FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Bayesian Variable Selection in Genome-wide Association Studies

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Medizinische Informatik

eingereicht von

Stephan Weinwurm BSc.
Matrikelnummer 0625729

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Mag.rer.nat. Dr.techn. Rudolf Freund
Mitwirkung: Univ.Prof. Dipl.-Ing. Dr.rer.nat. Johann Sölkner
Dr. Patrik Waldmann

Wien, 29. Januar 2013

(Unterschrift Stephan Weinwurm BSc.)

(Unterschrift Betreuung)

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Bayesian Variable Selection in Genome-wide Association Studies

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Medical Informatics

by

Stephan Weinwurm BSc.
Registration Number 0625729

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Dipl.-Ing. Mag.rer.nat. Dr.techn. Rudolf Freund
Assistance: Univ.Prof. Dipl.-Ing. Dr.rer.nat. Johann Sölkner
Dr. Patrik Waldmann

Vienna, 29th January 2013 _____ _____
(Signature of Author) (Signature of Advisor)

Erklärung zur Verfassung der Arbeit

Stephan Weinwurm BSc.
Reichergasse 172, 3400 Klosterneuburg

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum)

(Unterschrift Stephan Weinwurm BSc.)

Acknowledgements

I would like to thank everybody who has helped me through the ups and downs during this thesis and contributed in whatever way.
Especially I would like to thank my girlfriend, Kori, who was always there for me, motivated me in hard times and did a great job on correcting and editing - thank you for being mine. I also want to thank my parents and my grandfather for supporting me in so many ways and always having an open ear.
Furthermore, I would like to thank Patrik Waldmann and Johann Sölkner, from the BOKU Wien, for giving me the opportunity to contribute to their project and for answering all my technical questions. Finally, I want to thank Rudolf Freund for supervising this work and leading this thesis to a successful end. Without them the realization of the project would not have been possible.

Abstract

The work confronts a common challenge arising from genome-wide association studies(GWAS). The ultimate goal of GWAS is to identify the true subset of single-nucleotide polymorphisms(SNPs), specific locations within an organism's genome, strongly influencing a certain characteristic, such as a trait or disease. This problem has often been tackled by using methods such as hybrid correlation-based search(hCBS), a modification of a method called stochastic search variable selection, as well as penalized regression methods namely lasso and ridge regression. Due to their generality, these methods are not limited to genome analysis; in fact, they are applicable to a variety of large scale regression problems.

Typical state of the art genome-wide association studies comprise hundreds of thousands or even millions of SNPs in contrast with a much lower number of genomes. The above mentioned approaches are capable of dealing with situations where the number of variables (SNPs) exceeds the number of observations (phenotypes); also known as $p \gg n$ problems. The work at hand discusses modifications of the methods mentioned above to improve performance in terms of variable selection and prediction. Furthermore, all methods, as well as their modifications, are evaluated and compared in settings of highly correlated datasets, as is common in genome-wide association studies.

Kurzfassung

Die vorliegende Arbeit beschäftigt sich mit einer häufigen Problemstellung in genomweiten Assoziationsstudien (GWAS). Das Ziel dieser Studien ist es sogenannte Single-Nucleotid Polymorphismen (SNP), Stellen im Genomen von Organismen die sich zwischen Individuen unterscheiden, zu entdecken, welche ein bestimmtes Merkmal bzw. Charakteristik beeinflussen und prägen. Diese Merkmale werden auch Phänotyp genannt. Die untersuchten Merkmale variieren je nach Interesse und Forschungsfeld und reichen von gewissen Charakterzügen über das Auftreten bestimmter Krankheiten bis hin zu evolutionären Aspekten.
Für diese Aufgabenstellung werden oftmals Methoden wie Hybrid-Correlation-based Search(hCBS), Stochastic Search Variable Selection oder Penalized-Regression Methoden wie Lasso oder Ridge Regression verwendet. Diese Methoden können aufgrund ihrer Generalität nicht nur für Genomanalysen verwendet werden, sondern auch für viele andere Large-Scale Regressionsprobleme.
Heutige genomweite Assoziationsstudien beinhalten hunderttausend bis hin zu Millionen von Single-Nucleotide Polymorphismen im Gegensatz zu einer wesentlich geringeren Anzahl an sequenzierten Genomen. Die erwähnten Methoden sind in der Lage mit dieser Bedingungen umzugehen, wobei die Anzahl an Variablen (SNPs) die Anzahl der Beobachtungen (Phenotypen) bei weitem übersteigen, auch bekannt als $p \gg n$ Probleme. Die Arbeit behandelt Verbesserungen und Modifikationen der oben erwähnten Methoden um die Variablenselektion sowie die Vorhersage ungesehener Phänotypen zu verbessern. Des weiteren werden die Methoden, sowie die vorgeschlagenen Verbesserungen, anhand von hoch korrelierten Datensätzen, wie sie oft in genomweiten Assoziationsstudien auftreten, verglichen und evaluiert.

Contents

CHAPTER 1

Introduction

THE genome contains the blueprint of life for every organism. Understanding the genome is currently among the prevailing challenges in twenty-first century science. To understand the functionality and the mechanisms a great number of genome-wide association studies (GWAS) are conducted; whereby, the goal [17] is to detect variations in the genome and their effects on certain phenotypes. Phenotypes under consideration range from traits such as differences in the appearance of individuals to certain characteristics and complex diseases. Genome-wide association studies [48] identify these associations by comparing individuals with different manifestations of the phenotype. For example a group of individuals affected by a certain disease in contrast to a group not affected or observed differences in their phenotype such as height, eye color or blood groups just to name a few.

To this end [17], the genomes of the individuals are sequenced, using next-generation sequencing [46], and are studied with respect to their ability to explain the phenotype; hence, to find patterns of association between the genetic variations sequenced and the phenotype. The most common genetic variations [11] are variations of single positions in the genome. These single point variations, also called mutations [34], are changes of a single base pairs and are known as single-nucleotide polymorphisms (SNPs). SNPs vary between individuals and are mostly responsible for the variations in characteristics and appearance among individual. For example, it is estimated [2, 11], that the human genome contains approximately 10 million SNPs.

In most genome-wide association studies [23] many hundred thousands, up to one million, single-nucleotide polymorphisms are considered. Variations [17] of an SNP that appear statistically more frequently in a group of individuals with a certain phenotype are considered to influence this phenotype, hence these variations are reported to be associated with that phenotype.

This thesis aims at modifying and improving common methods in the field of genome-wide association studies. The methods, as well as their modifications, are compared and evaluated to alleviate the decision of which method to choose in future studies.

The purpose of this chapter is to introduce the intention and the idea of the thesis and to give an overview of the scope and the structure. In Section 1.1 the purpose of the work is outlined and gives a superficial introduction to the field of research. Section 1.2 addresses the main problems and challenges arising in genome-wide association studies; Section 1.3 defines the goals and purposes of the present work and is followed by a definition of the contribution of the thesis to the research area in Section 1.4. Finally, Section 1.5 outlines the structure of the remaining chapters.

1.1 Motivation

Within the last few decades, genetics has made great progress, due to advances in methodology and technology. Recent results and successes [35] in genetics by genome-wide association studies are a result of a germination period after the first proposals of genome-wide approaches in the nineteen eighties. Moreover, a great upsurge [46] is based on the development of high-throughput sequencing technologies, called next-generation sequencing(NGS), which allow a great number of individuals to be sequenced cost-effectively. As a consequence [48], enormous amounts of data are produced and need to be analyzed, necessitating the development [48] of new biostatistical methods.
It is considered [23, 35] that not only a single or a few single-nucleotide polymorphisms have a large influence on a certain phenotype, but many SNPs with small effects have a large influence together. Nevertheless, most of the GWAS [23] are carried out as single-SNP analysis, testing each SNP one at a time for association to the phenotype under consideration, due to the computational demands arising from more sophisticated approaches. However [39], a simultaneous analysis of multiple SNPs is crucial for the identification of sophisticated and complex associations between genetic and phenotypic variations.
Additionally, due to a phenomenon called linkage disequilibrium [34], single-nucleotide polymorphisms can be partially highly correlated, which impedes the identification of the phenotype-associated positions in the genome.
GWAS have contributed to the understandings in genetics. For example [17], *in human genetics there are already more than 30 SNPs known to be associated with the onset of the autoimmune disease Crohn's disease, around 20 SNPs associated with type 2 diabetes and more than 40 SNPs associated with the height of individuals*. Many more are to be discovered.

1.2 Problem description

The goal of genome-wide association studies [2] is to identify groups of SNPs that vary systematically between individuals with certain phenotypes. State of the art genome-wide association studies [23] comprise hundreds of thousands, sometimes millions of SNPs, in order to identify regions containing SNPs that affect the phenotype of interest. Due to the large number of SNPs used for state of the art studies, a fundamental problem [2] is that patterns can simply arise by chance. Therefore, many genotypes [35] have to be included into the study, to facilitate the identification. With increases in dataset size, computational demands rise significantly. As a consequence, many GWAS are carried out as single-SNP analysis [23], where each SNP is tested separately for its association to the phenotype, to reduce the computational burden. Finally, strong associations are interpreted as the SNP having an influence on the phenotype. This approach is considered [11] as being too simple to elucidate the complex architecture of the genome. Li *et.al.* [39] note that single-SNP analysis has major drawbacks in identifying all causal SNPs:

- Most phenotypes are believed to be polygenetic; that is, multiple genes influence a phenotype. Consequently single-SNP analyses only detect a small proportion of the causal SNPs
- Genes may interact to produce a phenotype, known as epistatic effects [11], which can not be detected by single-SNP analyses.

However, single-SNP analyses are faster [11] than more sophisticated approaches and therefore often used for analysis in genome-wide association studies.
As mentioned before, the human genome contains approximately 10 million SNPs. Depending on the study only a fraction [17] are included in the analysis with the reason that many SNPs are highly correlated and it is therefore not necessary to use all single-nucleotide polymorphisms in one study. This phenomenon is also known as the linkage disequilibrium (LD) [34], where a combination of SNPs is observed either more or less frequently than expected from their random formation.
This leads to the fact [17] that many single-nucleotide polymorphisms found to be associated with a phenotype are unlikely to be the real causal variants in the genome affecting the phenotype. Instead so called proxies or sentinels [17] are unveiled. According to Donnelly [17] a natural follow-up strategy is to include many more of the correlated SNPs from the associated genomic region into fine-mapping studies, where the causal SNPs ideally show a larger association than correlated ones.
As a consequence, more sophisticated approaches [11] are needed to simultaneously analyze large numbers of SNPs, especially in the presence of high correlation, where the number of SNPs usually far exceeds the number of phenotypes,

also known as p≫n problems.
Hybrid correlation based search is designed for the application to highly correlated datasets especially in the case of p≫n; whereas, Stochastic search variable selection as well as Bayesian lasso and Bayesian ridge regression have already been applied to GWAS datasets. All methods are able to perform multi-SNP analysis.

1.3 Goal

The goal of the present work is to improve and modify the common method hybrid correlation-based search (hCBS), in terms of its ability to perform variable selection and prediction. Furthermore, the second ambition is to compare and evaluate hCBS, along with the modifications, in settings of highly correlated datasets with Penalized regression methods such as Bayesian lasso and Bayesian ridge regression.
Conclusions drawn shall alleviate the decision of which method to employ in future studies.

1.4 Contribution

Genome-wide association studies [17,35] have contributed greatly to present knowledge in genetics. Nevertheless, there are still many unsolved questions and much knowledge remains to be unveiled. Therefore, even larger amounts of data need to be analyzed and the methods applied need to be able to identify weak patterns between genetic variations and the phenotypes. The work at hand addresses these challenges; thus, the scientific contribution of this thesis is twofold:

- To improve hybrid correlation-based search in terms of:
  - its ability to detect true positive variables; hence, identifying variables (SNPs) influencing the outcome (phenotype)
  - its ability to predict the outcome (phenotype), based on an unseen dataset; and,
- to conduct a detailed comparison between the methods, with respect to variable selection, prediction, as well as their computational load.

This thesis considers quantitative trait loci (QTL) were two or more positions in the genome influence a continuous phenotypes. Other phenotypes such as binary or ordinal traits are beyond the scope of this thesis.

1.5 Outline

The thesis is structured into several main chapters; the following Chapter 2 handling related topics and papers for further reading, similar methods and approaches, as well as comparative works. Subsequently, Chapter 3 explains the biological and genetic background relevant for understanding the purpose of genome-wide association studies. In Chapter 4, common statistical frameworks and methodologies such as hierarchical models and Markov chain Monte Carlo methods are discussed relevant to the methods and the modifications in Chapter 5. The latter includes the hybrid correlation-based search and its modifications as well as Bayesian lasso and Bayesian ridge regression. Chapter 6 compares the results obtained by applying the methods outlined in Chapter 5 to simulated and real datasets. Finally, Chapter 7 presents a discussion of the methods with respect to their application in genome-wide association studies.

