63.0	45.7	51.7
280	66.7	69.3	45.0	53.7
290	64.7	70.0	47.0	54.7
300	70.7	74.3	46.3	54.0
310	73.7	72.3	46.3	55.3
320	72.7	73.0	49.0	54.0
330	73.3	73.3	50.7	56.3
340	76.0	74.3	49.0	57.0
350	76.7	74.0	50.3	56.3
360	77.3	77.7	49.7	57.0
370	77.3	74.3	48.3	56.3
380	78.0	78.0	52.0	57.3

Table A.1: Detection rates of correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes depending on the sample size. (Displayed values are in percent [%]).

Sample Size	qMAP	MI-KDE	MIC	PLINK
10	0.0	0.0	0.6	0.0
20	1.9	0.0	1.3	0.0
30	1.9	1.3	3.8	0.0
40	0.6	1.3	0.6	0.6
50	1.9	1.9	1.3	0.6
60	0.6	1.3	0.6	0.6
70	5.6	1.3	1.9	2.5
80	6.9	1.9	2.5	1.3
90	8.8	3.1	5.0	3.1
100	8.1	3.8	1.3	3.1
110	5.6	5.6	3.8	3.1
120	10.6	8.1	5.6	3.8
130	6.3	7.5	5.0	5.6
140	14.4	12.5	6.3	6.9
150	11.3	13.1	11.9	6.3
160	16.3	15.6	4.4	5.0
170	19.4	15.0	9.4	10.0
180	15.6	20.0	11.3	8.8
190	21.9	23.1	8.8	8.8
200	25.0	23.1	8.1	13.1
210	22.5	30.6	11.3	13.1
220	26.9	30.0	15.0	10.6
230	25.6	28.1	10.0	13.1
240	29.4	28.1	15.0	14.4
250	28.1	30.0	10.6	11.3
260	37.5	31.3	13.8	13.1
270	37.5	35.6	13.8	15.0
280	39.4	43.8	15.0	15.0
290	38.8	45.6	15.0	16.9
300	48.1	53.1	16.9	15.0
310	51.9	50.6	17.5	17.5
320	51.9	50.0	20.6	16.3
330	53.8	51.9	20.0	19.4
340	56.9	52.5	20.0	20.0
350	58.7	53.1	21.3	19.4
360	58.1	58.1	22.5	20.0
370	61.9	54.4	20.0	19.4
380	60.0	58.8	24.4	20.0

Table A.2: Detection rates for weak associations (disease associated genes) in the interaction network obtained from the detection rate of correct SNP-gene transcript associations in the simulated eQTL dataset depending on the sample size. (Displayed values are in percent [%]).

Sample Size	qMAP	MI-KDE	MIC	PLINK
10	0.0	12.1	17.1	32.9
20	37.1	48.6	35.7	72.1
30	56.4	69.3	55.0	79.3
40	67.1	77.9	65.0	83.6
50	80.0	84.3	70.0	87.1
60	85.0	90.7	73.6	90.7
70	87.1	90.0	70.7	90.0
80	86.4	92.1	69.3	91.4
90	85.0	89.3	75.0	89.3
100	87.1	94.3	77.9	93.6
110	88.6	90.7	76.4	91.4
120	91.4	92.9	80.7	94.3
130	90.0	90.7	77.9	88.6
140	92.9	94.3	83.6	92.9
150	88.6	90.0	78.6	90.7
160	97.1	97.9	83.6	97.9
170	93.6	95.0	80.7	94.3
180	92.1	95.0	82.9	95.0
190	92.9	95.7	83.6	95.7
200	94.3	95.0	78.6	94.3
210	97.9	97.9	83.6	97.1
220	95.0	95.7	83.6	95.7
230	95.7	97.1	78.6	97.1
240	93.6	95.0	79.3	95.0
250	95.7	95.7	84.3	96.4
260	94.3	94.3	78.6	94.3
270	95.0	94.3	82.1	93.6
280	97.9	98.6	79.3	97.9
290	94.3	97.9	83.6	97.9
300	96.4	98.6	80.0	98.6
310	98.6	97.1	79.3	98.6
320	96.4	99.3	81.4	97.1
330	95.7	97.9	85.7	98.6
340	97.9	99.3	82.1	99.3
350	97.1	97.9	83.6	98.6
360	99.3	100.0	80.7	99.3
370	95.0	97.1	80.7	98.6
380	98.6	100.0	83.6	100.0

Table A.3: Detection rates for strong associations (disease unassociated genes) in the interaction network obtained from the detection rate of correct SNP-gene transcript associations in the simulated eQTL dataset depending on the sample size. (Displayed values are in percent [%]).