CHAPTER 2

Related Work

THE chapter contains references for further reading and similar topics addressed by other works. Stochastic Search Variable Selection (SSVS) is introduced by George and McCulloch [21] to facilitate the identification of a subset of variables in a multiple regression problem. The method was encouraged by the fact, that model comparisons using Akaike or Bayesian Information Criterion can be prohibitive when p is large. To explore the most probable combinations of the subset SSVS uses a Gibbs sampler, as explained in Section 4.2.2. Therefore, a Bayes hierarchical setup is used to model the regression coefficients β_i as having come from a mixture of two Normal distributions. The first Normal distribution is widespread; whereas, the second normal distribution yields a small variance and is clustered around zero. In every iteration of the Gibbs sampler the regression coefficients are assigned as either having come from the widespread normal distribution or else being clustered around zero. The variables assigned to the widespread distribution are considered to be included in the subset of relevant variables X_γ. Therefore, the latent variable γ, indicating whether predictor β_i is included by setting $\gamma_i = 1$ and excluded by $\gamma_i = 0$ respectively, is used.

Chipman [12] incorporates relationships between variables allowing to model interactions, polynomial effects, dummy variables for categorical factors and restrictions to model sizes into the SSVS again making use of a Gibbs sampler.

George and McCulloch [22] extends SSVS by using conjugate priors and setting variables not included in γ exactly to $\gamma_i = 0$, thereby improving computational speed. Moreover, a Metropolis-Hastings algorithm, as explained in Section 4.2.1, is used instead of the Gibbs sampler.

An extension to the multivariate case was proposed by Brown *et.al.* [7, 8] using different priors.

O'Hara *et.al.* [27] give a detailed overview over various approaches of Bayesian variable selection including a discussion of which method to prefer as well as their

implementations. Fridley [20] addresses various approaches for Bayesian model and Bayesian variable selection including SSVS, Bayesian Model averaging and reversible jump MCMC on genomic datasets. A comprehensive introduction to Bayesian variable selection is given by Guan *et.al.* [23] with focus on large-scale regression and its specific application in genome-wide association studies. Moreover, novel priors are introduced on hyperparameters such as the subset size and the variance of included variables. Liang *et.al.* [41] review various methods and approaches for analyzing highly correlated datasets. Baragatti and Pommeret [3] propose a novel method for variable selection in probit regression introducing an improved g-prior for the regression coefficients to overcome limitations in the case of $p \gg n$ and strong multicollinearity making use of a Metropolis-within-Gibbs sampler.

SSVS has been applied in a number of papers to genomic datasets. An excerpt of papers, from the substantial list of works using SSVS, includes Skarman *et.al.* [52] comparing SSVS and a model selection approach using ANOVA and the Akaike information criterion, Chen *et.al.* [11] comparing different methods to incorporate epistatic effects[1], Yang *et.al.* [64] using a two-step approach combining Bayesian probit regression and SSVS, Srivastava *et.al.* [54] applying both the Lasso and SSVS to identify genes influencing rheumatoid arthritis. First approaches using SSVS for genomic datasets are made by Yi [66] and Yi *et.al.* [65] using SSVS for gene mapping problems with quantitative trait loci and Meuwissen *et.al.* [47] making use of linkage disequilibrium in SSVS.

Ridge regression has first been introduced by Hoerl and Kennard [33] to improve prediction in the face of multicollinearity. Subsequently, the least absolute shrinkage and selection operator (Lasso) is introduced by Tibshirani [59] to enable subset selection and actively exclude variables. Various other penalized regression methods exists beside ridge regression and lasso for example group lasso [68], the fused lasso [60] and the elastic net [30].
Park and Casella [49] introduce a Bayesian formulation of the lasso where the lasso estimate is obtained as a posterior mode of the hierarchical model explored by Gibbs sampling. Variable selection is guided by the use of the interval estimates obtained from the posterior distribution. Hans [24] extends the Bayesian lasso by focusing on the prediction and introduces a slightly different model again making use of Gibbs sampling.
The Bayesian lasso is extended to a more general formulation by Kyung *et.al.* [37] to suit other penalized regression methods such as the fused lasso, group lasso and the elastic net and a discussion about problem arising from standard errors in a

[1]Epistasis is the phenomenon were the functionality of a gene and its effects is influenced and regulated by other genes.

non-bayesian formulation is given. A slightly different Bayesian elastic net is proposed by Li and Lin [40] as well as by Hans [26] using different priors.
Hans [25] also proposes a novel of Bayesian lasso able to actively perform variable selection making use of a similar approach as in SSVS.

Bayesian penalized regression is applied to genomic datasets by Yi and Xu [67] along with other Bayesian hierarchical models to identify a subset of quantitative trait loci. Cai *et.al.* [9] proposes a fast empirical Bayesian lasso and applies is to genomic datasets for the identification of multiple quantitative trait loci; whereas, Cleveland *et.al.* [13] compares Bayesian lasso to other methods for prediction of breeding values. A two-step approach is proposed by Li *et.al.* [39] first reducing the number of SNPs and subsequently applying Bayesian lasso to identify associated SNPs. Harris *et.al.* [28] compare the accuracy of predictions made by Bayesian lasso and Bayesian ridge regression using different SNP densities. Finally, Silva *et.al.* [51] discusses the accuracy of Bayesian lasso to predict the breeding value with respect to the choice of the shrinkage parameter λ.

CHAPTER 3

Genetic Background

THE purpose of this chapter is to clarify the main principles of genetics necessary to convey the idea of genome-wide association studies. Therefore, certain fundamentals of genetics and biochemistry are outlined to aid in the understanding of the underlying biology of the analysis of genomic data, obtained by sequencing genomes. The genetic information and its relevance to organisms are discussed as well as the reason for the great interest in understanding the genome in many areas of research. Due to the complexity of the deoxyribonucleic acid (DNA) and the genetic information contained therein, a detailed overview lies beyond the scope of this thesis and can be found in various books, for example Molekulare Genetik [34] by Rolf Knippers, Statistical Methods in Genetic Epidemiology by Duncan c. Thomas [58], especially Chapter 2, as well as Principles of Biochemistry [38] by Nelson and Cox.

The outline of this Chapter is as follows:
Section 3.1 gives a superficial explanation of the connection between the genome and the phenotypes and traits of individuals as well as the effects of changes in the genome. Subsequently, Section 3.2 explains the biological and biochemical background of the genome; whereas, in Section 3.3, together with Subsection 3.3.1 and Subsection 3.3.2, the assembly of proteins from the DNA is described. Section 3.4 addresses the reasons for genetic variability in populations coming from reproduction in Subsection 3.4.1 and mutations in Subsection 3.4.2. The Chapter concludes with the analysis of genomes and the expectations for discoveries as well as a brief introduction of single-SNP analysis in Section 3.6.

3.1 Introduction

Most characteristics of an organism are determined by its genome [34], which basically functions as a construction plan. Certain regions in the DNA, better known as **genes**, are used as the basis for the assembly of molecules called **proteins** which undertake various tasks in the organism and consequently, are responsible for a great number of processes in the organism. A change in the construction plan [34], during cell division reproduction, or another occurrence can lead to a change in the functionality of proteins, which, as a consequence, can lead to an alteration of an organism's traits. Researchers have great interest in determining the influence of genes on an organism's characteristics and in unveiling the impacts of changes in the DNA.
The following Sections gives a more detailed introduction to this process and outlines why the analysis of certain changes in the genome is used in GWAS.

3.2 DNA

The DNA [34] is a long molecule that is the source of genetic information in every living organism[1]. DNA is also referred to as **genome** and is mostly used to describe the entirety of an organism's hereditary information; i.e., the genetic information inherited from its ancestors. As can be seen in Figure 3.1, DNA has the form of a double helix with connections between the two outer boundaries, named **strands**. These two strands of the double helix are called the **backbone** and the connections are termed **base pairs**. Every base pair consists of two connected molecules each attached to one strand termed **nucleotide** [34]. Four different nucleotides exist namely Guanine (G), Cytosine (C), Adenosine (A) and Thymine (T). The only connection possible [38] is between Adenosine and Thymine and Guanine and Cytosine and vice versa; therefore, the sequence of the nucleotides on the two strands is said to be **complementary**. By knowing the sequence of one strand, the sequence of the complementary strand can easily be inferred.
The sequence of the base pairs attached to the backbone along the DNA encodes the genetic information and consequently determines the characteristics and traits of the individual. Accordingly, the genetic sequence in Figure 3.1 is:

TGAGACTCTGAGAC

Thereby giving the complementary strand the following sequence:

ACTCTGAGACTCTG

[1] Except for some special forms of viruses called RNA-Viruses

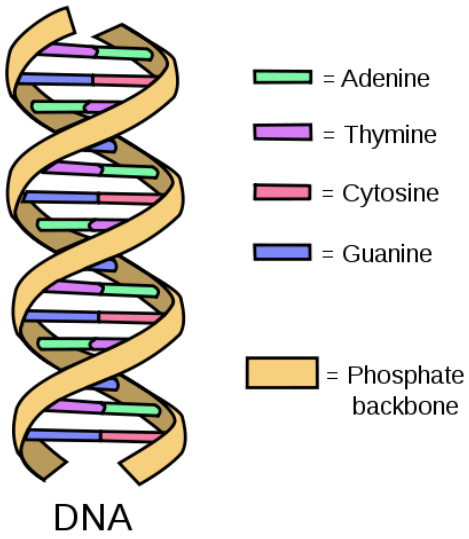

Figure 3.1: A simplified illustration of a short piece of double-stranded DNA - [from commons.wikimedia.org]

This sequence depicts how the sequence of a genome could look. Nevertheless, as seen in Table 3.1, real genomes comprise millions of base pairs, the number depending on the species as well as the affiliation to the group of Eukaryotes[2] or Prokaryotes[3].

Organism	**Size of genome**	**Number of chromosomes**	**Estimated number of genes**
Yeast	12 Millions	16	6 240
Common Fruit Fly	97 Millions	6	18 240
Maize/Corn	2 400 Millions	10	30 - 40 000
Mouse	3 000 Millions	20	25 000
Human	3 000 Millions	23	25 000

Table 3.1: The table shows genome sizes, the number of the haploid chromosomes and the number of estimated genes in various species.

Table 3.1 depicts the haploid genome of the organisms, due to the fact, that different organisms can have a varying number of chromosome sets[4].

[2]Any organisms having a complex cellular structure containing specialized organelles as well as a nucleus. Eukaryotes include all multicellular organisms, such as animals, plants and fungi.

[3]Prokaryotes are organisms, whose cells lack a cell nucleus as well as other organelles. Their DNA is present in the cell without being surrounded by a membrane.

[4]Most eukaryotic cells contain two sets of their genome (two sets of chromosomes), inherited

As stated in Section 3.1, the information in the genome is used to assemble macromolecule called proteins [34], which in turn undertake a great number of important tasks in the organism. Nevertheless, most of the information contained in the DNA is not used for the production of proteins. In fact, as depicted in Figure 3.2, only genes are used to assemble proteins. In many eukaryotic cells [34] only 5 - 10% of the DNA contains regions coding a protein. The remaining DNA contains many repetitive areas with either regulatory effects on the genes or unknown genetic function and is referred to as **noncoding DNA** [34]. New findings [18] indicate that over 80% of the noncoding DNA serves some biochemical purpose.

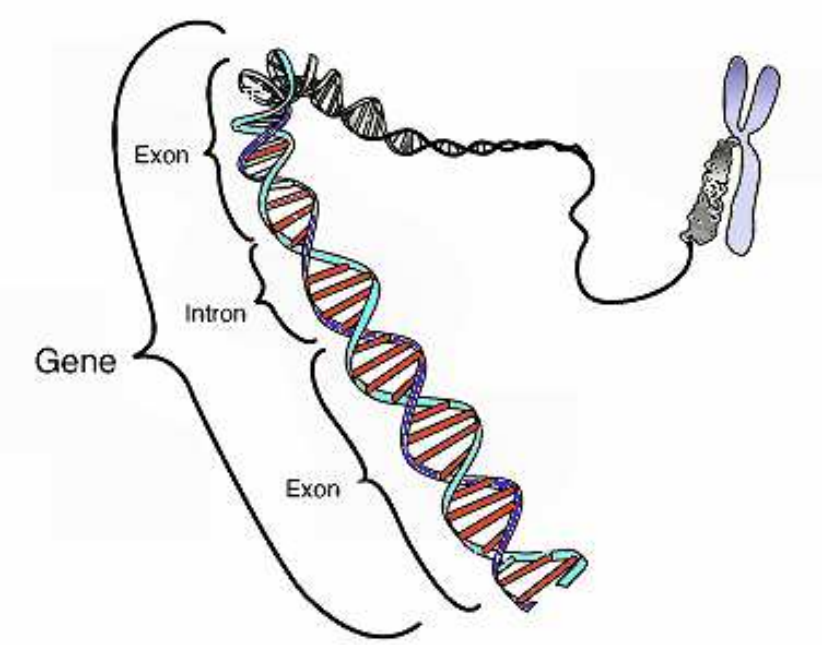

Figure 3.2: A gene, a protein encoding region on a chromosome - [from wikipedia.org]

3.3 Proteins

Proteins [38], large macromolecules present in every cell, are composed of a sequence of many amino acids and are responsible for many tasks, such as structure of the cells, transport of metabolites, which are intermediate substances and prod-

from the parents. A double set of chromosomes is referred to as **diploid** set and a single set is said to be **haploid**, which is depicted in Table 3.1. For example, the human genome consists of 23 chromosomes. This means, the genome is present twice in each cell, one set inherited from the mother and the second one from the father. This results a total of 46 chromosomes.

ucts of metabolism[5], catalysis of chemical reactions and detection of semiochemicals[6]. The function of proteins [38] is determined mainly by their shape, which is in turn determined by the sequence of amino acids of which the proteins are assembled. As discussed in Section 3.3.2 the sequence of amino acids is, with some restrictions, encoded in the sequence of base pairs in the DNA.
The necessary steps from the DNA-sequence to the assembled protein can be seen in Figure 3.3 and are outlined in the following Sections.