Sample Size	qMAP	MI-KDE	MIC	PLINK
10	1000000	756	637	918
20	787	766	641	964
30	650	936	496	841
40	527	694	531	843
50	525	626	509	767
60	343	456	588	809
70	376	435	559	769
80	252	279	364	763
90	124	214	376	647
100	163	134	429	768
110	132	145	361	687
120	83	147	325	717
130	54	103	387	539
140	49	52	314	576
150	59	72	280	653
160	42	62	272	568
170	30	55	278	496
180	26	60	295	523
190	14	54	167	548
200	13	53	247	508
210	8	25	245	369
220	12	19	213	372
230	6	14	194	605
240	7	19	211	457
250	6	17	190	404
260	5	8	173	437
270	4	13	190	344
280	5	16	208	457
290	5	17	160	380
300	5	23	174	378
310	5	12	150	399
320	4	7	159	360
330	3	9	186	399
340	3	16	137	342
350	3	5	233	333
360	3	8	245	313
370	3	6	191	259
380	3	4	154	320

Table A.4: Median ranks of correct SNP-gene transcript associations in the simulated eQTL dataset consisting of an interaction network of 15 genes depending on the sample size.

Sample Size	qMAP	MI-KDE	MIC	PLINK
10	1000000	1359	1156	1710
20	1454	1434	1198	1807
30	1217	1754	928	1575
40	988	1301	995	1579
50	983	1173	952	1438
60	642	854	1102	1516
70	704	815	1047	1441
80	472	523	681	1430
90	232	401	703	1212
100	304	249	804	1439
110	247	272	675	1287
120	156	275	608	1344
130	101	191	725	1009
140	91	97	588	1080
150	111	134	524	1224
160	78	115	508	1064
170	54	101	519	928
180	48	111	551	979
190	25	100	313	1027
200	24	98	461	951
210	15	47	458	691
220	21	35	398	697
230	11	25	363	1133
240	11	35	394	855
250	11	31	354	757
260	8	13	323	818
270	7	24	355	644
280	9	29	389	856
290	8	31	298	711
300	8	41	325	709
310	8	22	280	748
320	6	13	297	674
330	5	16	348	747
340	5	29	256	640
350	4	9	435	623
360	4	14	457	585
370	4	10	356	485
380	4	7	286	600

Table A.5: Median ranks of weak associations (disease associated genes) for correct SNP-gene transcript associations in the simulated eQTL dataset depending on the sample size.

Sample Size	qMAP	MI-KDE	MIC	PLINK
10	1000000	66	44	12
20	25	2	6	1
30	2	1	2	1
40	1	1	1	1
50	1	1	1	1
60	1	1	1	1
70	1	1	1	1
80	1	1	2	1
90	1	1	1	1
100	1	1	1	1
110	1	1	2	1
120	1	1	1	1
130	1	1	2	1
140	1	1	1	1
150	1	1	1	1
160	1	1	1	1
170	1	1	2	1
180	1	1	1	1
190	1	1	1	1
200	1	1	2	1
210	1	1	1	1
220	1	1	1	1
230	1	1	2	1
240	1	1	2	1
250	1	1	2	1
260	1	1	2	1
270	1	1	2	1
280	1	1	2	1
290	1	1	2	1
300	1	1	2	1
310	1	1	2	1
320	1	1	1	1
330	1	1	2	1
340	1	1	2	1
350	1	1	2	1
360	1	1	2	1
370	1	1	2	1
380	1	1	2	1

Table A.6: Median ranks of strong associations (disease unassociated genes) for correct SNP-gene transcript associations in the simulated eQTL dataset depending on the sample size.

Bibliography

[1] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[2] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, "The Sequence of the Human Genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 5th ed. Garland Science, November 2007.

[4] T. A. Manolio, "Genomewide Association Studies and Assessment of the Risk of Disease," *New England Journal of Medicine*, vol. 363, no. 2, pp. 166–176, 2010.

[5] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop, "Mapping complex disease traits with global gene expression," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 184–194, 2009.

[6] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, *et al.*, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.