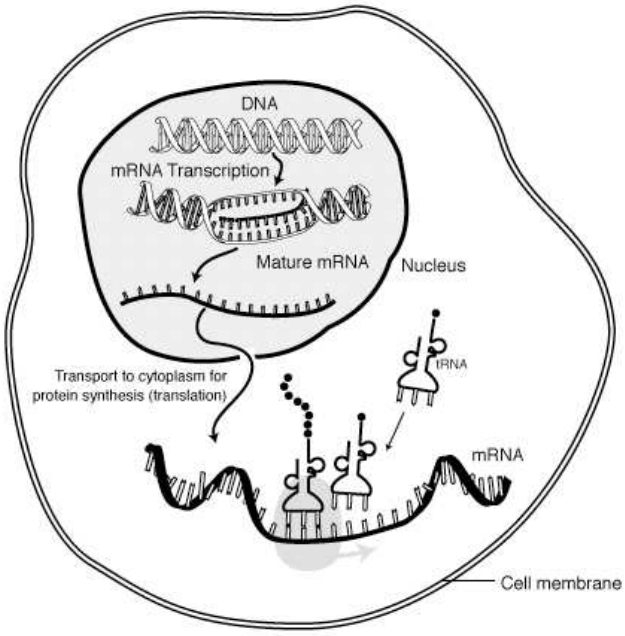

Figure 3.3: Overview of the assembly of proteins - [from wikipedia.org]

3.3.1 Transcription

The first step during the assembly of a protein is called **transcription** [34] and includes duplicating the genetic information from the DNA.
Therefore, the region of the DNA containing the genetic code for the protein to be produced, is unfolded and the two strands are separated, which is depicted in Figure 3.3. The sense strand is then used to copy the information of the gene onto a temporary transport molecule called ribonucleic acid (RNA), for the transport to the place in the cell where the actual proteins are then assembled.
During the process of transcription a molecule called DNA-dependent RNA poly-

[5]Metabolism is referred to as all the necessary chemical reaction to maintain life (growth, cell division, maintain structures and respond to the environment)

[6]Semiochemicals are chemical substances carrying information within or between organisms.

merase, or RNA polymerase for short, binds to the beginning of the gene, by identifying the start position, also known as the **promoter** [34]. Subsequently, the genetic sequence is copied to a new RNA strand by using the same base pairing principles as between the two strands of the DNA. However [34], instead of the nucleotide Thymine (T) the nucleotide Uracil (U) is used. Furthermore, the structure of the RNA differs slightly from a DNA strand. Note that the sequence present on the RNA corresponds to that of the antisense DNA strand, which has not been used to duplicate the genetic information from the DNA. The genetic information is copied to the RNA until a stop sequence, also known as the **terminator** [34], is encountered. The result from the process is called **precursor-messenger RNA (Pre-mRNA)** [34].
The pre-mRNA, as well as the gene itself, contains coding and non-coding regions [34], called **exons** and **introns**. As it can be seen from Figure 3.3.1 the introns, the non-coding regions, are cut out of the immature Pre-mRNA, during a procedure called **splicing**[7] [34]. The result is then called **mature mRNA** or simply **mRNA** and is the basis for the assembly of the final protein. As shown in Table 3.1, not only the size of genomes and the number of genes determine the complexity of an organism. A phenomenon called **alternative splicing** [58] is responsible for one gene encoding three to four different mRNAs and resulting in three or four different proteins. Manny proteins in humans [34] are for example assembled by alternative splicing, whereby, depending on the protein to be produced, different regions of the immature Pre-mRNA are cut out during splicing.

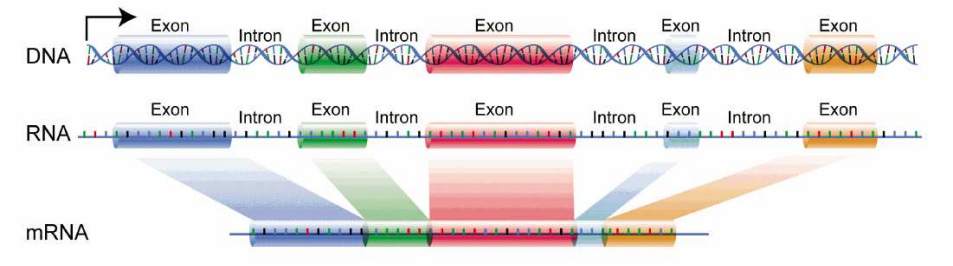

Figure 3.4: Splicing of a immature mRNA molecule by cutting out introns - [from wikipedia.org]

3.3.2 Translation

During the second main step, called **translation** [34], the previously assembled mRNA is translated into the sequence of amino acids, which finally forms the protein.

[7] In Prokaryotes usually no splicing takes place

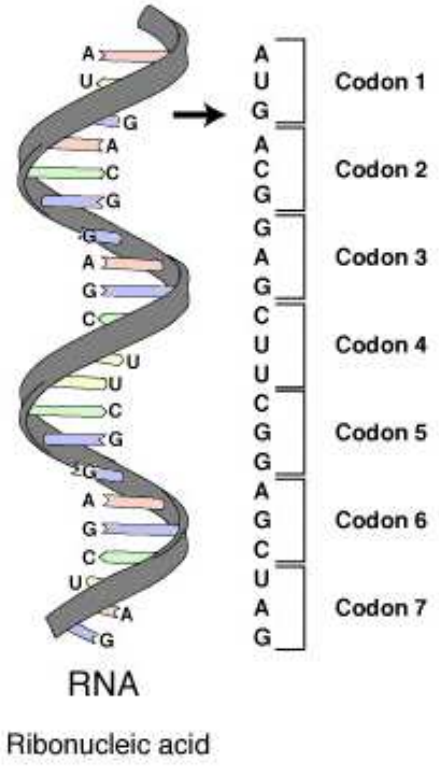

Figure 3.5: mRNA molecule and its translation into tRNA codons - [from wikipedia.org]

The translation is carried out on large molecules called **ribosomes**, which lie outside of the nucleus. Since transcription takes place in the nucleus[8] where the DNA is present, the mRNA first has to be channeled outside the nucleus. The mRNA is then bound to a ribosome and translation begins, as shown in Figure 3.3.
During translation [34], each amino acid is determined by three consecutive nucleotides on the mRNA, called a **codon** or **triplet**, as can be seen in Figure 3.5. Based on the fact that an mRNA transcript contains four different nucleotides (A,U,C and G), $4^3 = 64$ different amino acids can be coded by a nucleotide-triplet. Nevertheless, only 20 different amino acids [34], which are used for the assembly of proteins, exist, which leads to the conclusion, that some amino acids are encoded by more than one triplet of mRNA. Table 3.2 presents the **genetic code** [38], which is the convention on how the triplets are translated into amino acids, valid for every organism known.

[8]Prokaryotes do not have a nucleus and consequently the DNA is present in the cytoplasm, which is the substance inside the cell containing and holding all the cell's internal organelles. The transcription as well as the translation occurs in the cytoplasm.

Note that three codons represent the termination of the assembly of the protein. Whenever they are encountered in the sequence of mRNA, the ribosome stops the composition of the amino acid sequence. Correspondingly, the codon AUG serves as an initiation site. Thus, at the first appearance of AUG the translation of the protein is initiated.

In order to place the correct amino acid onto a corresponding triplet of nucleotides on the mRNA, another molecule, the **transfer-RNA (tRNA)** [38], is needed, as depicted in Figure 3.3. The tRNA is an adapter molecule, consisting of two main features. First, each tRNA molecule contains three nucleotides, which are complementary to the sequence encoded in the RNA. Second, the tRNA binds the corresponding amino acid to the sequence encoded in the tRNA, following the genetic code from Table 3.2. Consequently, a tRNA molecule that binds the amino acid 'glutamic acid' contains either the anticodon CUC or CUU and is able to bind to the mRNA sequence GAG or GAA. For example, as it can be seen in Figure 3.5, this tRNA would bind to the third codon and the ribosome and would add a glutamic acid to the chain of amino acids.

During translation each tRNA molecule binds to the appropriate codon on the mRNA in sequence. The amino acid which is connected to the current tRNA molecule, is then added to the end of the chain of the amino acids already processed. The addition of the amino acid is carried out by ribosomes, as depicted in Figure 3.3

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F Phenylalanine)	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine	UCA		UAA	Stop (Ochre)	UGA	Stop (Opal)	A
	UUG		UCG		UAG	Stop (Amber)	UGG	(Trp/W) Tryptophan	G
C	CUU		CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	AUG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

Table 3.2: Genetic Code

After the chain of amino acids as encoded in the DNA and the mRNA respectively, is completely assembled, the protein folds into a three-dimensional structure [38], which determines its shape and thus also its functionality.
In sum, proteins are responsible for a great number of important tasks in every cell and also in the organism as a whole. The functionality, and accordingly its shape, of a protein is determined by the sequence of amino acids, which is assembled from the sequence of the mRNA, originally from the DNA.

3.4 Genetic variability

As explained in the previous Sections, the genetic information in the DNA determine the phenotypes and traits of each individual through the transcription and translation of the DNA into proteins. The genetic information is therefore responsible for the variety of organisms and variability between individuals of certain species. The reason for the great diversity in organisms is the adaption and the diversification of organisms over time. During reproduction [34], the transfer of the genetic information to offspring, the genome is altered, which is known as **recombination** [34], to ensure adaption and improvement.
The remainder of this Section gives an overview of the processes of cell division and outlines the basic principles and causes of mutations, which are the basis for genetic variation.

3.4.1 Reproduction

Each cell passes on its genetic information to the next generation of cells during cell division[9]. During which the DNA is duplicated to provide the daughter cell with the genome, which in turn is the basis for the new cell. In order to perform cell division, the genetic information, started as DNA, has to be duplicated. In the first step, the DNA is unwound at certain locations known as **origins** [34] in order to enable special molecules known as **DNA polymerase** to copy the sequence of base pairs onto a new DNA strand. As depicted in Figure 3.6 both strands are simultaneously duplicated and each of the two strands serves as the complementary strand for the the second strand of the new double helix. Finally, two identical DNA strands are obtained.
In eukaryotic cells two different types of cell division exist, which are explained briefly:

[9] In Prokaryotes cell division is the formation of daughter cells and a form of reproduction. In Eukaryotes the reason for cell division is twofold. On one side is the proliferation of cells during the embryonic stage and on the other side the replacement of dead cells. The latter is for example the reaction to an injury or an inflammation. Cell division is stopped, when the desired amount of cells is reached.

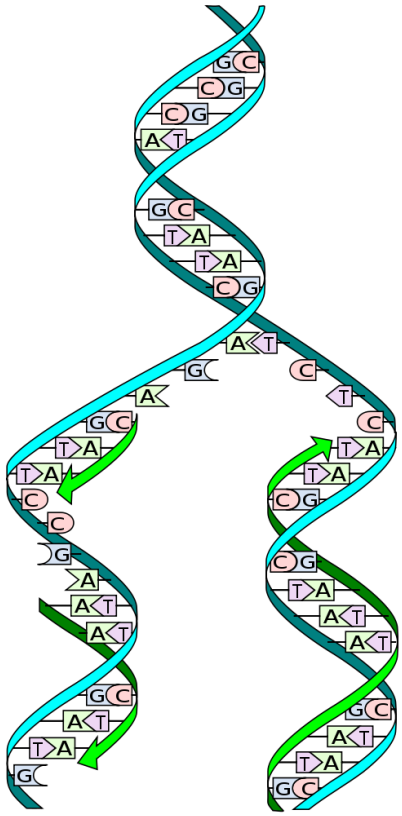

Figure 3.6: Duplication of the DNA - [from wikipedia.org]

- **Mitosis**: Mitosis [34] is the process where a cell first duplicates its genome, followed by the division of the cell, including the nucleus, the organelles and the cell membrane into two cells. After separation, each of the two daughter cells, contains the complete genome. This process occurs during development and growth, where the number of cells in the organism increases, or during the replacements of lost cells[10]. An overview can be seen in Figure 3.7, where the main steps of mitosis are shown.