[7] Y. Gilad, S. A. Rifkin, and J. K. Pritchard, "Revealing the architecture of gene regulation: the promise of eQTL studies," *Trends in genetics*, vol. 24, no. 8, pp. 408–415, 2008.

[8] C. W. Bartlett, S. Y. Cheong, L. Hou, J. Paquette, P. Y. Lum, G. Jäger, F. Battke, C. Vehlow, J. Heinrich, K. Nieselt, *et al.*, "An eQTL biological data visualization challenge and approaches from the visualization community," *BMC bioinformatics*, vol. 13, no. 8, p. 1, 2012.

[9] M. F. Moffatt, M. Kabesch, L. Liang, A. L. Dixon, D. Strachan, S. Heath, M. Depner, A. von Berg, A. Bufe, E. Rietschel, *et al.*, "Genetic variants

regulating ORMDL3 expression contribute to the risk of childhood asthma," *Nature*, vol. 448, no. 7152, pp. 470–473, 2007.

[10] H. H. Göring, J. E. Curran, M. P. Johnson, T. D. Dyer, J. Charlesworth, S. A. Cole, J. B. Jowett, L. J. Abraham, D. L. Rainwater, A. G. Comuzzie, *et al.*, "Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes," *Nature genetics*, vol. 39, no. 10, pp. 1208–1216, 2007.

[11] P.-R. Loh, M. Baym, and B. Berger, "Compressive genomics," *Nature biotechnology*, vol. 30, no. 7, pp. 627–630, 2012.

[12] K. Davies, *The $1,000 genome: the revolution in DNA sequencing and the new era of personalized medicine*. Simon and Schuster, 2010.

[13] S. D. Kahn *et al.*, "On the future of genomic data," *Science*, vol. 331, no. 6018, pp. 728–729, 2011.

[14] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure, "Exome sequencing as a tool for Mendelian disease gene discovery," *Nature Reviews Genetics*, vol. 12, no. 11, pp. 745–755, 2011.

[15] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, *et al.*, "Exome sequencing identifies the cause of a mendelian disorder," *Nature genetics*, vol. 42, no. 1, pp. 30–35, 2010.

[16] G. J. Annas and S. Elias, "23andMe and the FDA," *New England Journal of Medicine*, vol. 370, no. 11, pp. 985–988, 2014.

[17] "Human Genome Center Supercomputer," [Online: https://supcom.hgc.jp], Accessed: 2015-02-07.

[18] J. Hagenauer, Z. Dawy, B. Goebel, P. Hanus, and J. Mueller, "Genomic analysis using methods from information theory," in *Information Theory Workshop, 2004. IEEE*. IEEE, 2004, pp. 55–59.

[19] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J. C. Mueller, "Gene mapping and marker clustering using Shannon's mutual information," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 3, no. 1, pp. 47–56, 2006.

[20] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic acids research*, vol. 29, no. 1, pp. 308–311, 2001.

[21] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, *et al.*, "The international HapMap project," *Nature*, vol. 426, no. 6968, pp. 789–796, 2003.

[22] J. Craig, "Complex diseases: Research and applications," *Nature Education*, vol. 1, no. 1, p. 184, 2008.

[23] J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, R. J. Loos, *et al.*, "Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits," *Nature genetics*, vol. 44, no. 4, pp. 369–375, 2012.

[24] J. Hardy and A. Singleton, "Genomewide Association Studies and Human Disease," *New England Journal of Medicine*, vol. 360, no. 17, pp. 1759–1768, 2009.

[25] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9362–9367, 2009.

[26] T. A. Manolio, L. D. Brooks, and F. S. Collins, "A hapmap harvest of insights into the genetics of common disease," *The Journal of clinical investigation*, vol. 118, no. 5, pp. 1590–1605, 2008.

[27] H. Kitano, "Systems Biology: A Brief Overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[28] J.-B. Veyrieras, S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, and J. K. Pritchard, "High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation," *PLoS genetics*, vol. 4, no. 10, p. e1000214, 2008.

[29] V. Emilsson, G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. B. Walters, S. Gunnarsdottir, *et al.*, "Genetics of gene expression and its effect on disease," *Nature*, vol. 452, no. 7186, pp. 423–428, 2008.

[30] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature methods*, vol. 5, no. 7, pp. 621–628, 2008.

[31] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature methods*, vol. 8, no. 6, pp. 469–477, 2011.