- **Meiosis**: Meiosis [34] is a specific type of cell division, which is necessary for the sexual reproduction in eukaryotes. Most of the steps involved are similar to mitosis; although, instead of two four daughter cells are produced. These four cells contain only a haploid set of chromosomes, which means that the chromosomes are duplicated, and only one set chromosomes is passed to each daughter cell. Furthermore, an important aspect of meiosis is, that the duplicated chromosomes are recombined. Recombination is the exchange of regions between the two sets of chromosomes in order to gener-

[10]Mitosis also occurs during regeneration (only a few organisms are able to regenerate lost parts and during asexual reproduction (or vegetative reproduction) in plants.

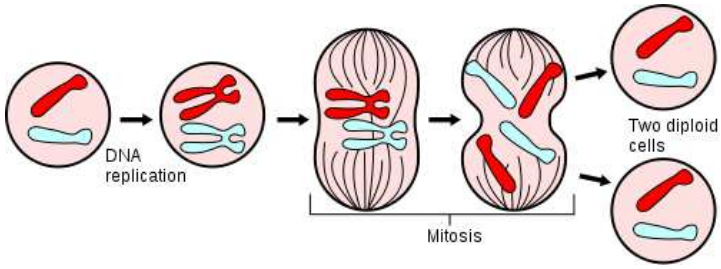

Figure 3.7: Overview of the most important steps during mitosis - [from wikipedia.org]

ate a slightly different genetic combination which can be seen in the second step in Figure 3.8. Moreover, the distribution of the recombined chromosomes are randomly selected[11] to form a haploid set of chromosomes. This is shown in step three and four in Figure 3.8.

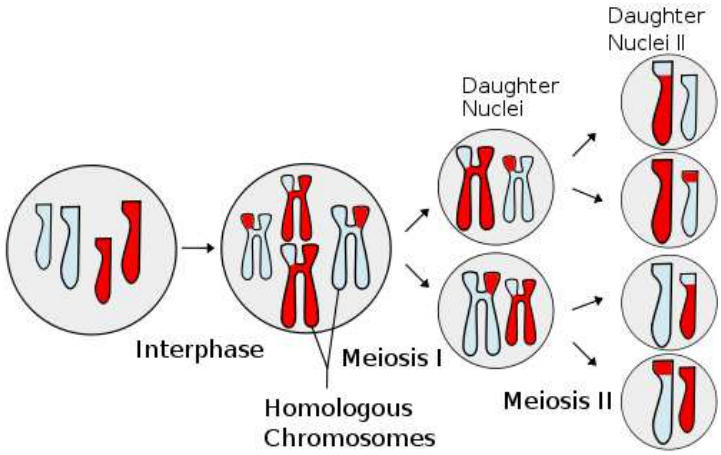

Figure 3.8: Overview of the most important steps during meiosis - [from wikipedia.org]

3.4.2 Mutations

In Section 3.4 the great diversity of existing organisms as it evolved over time is mentioned. This adaption and diversification of organisms is based on random changes in the DNA. These changes are called **mutations** [58] and are inheritable changes in genetic information. Mutations alter the sequence of the DNA, either having no effect, changing the protein encoded by the gene, or preventing the gene completely from functioning.

[11] Assuming a human chromosome set of 23 chromosomes, this leads to $2^2 3 \approx 8.4 * 10^6$ possible combinations.

Mutations are rare occurrences; otherwise, the transfer of the genetic information to the offspring would not result in similar offspring. Nevertheless, mutations are the basis for evolution; natural selection is the survival of organisms, and also their genes, which are thereby better adapted to their environment. However, each cell has complex mechanisms [34] for the repair of a damaged or altered DNA. Different versions of a gene are also referred to as **alleles**[12].
Mutations can be divided into two main groups [34]:

- **Chromosome mutations**: Chromosome mutations are changes in the number, shape or structure of chromosomes.
- **Gene mutations**: Gene mutations are alterations of the sequence of base pairs within a gene or outside the coding regions.

Within the scope of this work, as in GWAS, chromosome mutations do not play a great role; therefore, more attention is given to gene mutations. This type of changes to the DNA can have various causes, which will be discussed in the remainder of this Subsection.
During the replication of the DNA, every ten- to every hundred-thousandth nucleotide [34] placed on the newly assembled strand, is not complementary to the nucleotide on the strand being duplicated. For example, on the left strand in Figure 3.6: if the subsequent nucleotide added to the left strand were not a C, then a mutation would have occurred.
As mentioned above, every cell contains complex mechanisms to detect and repair mutations. In this case a substructure of the molecule duplicating the strand detects and removes the falsely positioned nucleotide so the correct nucleotide can be attached. This repair mechanism is called **mismatch-repair** [38] and it also plays a role in error detection during recombination in meiosis.
Not only errors during the replication and recombination of the DNA can alter the sequence of the DNA [34]. Due to the fact that DNA is a very complex and fragile molecule, metabolic products and external influences such as radiation, chemicals or toxins can interact with the DNA and change its structure or the sequence contained. Such events occur hundreds to thousands of times each day in every cell and without an effective repair mechanism frequent, severe mutations would be the consequence.
Nevertheless, certain mutations on single positions in the DNA occur without being recognized by the repair mechanisms. The result is a **single point mutation** [34], a permanent mutation, which becomes affixed during the next cell division. These mutations only take place in certain sequences and regions of the DNA. Around three-fourths of the gene mutations are exchanges of single nucleotides -

[12] A gene that has the sequence ATCTTA in one population and CTCTTA in another population are called alleles of the gene. Both encode the same protein, but the protein is not identical since the sequence of amino acids has changed.

also known as **single-nucleotide polymorphisms (SNPs)**[13]. Depending on the position of the SNP, the mutation can be:

- **Neutral / Silent** A neutral or silent mutation is the exchange of a nucleotide which either lies in the noncoding region of the genome or, if it lies within an exon of a gene, does not lead to a change in the translated amino acid. The mutation has no effect.

- **Missense** A missense mutation changes an amino acid in the assembled protein, but the resulting protein can either be conservative, meaning it does not change its functionality, or non-conservative if the properties of the proteins are altered, which can lead to a disease or a change of a trait.

- **Nonsense** The last possibility for a SNP is to be translated into a stop codon, which leads to a shortened protein. The protein can be functional or not depending on the sequence of lost amino acids. Usually the protein loses its functionality.

For example, in the human genome approximately every thousandth base pair is altered and these mutations are responsible for the diversity between individuals. Consequently, approximately $3 * 10^6$ SNPs are known in the human genome and more than 10 million SNPs are estimated [34]. Many SNPs occurred during evolution and have been present in the population for a long time. SNPs lying close together in the DNA are less likely to be separated by recombination during meiosis, which is a phenomenon called **linkage disequilibrium** [34, 58]. The fact that proximate SNPs are rarely separated accounts for the high correlations present in the genomic datasets in genome-wide association studies.
SNPs play a key role in the identification of complex traits and diseases, which do not follow the classic rules of inheritance. Therefore, great effort is put into sequencing and identifying the SNPs responsible for traits and diseases.

3.5 Analysis

Analysis of genomic data aims to find variations - mutations - in the DNA sequence influencing a certain phenotype. Hence [35], the primary goal is the identification of the „correct“ subset of SNPs showing similar patterns with the phenotype.
As explained Section 3.4, mutations are an important factor in genetic variability. The analysis of correspondence between mutations and the change in phenotypes reveals new biological connections. According to Thomas [58] and Knippers [34] single-nucleotide polymorphisms are the most common type of mutations; as a consequence, SNPs from the whole genome are analyzed in studies known as

[13] This term is used throughout the thesis for consistency.

genome-wide association studies (GWAS).
According to the National Human Genome Research Institute[14], genome-wide association studies [32] have identified the influence of around 7000 SNPs on various phenotypes. Figure 3.9 depicts the 23 human chromosomes and a selection of the strongest associated SNPs on various phenotypes ranging from diseases to the physical appearance and various other characteristics.
For example, according to the guidelines of the NHGRI [32] genome-wide association studies need to include more than 100,000 SNPs of the human genome in the initial phase to be considered meaningful. Balding [2] states that at least 300,000 SNPs are needed. However, GWAS [23] sometimes include one million or greater SNPs in contrast to the number of individuals sequenced, which ranges from thousands to tens of thousands sequenced. This size of dataset is necessary to capture the genetic variation in the human genome.
The development of high-throughput next-generation sequencing [46] and the decline [48] in genome sequencing costs have facilitated the production of large amounts of genomic data.
GWAS often make use of linkage disequilibrium, as explained in Section 3.4.2, to reduce the amount of data to be analyzed. Linkage disequilibrium [34] is the shared evolutionary history of two SNPs. The closer two SNPs are in the genome, the less likely it is that they are separated during the recombination phase in meiosis. Hence [17], proximate SNPs are often highly correlated; therefore, it is usually enough to include one of the highly correlated SNPs into the study, in order to identify the genomic region influencing the trait under study. The identification of the correlated SNP [42, 61], also called **proxies** or **sentinels**, is often sufficient. Therefore, linkage disequilibrium needs to be somehow estimated [2, 15] in the complete genomic dataset to assess the power of a study. This process [57] is also known as SNP tagging.
According to Donnelly [17] one of the major challenges is to investigate the regions where proxies have been identified more accurately to reveal the true causal SNPs. These studies [17, 44, 56], also known as **fine-mapping studies**, involve many more highly correlated SNPs in the associated region as well as a greater [35] number of individuals sequenced. The causal SNPs then show a higher association than the correlated ones, except for perfectly correlated proxies. Ideally, fine-mapping studies include a great number of correlated SNPs such that results narrow down the set of possible causal SNPs.
Prior to the analysis of the genomic dataset some preceding steps have to be carried out. Preliminary analyses [2] ensure the necessary quality of datasets and aim at avoiding biases and other systematic errors introduced by poor quality of the

[14]The National Human Genome Research Institute(NHGRI) hosts a comprehensive catalogue of genome-wide association studies where associations between genome and phenotypes are available. `http://www.genome.gov/gwastudies/`

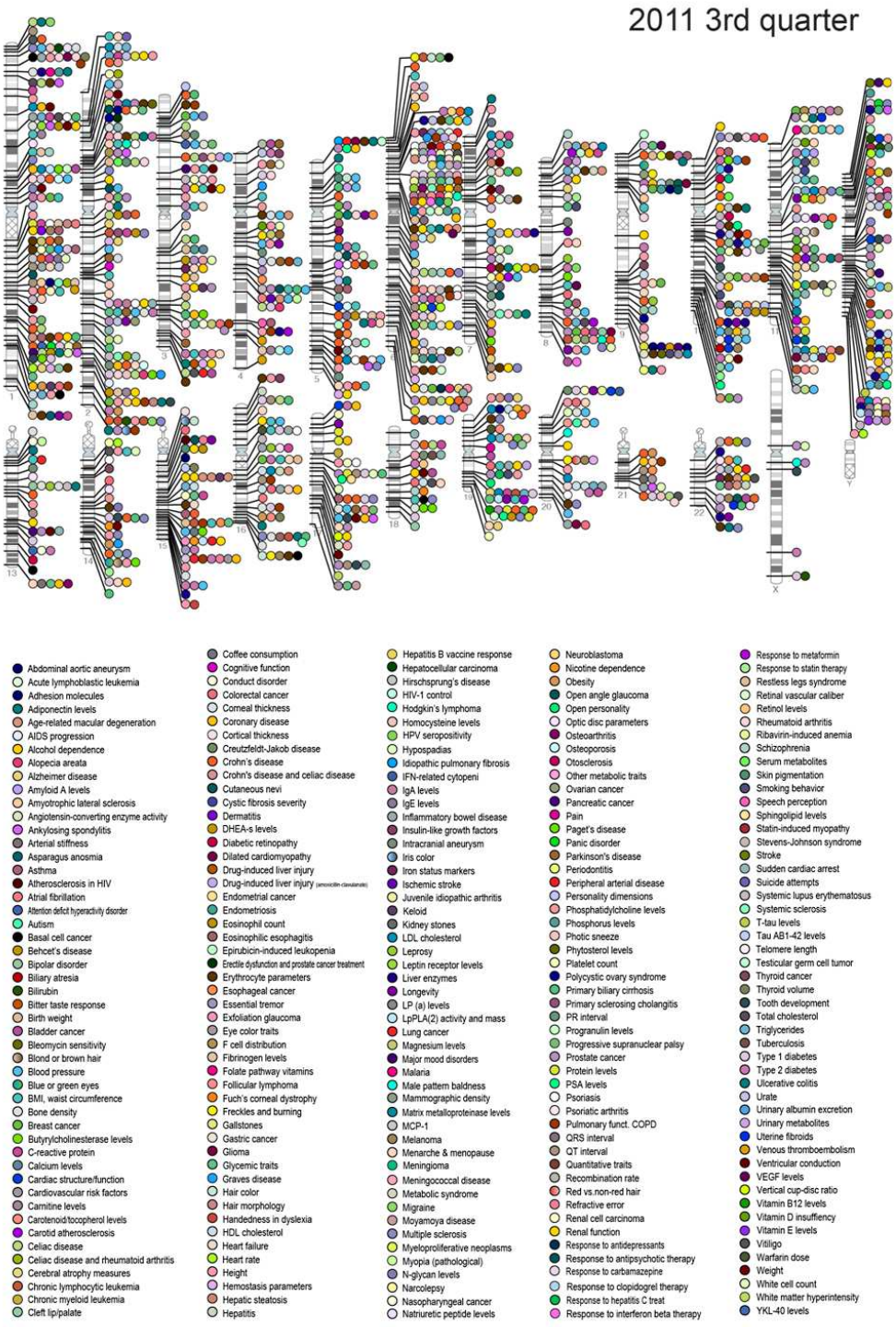

Figure 3.9: SNPs identified through GWAS by 06/2011 - Credit: Darryl Leja and Teri Manolio

dataset. Moreover [42], since different next-generation sequencing technologies do not sequence the exact same set of SNPs, missing values for SNPs need to be imputed [43] to compile them for example into one large GWA study. The necessity of imputing missing SNP values [43], instead of discarding missing data, originates from the improved statistical significance, the enhanced results in fine-mapping studies, the meta-analysis from different datasets as mentioned before and the sporadic missing genotype data from sequencing errors. Missing SNP data do not play a great role in single-SNP analyses [2], but are more problematic in multiple SNP analyses. However, various methods and approaches exists for genotype imputations [43, 53], but the discussion lies beyond the scope of the work.