[32] K. Yoshida, M. Sanada, Y. Shiraishi, D. Nowak, Y. Nagata, R. Yamamoto, Y. Sato, A. Sato-Otsubo, A. Kon, M. Nagasaki, *et al.*, "Frequent pathway mutations of splicing machinery in myelodysplasia," *Nature*, vol. 478, no. 7367, pp. 64–69, 2011.

[33] A. J. Myers, J. R. Gibbs, J. A. Webster, K. Rohrer, A. Zhao, L. Marlowe, M. Kaleem, D. Leung, L. Bryden, P. Nath, *et al.*, "A survey of genetic human cortical gene expression," *Nature genetics*, vol. 39, no. 12, pp. 1494–1499, 2007.

[34] J. C. Barrett, S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada, *et al.*, "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease," *Nature genetics*, vol. 40, no. 8, pp. 955–962, 2008.

[35] C. Libioulle, E. Louis, S. Hansoul, C. Sandor, F. Farnir, D. Franchimont, S. Vermeire, O. Dewit, M. De Vos, A. Dixon, *et al.*, "Novel Crohn Disease Locus Identified by Genome-Wide Association Maps to a Gene Desert on 5p13.1 and Modulates Expression of PTGER4," *PLoS Genet*, vol. 3, no. 4, p. e58, 2007.

[36] J. Mueller, E. Bresch, Z. Dawy, T. Bettecken, T. Meitinger, and J. Hagenauer, "Shannon's Mutual Information Applied to Population-Based Gene Mapping," in *American Journal of Human Genetics*, vol. 73, no. 5, 2003, pp. 610–610.

[37] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, October 1948.

[38] J. F. Crow, "Shannon's Brief Foray into Genetics," *Genetics*, vol. 159, no. 3, pp. 915–917, 2001.

[39] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[40] D. J. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge university press, 2003.

[41] "Information Theory, Inference, and Learning Algorithms," [Online: http://www.inference.phy.cam.ac.uk/mackay/itila/], Accessed: 2015-02-08.

[42] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pac Symp Biocomput*, vol. 5. Citeseer, 2000, pp. 418–429.

[43] B. Goebel, Z. Dawy, J. Hagenauer, and J. C. Mueller, "An Approximation to the Distribution of Finite Sample Size Mutual Information Estimates," in *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1102–1106.

[44] H. Ueda, J. M. Howson, L. Esposito, J. Heward, G. Chamberlain, D. B. Rainbow, K. M. Hunter, A. N. Smith, G. Di Genova, M. H. Herr, *et al.*, "Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease," *Nature*, vol. 423, no. 6939, pp. 506–511, 2003.

[45] P. Hanus, B. Goebel, J. Dingel, J. Weindl, J. Zech, Z. Dawy, J. Hagenauer, and J. C. Mueller, "Information and communication theory in molecular biology," *Electrical Engineering*, vol. 90, no. 2, pp. 161–173, 2007.

[46] M. Sarkis, B. Goebel, Z. Dawy, J. Hagenauer, P. Hanus, and J. C. Mueller, "Gene mapping of complex diseases - A comparison of methods from statistics informnation theory, and signal processing," *Signal processing magazine, IEEE*, vol. 24, no. 1, pp. 83–90, 2007.

[47] J. Kasturi, R. Acharya, and M. Ramanathan, "An information theoretic approach for analyzing temporal patterns of gene expression," *Bioinformatics*, vol. 19, no. 4, pp. 449–458, 2003.

[48] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context," *BMC bioinformatics*, vol. 7, no. Suppl 1, p. S7, 2006.

[49] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting Novel Associations in Large Data Sets," *science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[50] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[51] J. Rissanen and G. G. Langdon Jr, "Universal modeling and coding," *Information Theory, IEEE Transactions on*, vol. 27, no. 1, pp. 12–23, 1981.

[52] J. Rissanen, "Universal coding, information, prediction, and estimation," *Information Theory, IEEE Transactions on*, vol. 30, no. 4, pp. 629–636, 1984.

[53] ——, "Stochastic Complexity and Modeling," *The annals of statistics*, pp. 1080–1100, 1986.

[54] ——, "Stochastic Complexity," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 223–239, 1987.

[55] ——, "Hypothesis Selection and Testing by the MDL Principle," *The Computer Journal*, vol. 42, no. 4, pp. 260–269, 1999.

[56] ——, "Strong optimality of the normalized ML models as universal codes and information in data," *Information Theory, IEEE Transactions on*, vol. 47, no. 5, pp. 1712–1717, 2001.

[57] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2743–2760, 1998.

[58] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *Information Theory, IEEE Transactions on*, vol. 37, no. 4, pp. 1034–1054, 1991.

[59] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

[60] P. D. Grünwald, *The minimum description length principle*. MIT press, 2007.

[61] P. D. Grünwald, I. J. Myung, and M. A. Pitt, *Advances in minimum description length: Theory and applications*. MIT press, 2005.

[62] I. Tabus and J. Astola, "On the use of MDL principle in gene expression prediction," *EURASIP Journal on Applied Signal Processing*, vol. 2001, no. 1, pp. 297–303, 2001.

[63] J. Dougherty, I. Tabus, and J. Astola, "Inference of gene regulatory networks based on a universal minimum description length," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2008, p. 5, 2008.

[64] W. Zhao, E. Serpedin, and E. R. Dougherty, "Inferring gene regulatory networks from time series data using the minimum description length principle," *Bioinformatics*, vol. 22, no. 17, pp. 2129–2135, 2006.

[65] R. Jörnsten and B. Yu, "Simultaneous gene clustering and subset selection for sample classification via MDL," *Bioinformatics*, vol. 19, no. 9, pp. 1100–1109, 2003.

[66] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, *et al.*, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[67] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problemy Peredachi Informatsii*, vol. 23, no. 3, pp. 3–17, 1987.

[68] P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters*, vol. 103, no. 6, pp. 227–233, 2007.

[69] P. Kontkanen, W. Buntine, P. Myllymäki, J. Rissanen, and H. Tirri, "Efficient computation of stochastic complexity," in *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*, 2003, pp. 233–238.

[70] P. Kontkanen and P. Myllymäki, "MDL histogram density estimation," in *International Conference on Artificial Intelligence and Statistics*, 2007, pp. 219–226.

[71] Y. Kameya, "Time series discretization via mdl-based histogram density estimation," in *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*. IEEE, 2011, pp. 732–739.

[72] "PLINK," [Online: http://pngu.mgh.harvard.edu/~purcell/plink/], Accessed: 2015-02-07.

[73] J. J. Corneveaux, A. J. Myers, A. N. Allen, J. J. Pruzin, M. Ramirez, A. Engel, M. A. Nalls, K. Chen, W. Lee, K. Chewning, *et al.*, "Association of CR1, CLU and PICALM with Alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals," *Human molecular genetics*, p. ddq221, 2010.

[74] C. Liu, L. Cheng, J. A. Badner, D. Zhang, D. W. Craig, M. Redman, and E. S. Gershon, "Whole genome association mapping of gene expression in the human prefrontal cortex," *Molecular psychiatry*, vol. 15, no. 8, p. 779, 2010.

[75] . G. P. Consortium *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[76] ——, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[77] S. Wright, "An Analysis of Variability in Number of Digits in an Inbred Strain of Guinea Pigs," *Genetics*, vol. 19, no. 6, p. 506, 1934.

[78] "MIC," [Online: http://www.exploredata.net/], Accessed: 2015-02-07.

[79] E. Jones, T. Oliphant, P. Peterson, *et al.*, "SciPy: Open source scientific tools for Python," [Online: http://www.scipy.org/], 2001–, Accessed: 2015-02-07.

[80] T. E. Oliphant, "Python for Scientific Computing," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 10–20, 2007.

[81] K. J. Millman and M. Aivazis, "Python for Scientists and Engineers," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 9–12, 2011.

[82] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 2015.

[83] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. CRC press, 1986, vol. 26.

[84] D. M. Bashtannyk and R. J. Hyndman, "Bandwidth selection for kernel conditional density estimation," *Computational Statistics & Data Analysis*, vol. 36, no. 3, pp. 279–298, 2001.

[85] K. Xia, A. A. Shabalin, S. Huang, V. Madar, Y.-H. Zhou, W. Wang, F. Zou, W. Sun, P. F. Sullivan, and F. A. Wright, "seeQTL: a searchable database for human eQTLs," *Bioinformatics*, vol. 28, no. 3, pp. 451–452, 2012.

[86] A. J. Myers, A. M. Pittman, A. S. Zhao, K. Rohrer, M. Kaleem, L. Marlowe, A. Lees, D. Leung, I. G. McKeith, R. H. Perry, *et al.*, "The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts," *Neurobiology of disease*, vol. 25, no. 3, pp. 561–570, 2007.