Genomic datasets usually do not contain the nucleotide sequence contained in the DNA. Instead, assuming a diploid set of chromosomes, the combination of each nucleotide on both DNA sequences is used. The same nucleotide on the same location in both sequences is referred to as homozygote [34]; whereas, different nucleotide are termed heterozygote [34]. Moreover, usually a reference is used for every SNP value and as a consequence two forms of homozygote SNPs exist. The first represents both nucleotide are the same as the reference SNPs, whereas, the second form of homozygte means both nucleotides are different from the reference. Accordingly heterozygosity refers to one of the two nucleotides being different from the reference.

In conclusion, the main challenge of genome-wide association studies is to identify the true single-nucleotide polymorphisms influencing a certain phenotype. The identification is impeded by high correlations between the SNPs due to linkage disequilibrium. A second purpose of GWAS [14] arises from the prediction of phenotypes based on a set of SNPs, which is useful in animal husbandry, for example.
Due to the fact [23, 27, 42] that most causal SNPs have a small effect and that the genome is very large [2], and that patterns and apparent associations can arise by chance, it is unlikely to identify the true subset of SNPs.

3.6 Single-SNP

Single-SNPs regression is the most frequently used approach in GWAS [11, 23], which directly tests the association between a single SNP and the phenotype. Every SNP is examined separately and a strong association is an indication for its influence, or the influence of a proximate correlated SNP (perhaps not included in the analysis), on the phenotype [23]. Single-SNP analyses have the major advantage of being easily parallelizable and can therefor be applied to large genomic

datasets. The intention of this approach is often to identify relevant genes containing the associated SNPs and to glean some insight into the biology of the trait under consideration.
Since the area of application is broad and many studies have applied single-SNP analysis to genomic context; thus, a great range of statistical methods exist.

Depending on the type of measured phenotype the statistical methods vary. A natural and common studies are case-control studies [2], where phenotypes are of binary nature. Therefore, every SNP is tested for the null hypothesis [2] of no association to the phenotype, where usually a 2x3 matrix is considered containing the counts of the two homozygote genotypes and the heterozygote genotypes for control as well as for the case group. Different statistical tests can then be used to test for the acceptance or rejection of the null hypothesis for each SNP.
Another type are continuous phenotypes measuring quantitative characteristics. For this purpose [2] linear regression, where a relationship between mean value of the trait and genotype is tested against the null hypothesis, as well as analysis of variance (ANOVA), where the mean of the three genotype groups[15] are tested for equality, are common choices.
For studies analyzing ordinal phenotypes linear models are adapted to logistic regression where the outcome is categorical. Usually [2] the difficulties arising from non-continuous phenotypes are overcome by transforming the phenotype to a continuous scale using a logit-transformation. Subsequently the three groups of genotypes are again tested for their influence. The null hypothesis is that all three groups have no influence.
Usually the p-value is computed to assess the evidence for an association between a SNP and the phenotype assuming that the null hypothesis is true and for example [32] only SNPs reported with a p-value below 10^{-5} are considered for the NHGRI catalogue. Despite their widespread use [56], the frequentist approach has some limitations, such as the threshold [44] for SNP to be considered as associated as well as as the size of the study and factors like the minor allele frequency (MAF)[16]. This drawback arises from the fact [56] that an association of a SNP with a given p-value does not only depend on how unlikely that p-values is under H_0 but also on how unlikely it is under alternative hypothesis H_1. One response to such issues [56] is to avoid performing tests with low power by for example discarding low-MAF SNPs, which is sometimes inadequate since causal SNPs might be discarded. Uncertainty introduced by imputed data [43], especially rare SNPs, can also degrade the power of frequentist tests and can lead to spuriously low p-values. Another impediment arises from multiple testing [2] and the identification of false positive detections because every SNP is a priori unlikely to be causally

[15] The three groups are again the two homozygote and the heterozygote forms of a SNP.
[16] The minor allele frequency is the frequency of the less common allele of a SNP.

associated. Therefore strong evidence is needed to overcome the skepticism about an association. To control the number of falsely rejected H_0 hypotheses [42, 56] more stringent significance is required as the number of tests increases. This can sometimes lead to less analyses and tests performed [56] to avoid the additional multiple testing penalty imposed.

Alternatively Bayesian methods are used with increasing frequency to alleviate the limitations of p-values but with the drawback of additional modeling assumptions and increasing compuatational demands. Bayesian approaches have the benefit [5, 56] of providing a unified approach to data analysis with uncertainties in the model leading to directly interpretable and comparable results among SNPs within and across studies. Additionally biological knowledge, such as the number of expected true associations, the MAF of every SNP, or the proximity to genes of interest, as well as other prior information can be incorporated and different genetic models can be considered in a single analysis. Furthermore [44, 56], instead of specifying a threshold for the p-values, measures like the **Bayes factor** (BF) or **posterior probabilities** are used. The BF considers the ratio between the marginal likelihoods of the data under H_1 and under H_0. The result can be interpreted as that the observed genomic data are by the Bayes Factor more likely under H_1 than under H_0; the larger the BF, the stronger is the support for H_1. In contrast [56], the posterior probability can be directly interpreted as probability regardless of influences like for example the sample size, the number of analyzed SNPs, or the MAF of every SNP. The posterior probability combines the evidence that a SNP is associated with the phenotype based on the data as well as the prior knowledge assumed.
A common used advantage of the Bayesian approach is the averaging over different genetic models[17].
However, the need to specify prior knowledge can lead to spurious and distorted results.

An inherent drawback of the frequentist and the Bayesian single-SNP analysis is [11] that only single-SNP effects are identified and epistatic effects and groups of associated SNPs are neglected. The approaches outlined in this Section work well for traits strongly influenced by only a single or a few SNPs but are not able to reveal the biology of complex traits. More complex Bayesian approaches found to perform superior [23, 39] even in the case of a few causal SNPs. Nevertheless, more sophisticated methods are computationally more demanding [48] and are not yet able to identify the majority of causal SNPs [17]. The thesis contributes to the improvement of this issue and to identify the more promising method.

[17] Usually [56] different genetic models represent additive, dominant or recessive genetic effects and are incorporated for every SNP using different weights.

CHAPTER 4

Statistical Background

THE current chapter outlines common statistical methods relevant for the analysis using large-scale regression methods applied to genome-wide association studies to model the complexity of genomes and to alleviate the computational burden arising therein. Hierarchical models, describing complex relationships and processes, together with Markov chain Monte Carlo methods which facilitate the analysis, are frequently used [10] in GWAS among various other applications.

Section 4.1 gives an outline of Bayesian hierarchical models and their ability to model complex contexts. In Section 4.2 the concept of Markov chains is first explained and second an introduction to Markov chain Monte Carlo methods is given with respect to practical applications and with focus on the Metropolis-Hastings algorithm in Subsection 4.2.1 and the Gibbs sampler in Subsection 4.2.2.

4.1 Bayesian Hierarchical Models

Bayesian inference provides the possibility to combine prior beliefs with observed data to obtain knowledge about underlying stochastic processes as well as its uncertainty. Bayesian hierarchical models represent the dependencies of random variables from which inferences is made.
An import characteristic of Bayesian hierarchical models [10] is the ability to model a great variety of complex processes and interrelationships between stochastic components and to capture the behavior of the processes with respect to the inherent uncertainty. Bayesian methods are becoming more popular in genome-wide association studies [2, 4] because of the improving means for tackling the high computational demands [56] as well as the unified approach of data analysis. The purpose of the current Section is to outline the ideas of Bayesian statistics

and inference necessary for the subsequent Sections. More details can be found in Bishop [5] and Carlin *et.al.* [10] for example.

Bayesian models provide a unified approach [10] to data analysis and inference as well as a consistent way to incorporate prior beliefs into the model. Uncertainty captured in the model includes both uncertainty in the data as well as in the model parameters. To achieve this both the observed data as well as any unknown or latent variables are modeled as random variables having a probability distribution.
The main difference to the frequentist approach [5] is that the Bayesian model summarizes the observed data with respect to prior beliefs in the form of **posterior probabilities**, instead of making point estimates and estimating the uncertainty separately. The posterior probability is obtained by

$$\textbf{posterior} \;\propto\; \textbf{likelihood}\; x\; \textbf{prior}. \tag{4.1}$$

Bayesian hierarchical models yield the attractive feature that uncertainty is propagated through the complex models, affecting the certainty of inferred posterior probabilities.
The central paradigm of Bayesian statistics is the Bayes theorem [5], shown in Equation 4.2, which converts prior beliefs about the variables in the model into a posterior distribution by incorporating information contained in the observed data.

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} \tag{4.2}$$

In Equation 4.2 θ represents a model parameter and $p(\theta)$ is considered to be a prior belief of the probability of certain values of θ. $p(Y|\theta)$ is the **likelihood function** and expresses the probability of observing a dataset under the parameter θ. $p(Y)$ is the probability of the dataset to be observed and is usually obtained by marginalization. It ensures the left-hand side of Equation 4.2 is a valid probability distribution integrating or summing to 1. The posterior probability [5] $p(\theta|Y)$ summarizes the knowledge obtained from the prior beliefs and the likelihood of the observed data given the prior beliefs for the parameter. The posterior captures the uncertainty in the parameter after the data have been observed.
The prior is often governed by another parameter [10], also termed **hyperparameter**, which is mostly unknown; therefore, a second stage [10] is introduced to assign **hyperprior** $p(\lambda)$ to the parameters of the prior distribution. The Bayes theorem is then augmented with the hyperprior as depicted in Equation 4.3[1]

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta|\lambda)p(\lambda)}{\int p(Y|\theta)p(\theta|\lambda)p(\lambda)d\lambda} \tag{4.3}$$

[1]The same equation holds for the discrete random variables where the integral is substituted by a summation.

As a side note, a delicate issue [5] of the Bayesian approach arises from the choice of priors, because the selection of priors is often based on convenience and subjective believes. Stephens *et.al.* [56] note that it is often desirable to avoid subjectivity in the form of noninformative priors, but that the real problem is the hidden subjectivity and missing clarification of the assumptions. However, this discussion lies beyond the scope of this thesis.

Due to the complexity of most models [10], the integral in Equation 4.3 is often not tractable and cannot be solved analytically, which is required for the inference from the hierarchical model. Special forms of priors called **conjugate priors**[2] alleviate the analytic evaluation. Nevertheless [5,10], intractable integrations remain and need to be approximated.
Other situations [1] require the optimization of the posterior distribution; this is, the identification of the optimal values for the parameters. Often exhaustive computation of the posterior distribution is infeasible as it is impossible to compute and compare all solutions. A more detailed discussion is referred to Chapter 5 and Chapter 6.
Popular methods [5] exploring the posterior distribution are **Markov chain Monte Carlo** algorithms [62], which obtain samples by directly sampling from the distribution. A major advantage [10] is that, making use of high-speed computing equipment, high-dimensional distribution can be accurately approximated.

4.2 Markov Chain Monte Carlo

As outlined in the previous Section, Bayesian methods require the computation of the posterior probability to enable inference of stochastic processes. Computational challenges [4, 10] arise from intractable integrations, especially as occuring in Bayesian hierarchical models, and from the identification of the best values in optimization problems.
The most popular computational tools [10] are Markov chain Monte Carlo (MCMC) methods due to their ability to enable inference even in high dimensional posterior distributions, and to, for all intents and purposes, break the curse of dimensionality[3]. MCMC methods do not produce a closed-form solution but instead simulate draws of samples from the posterior distribution, thereby generating a **Markov chain**.
Although [10] these samples do not contain as much information as a closed form

[2] Conjugate priors are prior probability distributions that belong to the same family of probability distribution as the posterior probability distribution.

[3] The „curse of dimensionality“ [5, 10] states that, with increasing dimensions, the amount of data to be computed, for example the full posterior distribution, grows exponentially with the dimensionality.

solution the estimation that results from sampling can be made arbitrarily accurate by increasing the number of samples drawn from the posterior distribution. A summation of the posterior distribution is usually sufficient to approximate the posterior distribution enough to allow reliable inference. As for most situations [5] the identification of the most probable values is satisfactory.
As a side note, a common criticism [10] of MCMC methods is that no two inferences will obtain the same approximation since different samples are drawn from the posterior distribution.

MCMC explores the distribution by simulating random draws from the the target distribution $\pi(\theta)$ resulting in a Markov Chain $\theta^0, ..., \theta^t$. This exploration is also called a **Markov process**. An important property [62] during the construction of the Markov chain is that each value θ^i for $i = 1, ..., t$[4] only depends on the preceding state, which can be seen from Equation 4.4.

$$p(X_{t+1} = s_{t+1}|X_0 = s_0, ..., X_t = s_t) = p(X_{t+1} = s_{t+1}|X_t = s_t) \tag{4.4}$$

This means that the only information necessary for obtaining the next sample θ^{t+1} from the distribution $\pi(\theta)$ is the current state θ^t. The Markov chain also needs a transition probability [16] T, as noted in Equation 4.5, which defines the probability of the transition from a current state θ^t to the next state θ^{t+1} such that the desired distribution is invariant.

$$T(\theta^t, \theta^{t+1}) = T(\theta^t \rightarrow \theta^{t+1}) = p(\theta^{t+1} = s^{t+1}|\theta^t = s^t) \tag{4.5}$$

A fundamental theorem of Markov chains states [16] that from any given starting point θ^0 the Markov chain has a probability of $\pi(\theta^t)$ of being in the state θ^t after sufficiently large number of steps. Hence, the probability of a state only depends on the probability $\pi(\theta^t)$ of reaching θ^t and is independent from the initial value θ^0. This property [5] is called **ergodicity** and thus the Markov chain is said to have a **stationary**, or **equilibrium**, distribution which corresponds to the distribution to be approximated. Consequently, independent of the initial value θ^0 the Markov chain will, after a finite number of steps, sample directly from the equilibrium distribution. Each Markov chain can only have one equilibrium distribution.
Moreover, besides ergodicity, Markov chains need to have certain other properties in order to converge to the invariant distribution $\pi(\theta)$. First [1], the Markov chain has to be **irreducible**, which means that from any given state θ^t there must be positive probability to reach all other states possible for θ^{t+1}. Hence, each state, in the case of discrete states, and each value, in the case of continuous states, is reachable with a certain probability greater than 0. Second [1], the Markov chain

[4]Note that the first sample θ^0 does not depend on any other state but is rather assigned an initial value.

needs to be **aperiodic** such that the chain cannot get trapped in cycles.
Given the properties mentioned [1] the MCMC sampler is a Markov process generating a Markov chain that has the target distribution $\pi(\theta)$ as its equilibrium distribution.

After convergence of the Markov chain to the stationary distribution, which is addressed later in this section, the samples drawn are used to summarize the posterior distribution and thus allow inference.
The samples obtained can be summed in any way [37]; however, common choices [10] are the posterior mean in Equation 4.6,

$$\hat{\theta} = \mathcal{E}[\theta|Y] \tag{4.6}$$

the posterior median in Equation 4.7,

$$\hat{\theta} = \int_{-\inf}^{\hat{\theta}} p(\theta|Y)d\theta \tag{4.7}$$

or for example the posterior mode in Equation 4.8.

$$\hat{\theta} = \sup_{\theta} p(\theta|Y) \tag{4.8}$$

Moreover, the uncertainty captured can be assessed by an estimation of the variance of the samples. The most probable regions of the distribution [5] can be identified after a rather small number of samples; however, to approximate the tails of the distribution a much larger number of samples is required. However, the accuracy of the estimation does not solely depend on the dimensionality of θ, but on the number of samples.

The convergence of the Markov chain to the desired equilibrium distribution [10] can be ensured for large number of posterior distribution. Nevertheless [10], a crucial point for the application of MCMC methods is the decision of when it is acceptable to stop sampling from the equilibrium distribution to obtain a sufficient approximation of the distribution. This issue is also called **convergence diagnosis** and deals with the estimation of the point when the Markov chain directly samples from the equilibrium function and enough samples have been obtained.
The first samples θ^0 to θ_{BurnIn} are referred to as the **burn-in** period [1], where the Markov chain has not yet reached stationarity and thus the samples do not represent direct draws from the target distribution. The length of the burn-in period is difficult to assess and depends on the pace of the chain moving away from the initial value as well as the autocorrelation of the chain which will be discussed shortly.
A common solution [1, 10] is to remove the beginning of the chain and to start using the subsequent samples. In practice, however, it is difficult to assess the

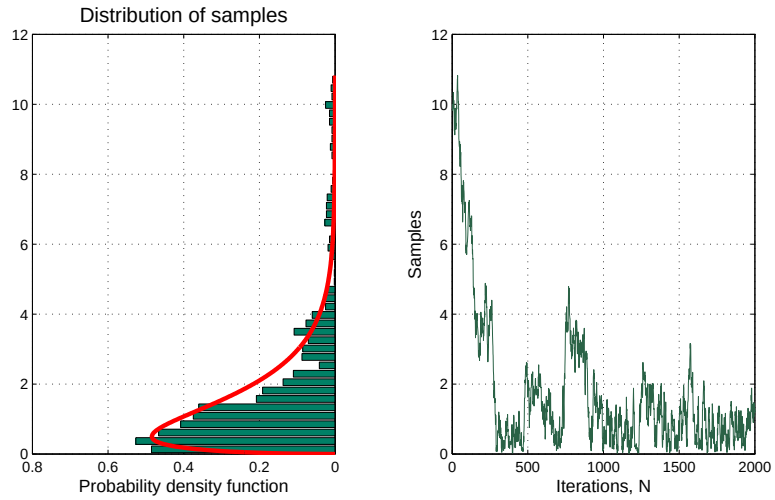

Figure 4.1: An example for a Markov chain where the around the first eighth of the sequence is the burn-in period

best length of the burn-in period because no completely reliable diagnostics [1] for convergence exists.

An example for the burn-in period can be seen in Figure 4.1. In the figure on the left, the distribution to be approximated is shown as a red curve; whereas, the greenish bars represent the summarized samples. The right-hand figure shows the samples of the Markov chain also referred to as **trace**. It can be easily seen that the beginning of the chain is the burn-in period where the Markov chain has not yet converged to the equilibrium distribution. The samples following are then considered to be direct draws from the target distribution and can be used for the estimation of the distribution. Obviously, an optimal initial state would be close to the center of the probability distribution.

Another difficulty arises from the quality of samples obtained. Ideally [1], the samples drawn by the MCMC methods should be i.i.d samples[5]. As shown in Equation 4.4 the samples, since each sample depends on the previous state, are not independent draws and thus are expected to be positively correlated [10] also referred to as the **autocorrelation**. Nevertheless, even though these samples are correlated it can be shown [62] that the draws are still from the equilibrium distribution and therefore present an unbiased picture of the distribution, if the number of samples is sufficiently large. The higher the autocorrelation of a Markov chain, the more samples are required to obtain the same accuracy. For example [62], a very high autocorrelation can require up to a few hundred times more samples to obtain the same accuracy as if i.i.d. samples are available.

[5]Independently and identically distributed samples are samples all drawn from the same probability distribution and do not influence one another.

A common approach to deal with autocorrelation is to **thin** [1] the Markov chain; that is, only using every i^{th} sample from the chain, thereby reducing the correlation between the samples until it becomes insignificant. However, a common criticism [10] is that thinning is not favourable since it increases the variance of the estimation. Instead, using all samples [10] is a more preferable approach along with the estimation of the **effective sample size (ESS)** which is usually much smaller [5] than the total number of samples. Equation 4.10 shows the estimation of the ESS where $\rho_k(\theta)$ refers to the autocorrelation at a distance k.

$$ESS = \frac{N}{\kappa(\theta)} \tag{4.9}$$
$$\kappa(\theta) = 1 + 2\sum_{k=1}^{\inf} \rho_k(\theta)$$

The issue of the effective sample size is closely related with the issue of when the Markov chain has reached the equilibrium distribution as well as when a sufficient number of samples has been collected.
Figure 4.2 depicts the difference between the approximations using samples sizes of 1000, 5000, 10000, and 25000 from the upper left to the bottom right figure. As the number of samples increases, the approximation becomes more and more accurate. Note from the right bottom plot that the tail has been explored by the Markov chain which has not happened in the Markov chain with the lower number of samples.
Various approaches exist to estimate the convergence of the chain for example the **Geweke test** [62] which compares the mean of the first 10% of the chain with the mean of the last 50% of the chain and compares for their equality using a hypothesis test, or the **Raftery-Lewis test** [62].
The most popular approach for diagnosis on the MCMC samples [19] is the **Gelman-Rubin diagnostic** $\hat{R}$, also known as **potential scale reduction factor** [10]. It estimates the equality of the variation within the sequence of samples [10], usually obtained by running many parallel Markov chains. Ideally $\hat{R}$ should be 1 as the number of samples converges to ∞.
An important fact of Markov chains is that as long as all requirements, such as ergodicity and aperiodicity, are fulfilled, the Markov chain always converges to the equilibrium distribution and the samples can be used for approximation. However, the time until convergence and the pace of the chain exploring the equilibrium distribution are the crucial factors for an efficient approximation.

4.2.1 Metropolis-Hastings-Algorithm

Among the most frequently used MCMC methods [1] is the **Metropolis-Hastings** algorithm. The algorithm was first introduced by Metropolis, Rosenbluth, Rosen-

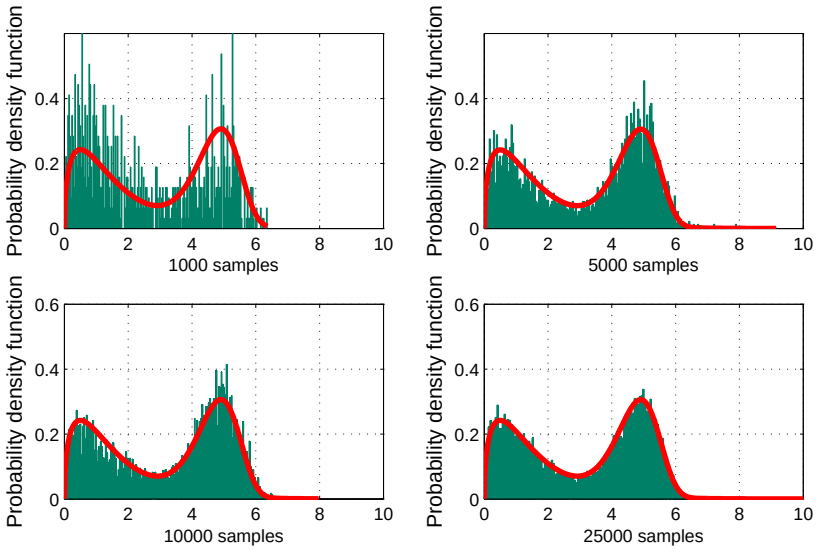

Figure 4.2: Approximation of an arbitrary distribution by a Markov chain using samples sizes of 1000, 5000, 10000, 25000

bluth, Teller and Teller [45] and was generalized by Hastings [31]. As outlined in the previous section, MCMC methods build a Markov chain by generating samples from the equilibrium distribution. The Metropolis-Hastings algorithm is a rejection algorithm [10] as it circumvents the problem of sampling directly from the distribution by generating new samples given the current state of θ. It subsequently accepts or rejects the newly proposed sample with a certain probability. The Metropolis-Hastings algorithm offers more flexibility [10] in contrast to other MCMC methods due to the variety of proposal distributions; however, only a careful choice yields a quickly converging chain.
Therefore, a Markov transition kernel [22] also called **proposal density** [10], **proposal**, or **candidate-generating distribution** [62] $q(\theta^*|\theta^t)$, as defined in Equation 4.5, is used to generate a new candidate sample.
Subsequently the proposed sample is either accepted or rejected with the probability [1] given by Equation 4.10. If the sample is rejected [5], then the previous sample is used as the current state, leading to multiple copies of samples.

$$\alpha_{MH} = min\left\{1, \frac{q(\theta^t|\theta^*)\pi(\theta^*)}{q(\theta^*|\theta^t)\pi(\theta^t)}\right\} \tag{4.10}$$

As can be seen in Equation4.10 the proposed sample θ^*, having a higher probability than the current sample is always accepted; whereas, in the case of the new sample, is accepted with α_{MH}. The complete Metropolis algorithm is shown by Algorithm 4.1.

Algorithm 4.1: Metropolis-Hastings algorithm

Result: Markov chain of length N generated by Metropolis algorithm

$\theta^0 = \textit{Initial Value}$
for $i = 1 : N$ **do**
  $\theta^* = q(\theta^*|\theta^t)$
  $u = \mathcal{U}_{[0,1]}$
  $\alpha_{MH} = min\left\{1, \frac{q(\theta^t|\theta^*)\pi(\theta^*)}{q(\theta^*|\theta^t)\pi(\theta^t)}\right\}$
  if $u < \alpha_{MH}$ **then**
    `// Accept proposed sample`
    $\theta^{t+1} = \theta^*$
  else
    `// Reject proposed sample`
    $\theta^{t+1} = \theta^t$
  end
end

A common variation [5] to the Metropolis-Hastings algorithm is the **Metropolis algorithm**. It is used [22] if the proposal distribution is symmetric; hence, $q(\theta^*|\theta^t) = q(\theta^t|\theta^*)$. For example, in the case of a normally distributed proposal distribution. The acceptance probability simplifies to Equation 4.11.

$$\alpha_M = min\left\{1, \frac{\pi(\theta^*)}{(\pi(\theta^t)}\right\} \tag{4.11}$$

The crucial point for efficient approximation by the Metropolis-Hastings algorithm[6] is the specified proposal distribution [5] $q(\theta^*|\theta^t)$, which is chosen to be easy to generate candidate samples θ^* from. The proposal distribution is an important tuning parameter [62] and strongly influences the speed of convergence to the equilibrium distribution as well as the quality of approximation with a limited number of samples.

An example is shown in Figure 4.4 where three choices of proposal distributions are compared. For each column, 7,000 samples are computed using a normal distribution as the proposal distribution.

The column on the left is computed using a narrow proposal distribution $q(\theta^*) \sim \mathcal{N}(\theta^i|0.1)$. Note that the samples are strongly centered around the mode of the target distribution and the chain baby-steps around the center. Note that the chain has an acceptance ratio[7] of around 0.985. Although only a few samples are not

[6]This also accounts for the Metropolis algorithm as well as for other variations of the Metropolis-Hastings algorithm.

[7]The acceptance ratio is the proportion of samples accepted in contrast to the total number of samples.

used, the chain only explores the most probable regions and problems can arise from distant peaks not explored sufficiently. A too-narrow proposal distribution can also get trapped in local peaks such that stationarity seems to be reached. Furthermore, it also leads to a high autocorrelation in the Markov chain, slowing down convergence.

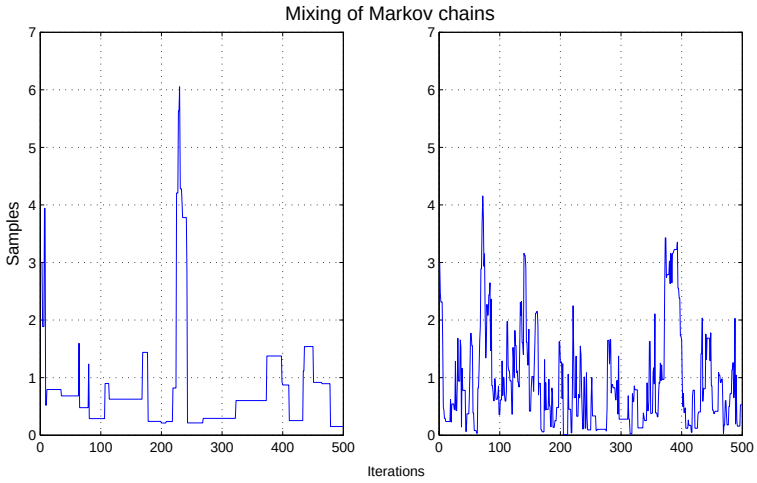

Figure 4.3: Difference between a poorly mixing (left Figure) and a well mixing Markov chain (right Figure).

In Figure 4.4, the middle column shows an overly wide proposal distribution $q(\theta^*) \sim \mathcal{N}(\theta^i|15)$, generating proposal samples far off from the current state which are likely to lie far from the distribution's center [10]. This leads to a high number of rejections (in this particular example the acceptance rate was only around 0.13, thus 87% of the samples are discarded) which in turn leads again to high autocorrelation. The chain is said to be **poorly mixing**. The difference between a poorly mixing and a **well mixing** Markov chain is shown in Figure 4.3. Note that the figure on the left shows long flat periods where the samples are rejected. A well mixing chain [62] looks similar to white noise.
The third column in Figure 4.4 shows a well mixing chain using a proposal distribution $q(\theta^*) \sim \mathcal{N}(\theta^i|4)$. In this case the acceptance ratio is around 0.4 which is considered to be favourable [10].

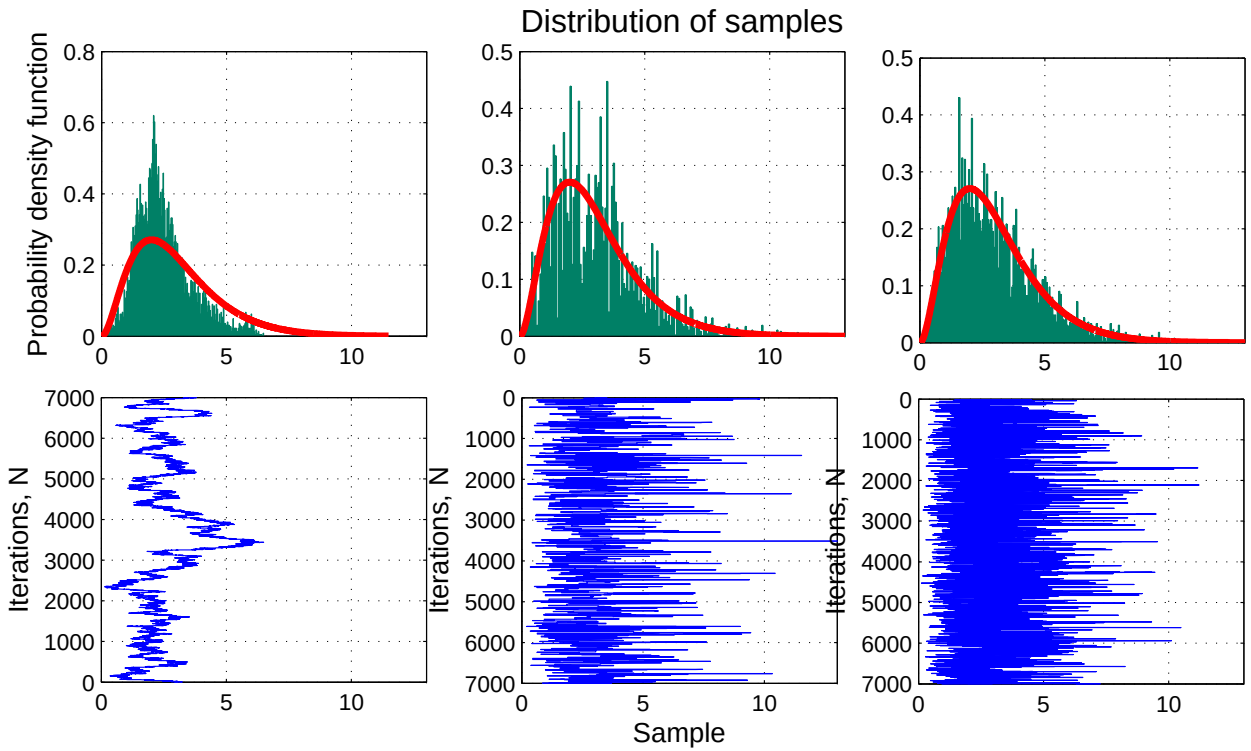

Figure 4.4: Various choices of the proposal distribution for a Metropolis-Hastings algorithm

4.2.2 Gibbs-Sampling

Besides Metropolis-Hastings the Gibbs sampler is a widely applicable [5] and common MCMC method which can be regarded as a special case [62] of the Metropolis-Hastings algorithm, having an acceptance ratio of 1. Here, no proposal distribution has to be tuned to obtain a well mixing Markov chain. All samples are used and none are rejected, thus saving computation time. The Gibbs sampler is used [29] when it is difficult to sample from the full joint distribution, but feasible to sample from the conditional distributions of every variable θ_i. Therefore [62], in contrast to the Metropolis-Hastings algorithm, the Gibbs sampler uses an univariate conditional distribution for each variable, where all variables but one are assigned fixed values, as the proposal distribution as shown in Equation 4.12. Thus [62], in every iteration all θ_i for $i = 1, ..., p$ are sampled from their univariate conditional distribution rather than to generate θ from the full joint distribution. This process is repeated N times until the Markov chain converges and the samples, after the burn-in period is removed, are used to approximate the target distribution.

$$p(\theta_i|\theta_1, ...\theta_{i-1}, \theta_{i+1}, ..., \theta_p) \tag{4.12}$$

The complete procedure for the Gibbs sampler is shown by Algorithm 4.2.
It is assumed [10] that in practice all conditional distributions uniquely determine the target distribution. However [10], if the conditional distribution is not conjugate and thus not available in closed form it is favourable to employ a Metropolis-Hastings algorithm.

Algorithm 4.2: Gibbs sampling algorithm

Result: Markov chain of length N generated by Gibbs sampler

$\theta^0 =$ *Initial Value*
for $i = 1 : N$ **do**
  Sample $\theta_1^{i+1} \sim p(\theta_1|\theta_2^i, ..., \theta_p^i)$
  Sample $\theta_2^{i+1} \sim p(\theta_2|\theta_1^{i+1}, \theta_3^i, ..., \theta_p^i)$
  .
  .
  .
  Sample $\theta_j^{i+1} \sim p(\theta_j|\theta_1^{i+1}, ..., \theta_{j-1}^{i+1}, \theta_{j+1}^i ..., \theta_p^i)$
  .
  .
  .
  Sample $\theta_{p-1}^{i+1} \sim p(\theta_{p-1}|\theta_1^{i+1}, ..., \theta_{p-2}^{i+1}, \theta_p^i)$
  Sample $\theta_p^{i+1} \sim p(\theta_p|\theta_1^{i+1}, ..., \theta_{p-1}^{i+1})$
end

Figure 4.5 shows an example of an applied Gibbs sampler where a bivariate Normal distribution

$$\mathcal{N} \sim (\begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix})$$

is to be approximated. The conditional distributions are given by Equations 4.13 and 4.14. The Gibbs sampler is run for 2,000 iterations.

$$p(x^{i+1}|y^i) \sim \mathcal{N}(\mu_1 + \rho(y^i - \mu_2), \sqrt{1-\rho^2}) \tag{4.13}$$

$$p(y^{i+1}|x^{i+1}) \sim \mathcal{N}(\mu_1 + \rho(x^{i+1} - \mu_2), \sqrt{1-\rho^2}) \tag{4.14}$$

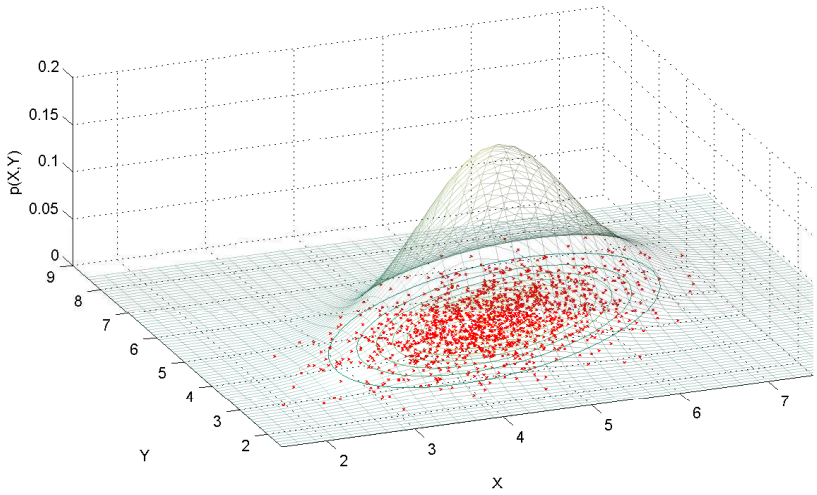

Figure 4.5: Result from a Gibbs sampler after 2000 iterations on a bivariate Normal distribution

CHAPTER 5

Methods

As outlined in Chapter 3 only a small number of single-nucleotide polymorphisms are considered to influence a phenotype. The purpose of analysis, as explained in Section 3.5, is to identify the true subset of SNPs influencing the phenotype on one hand and on the other hand to predict phenotypes based on a dataset using the identified subset. The former point [22] naturally arises from genomic datasets where SNPs with no or only a negligible influence on the phenotype can be excluded from the set of relevant variables. However [23], identification of the single best model is very unlikely to be successful because of small associations between SNPs and phenotypes and datasets with more predictors than observations $p \gg n$. Additionally [36], the datasets, as explained in more detail in Section 3.4.2, are often highly correlated.
The thesis considers two common methodologies, namely stochastic search variable selection (SSVS) and Bayesian penalized regression. Both methods are applied to genome-wide association studies; for SSVS see for example Guan *et.al.* [23], Chen *et.al.* [11], Srivastava *et.al.* [54] or Yi *et.al.* [66]. Bayesian penalized regression is used for example by Li *et.al.* [39], Silva *et.al.* [51], using Bayesian Lasso for the prediction of unseen traits, or Yi *et.al.* [67]. Methods considered in this chapter are able to circumvent the restrictions arising from single-SNP regression as outlined in Section 3.6, by considering the complete dataset and thereby combinations of SNPs for their association. Moreover, because of the Bayesian representation [37] easily-interpretable results in combination with valid standard errors are obtained and methods are able to partially model the complexity of the genome and its influences on traits and phenotypes.

The purpose of this chapter is to introduce and outline the methodology considered within this thesis and to propose modifications to the methods under consideration. Section 5.1 gives a general introduction to the regression problem as it is assumed

in most of the work about genome-wide association studies. Section 5.2 introduces hybrid correlation-based search with its two parts, stochastic search variable selection in Subsection 5.2.1 and correlation-based search in Subsection 5.2.2; whereas, the modifications are discussed in Subsection 5.2.3. Subsequently, Section 5.3 explains Bayesian penalized regression with Bayesian lasso in Subsection 5.3.1 and Bayesian ridge regression is discussed in Subsection 5.3.2.

5.1 General Regression Model

In this chapter details of the methods examined are presented. In case of hybrid correlation-based search and Bayesian penalized regression a multivariate linear regression model is considered consisting of n observations and p predictors, which is the natural choice [2] for this purpose and is common [23, 27, 36, 37, 49] among related work. The linear contribution of every causal SNP [2, 63, 69] is widely adopted.
Hastie *et.al.* [29] mentioned that linear models are a reasonable approach in situations with a small number of training cases and data with low patterns of associations.

$$Y = \alpha + X^{*'}\beta + \epsilon \tag{5.1}$$

Let $X = \{X_1, X_2, ..., X_p\}'$ denote the standardized predictor variables, an nxp matrix. In the scope of GWAS predictor variables are usually single-nucleotide polymorphisms, where every predictor represents one SNP from the genomic dataset. The combination of the two alleles from a genetic location on the chromosomes[1] determines the value of a predictor, as explained in Section 3.6.
Equation 5.1 depicts the assumed linear model where a subset of all SNPs contributes and influences the measured outcome (the phenotype). X^* is a small subset of X with $X^* = \{X_1, X_2, ..., X_{p^*}\}$, where $p^* \ll p$, associated to the outcome Y. Hence, Y is a linear combination of the predictors in X^*. Note that there are 2^p different combinations of X^*. According to O'Hara [27], in a Bayesian framework the selection of the 'best' subset is often determined in a variable-specific form where every variable is either included or excluded from the subset. More details are discussed in the remainder of this section.
The strength of the influence of every genetic location (SNPs) on the phenotype is represented by its regression coefficient β_i, hence $\beta = \{\beta_1, \beta_2, ..., \beta_p\}'$. The purpose of variable selection [22] (selecting strongly associated genomic locations) is to identify the group of variables with small regression coefficients, where it would be preferable to ignore them and instead to include variables in X_* having a regression coefficient different than 0.

[1] Assuming a diploid set of chromosomes, therefore every chromosome is present in every cell twice.

The last term in Equation 5.1 is ϵ and depicts the independent error term, also known as noise.
It is assumed, as typical for GWAS, that the number of predictors is larger than the number of observations, therefore p$\gg$n.

5.2 Hybrid Correlation-based Search

The hybrid correlation-based search (hCBS) [36] is an iterative stochastic search method comprising two parts. The first part follows the stochastic search variable selection [7, 8, 21, 22]; while, the second part is a newly proposed method named correlation-based Search (CBS). Both methods are used in the hybrid correlation-based search method, where in every iteration either the stochastic search variable selection or the correlation-based search method is used. The method is designed to identify a subset X_γ to approximate Equation 5.1 by Equation 5.2; hence, to approximate the 'best' model X^*. To this end it is given a set of predictor variables $X = X_1, ..., X_p$ and an outcome variable Y depending on X, as defined in Section 5.1. γ indicates which variables are included in the current subset by setting $\gamma_i = 1$ if the variable X_i is included in the subset and $\gamma_i = 0$ if not. The purpose of this is explained later in this Section.

$$Y = \alpha_\gamma + X_\gamma^{'}\beta_\gamma + \epsilon \tag{5.2}$$

The noise term in Equation 5.1 and Equation 5.2 is defined as Normal distribution $\epsilon \sim \mathcal{N}(0, \sigma^2)$. α is a normal distribution $\mathcal{N}(\alpha_0, h\sigma^2)$; where α_0 and h are hyperparameters. The variance of the noise term σ^2 is assigned an inverse gamma distribution as shown in Equation 5.3.

$$\sigma^2 \sim \mathcal{IG}(\frac{v}{2}, \frac{v\lambda}{2}) \tag{5.3}$$

Furthermore, a conjugate prior for the regression coefficients β_γ is used, as shown in Equation 5.4, given the current subset γ and σ^2. For H_γ an independent prior cI_{p_γ} is used, since it is computationally favourable [36].

$$\beta_\gamma|\gamma, \sigma^2 \sim \mathcal{N}(\beta_0, \sigma^2 H_\gamma) \tag{5.4}$$

Given σ^2, α is a normal distribution with $\mathcal{N}(\alpha_0, h\sigma^2)$ with hyperparameters α_0 and h. The posterior distribution [36] can by calculated by using the specified priors and gathering the information about the most probable subsets.

$$\begin{aligned} p(\gamma|X, Y) \propto g(\gamma) = \\ = |I_n + X_\gamma H \gamma X_\gamma^{'}|^{-\frac{1}{2}} |Q_\gamma|^{-\frac{v+n}{2}} p(\gamma) \\ \text{where} \quad Q_\gamma = v\lambda + Y^{'}(I_n - X_\gamma K_\gamma^{-1} X_\gamma^{'})Y \\ \text{and} \quad K_\gamma = X_\gamma^{'} X_\gamma + H_\gamma^{-1} \end{aligned} \tag{5.5}$$

Since the posterior distribution has to be evaluated for 2^p different models in order to find the γ with the largest posterior distribution $p(\gamma|X,Y)$, computation becomes infeasible with larger values of p.
In hybrid correlation-based search the vector γ can be used to obtain the most-oft included variables in the subset. Since $\gamma_i = 1$ for every variable included in each iteration the variable selection can be performed simply by computing the ratio of a variable being included to the number of iterations of the Markov chain. The result obtained is called **mean posterior inclusion probability** and is addressed in Chapter 6.
The remainder of this chapter includes definitions of stochastic search variable selection in Subsection 5.2.1 and the correlation-based search in Subsection 5.2.2.

5.2.1 Stochastic Search Variable Selection

Stochastic search variables selection randomly explores a fraction of the possible models of γ to identify the model with the largest posterior probability $p(\gamma|X,Y)$. As previously outlined [23, 36], SSVS does not incorporate any information about the relationships between variables for the generation of a new subset γ. At any iteration of the Markov Chain, SSVS alters the subset, from the preceding iteration. Therefore, a new vector γ^* is created from the current γ by either adding or removing a randomly chosen predictor from the current subset with probability ϕ. With probability $1-\phi$, one predictor that is currently included in the subset is being exchanged for a randomly chosen predictor that is currently excluded. This leads [36] to the following proposal distribution

$$q(\gamma^*|\gamma) = \begin{cases} \frac{\phi}{p}, \text{if} |p_\gamma - p_{\gamma^*}| = 1 \\ \frac{1-\phi}{p_\gamma(p-p_\gamma)}, \text{if} |p_\gamma - p_{\gamma^*}| = 0 \end{cases} \tag{5.6}$$

George and McCulloch [21] introduce a widely adopted prior for γ taking the form of an independent Bernoulli distribution as depicted in Equation 5.7, where p_γ denotes the number of variables currently selected into the subset; $p_\gamma = \sum_{i=1}^{p} \gamma_i$.

$$p(\gamma) = \omega^{p_\gamma}(1-\omega)^{(p-p_\gamma)} \tag{5.7}$$

ω is considered as a prior assumption [6–8, 21, 27] of the size of the subset, more specifically the ratio of variable included into the selected subset to the total number of variables. In the majority of GWAS [17] the number of expected SNPs associated and therefore relevant has shown to be rather small, hence ω is set to a small value.

5.2.2 Correlation-based Search

Correlation-based search uses a similar approach to stochastic search variable selection, except that correlation-based search does not consider every variable as

independent. As outlined in Section 3.4.2, genomic data show high correlations. Not considering correlation during variable selection [36] can result in the inclusion of highly correlated variables at the cost of variables being ignored, which are part of the true underlying subset.
SSVS [36] is modified to incorporate information about relationship between variables for the proposal of new subsets γ. While SSVS chooses the variables for the inclusion/exclusion step as well as the swap step randomly, CBS considers correlation between variables in every iteration of the Markov Chain to propose the altered subset γ. Therefore, only variables having a low correlation are added to the current subset; whereas, highly correlated variables are excluded from the current subset.
Let Υ_X denote the correlation matrix of predictors X with entries $\Upsilon_{X_{ij}} = \rho_{ij}$. $\mathcal{L}_\gamma$ representing the predictors currently included in the subset, hence $\mathcal{L}_\gamma = \{i : \gamma_i = 1, i = 1, ..., p\}$ and respectively $\mathcal{E}_\gamma$ denote the set of predictors excluded.
During the addition step an index $i' \in \mathcal{L}_\gamma$ is randomly chosen, and subsequently the predictor $x_{j'}$, where j' satisfies $\{j \in \mathcal{E}_\gamma : |\rho_{i'j'}| = min|\rho_{i'j}|\}$, is included in the subset. The deletion move is similar, therefore $x_{j'}$ satisfying $\{j \in \mathcal{E}_\gamma : |\rho_{i'j'}| = max|\rho_{i'j}|\}$ is excluded. The swap move is simply a combination of an addition and a removal step. Due to these changes, components of γ are no longer independent Bernoulli variables and therefore the prior is modified in Equation 5.8.

$$p(\gamma) = \binom{p}{p_\gamma}^{-1} \frac{1}{p_\gamma} \tag{5.8}$$

Consequently [36], as denoted in Equation 5.9, the proposal distribution $q(\gamma^*|\gamma)$ is altered as well, since the proposal of new subset is no longer symmetrical.

$$p(\gamma^*|\gamma) = \begin{cases} \frac{\phi}{2p_\gamma}, \text{if} |p_\gamma - p_{\gamma^*}| = 1 \\ \frac{1-\phi}{p_\gamma}, \text{if} |p_\gamma - p_{\gamma^*}| = 0 \end{cases} \tag{5.9}$$

5.2.3 Modifications

This section outlines the modifications and improvements made to the hybrid correlation-based search method to improve variable selection and prediction.

5.2.3.1 Variable Selection

Correlation-based search considers only the most correlated variables during the alterations of the subset γ. As explained in the previous section, Section 5.2.2, CBS considers variables whith a very high correlation to another variable in the subset during exclusion of a variable. Nevertheless, the correlation structure within the DNA [23] tends to be 'local'. Typically, every SNP is highly correlated with a small number SNPs in their surroundings and correlation [23, 34] decreases with

distance along the DNA. For example in fine-mapping studies, where many highly correlated variables in a genomic region are examined, only considering the highest correlated variable can be adverse and disregards slightly less correlated variables.
Therefore, a modification to CBS is proposed in order to improve the mixing of highly correlated variables. As in Section 5.2.2 $\mathcal{E}_\gamma$ is the set of predictors currently excluded from the subset and correspondingly, $\mathcal{L}_\gamma$ is the subset of predictors included. During the deletion moves as well as the deletion part of the swap move not only the highest correlated variables are considered, instead other predictors are given a chance to be chosen as well depending on their correlation. Again a predictor $i^{'} \in \mathcal{L}_\gamma$ is chosen randomly. Thereupon a discrete probability distribution is constructed where each bucket represents a predictor in the subset $j \in \mathcal{L}_\gamma \backslash i^{'}$ having a probability as shown in Equation 5.10 for being excluded.

$$p(j) = \frac{\rho_{i'j}}{\sum_{m=1}^{\mathcal{L}_\gamma \backslash i^{'}} \rho_{i'm}} \tag{5.10}$$

The modification leads to a reduced number of false positive detections as examined in Chapter 6.

Interestingly, various other strategies for the inclusion and exclusion steps have been considered, but led to an increased number of false positive detections. For example, considering the highest correlated variable $x_{j'}$ during the inclusion step $j \in \mathcal{E}_\gamma : |\rho_{i'j'}| = max|\rho_{i'j}|\}$ and excluding the least correlated variable $x_{j'}$ satisfying $\{j \in \mathcal{L}_\gamma : |\rho_{i'j'}| = min|\rho_{i'j}|\}$ from the current subset γ, showed inferior results. This was unexpected, since it is reasonable that correlated variables being associated should be kept in the subset and not associated correlated variables should be rejected.
Another strategy examined is to consider not only the least correlated variable during the inclusion steps, but, analogous to the modification in Equation 5.12, assigning other slightly correlated variables a probability to be selected as well as shown in Equation 5.12

$$p(j) = \frac{\rho_{i'j}}{\sum_{m=1}^{\mathcal{E}_\gamma} \rho_{i'm}} \tag{5.11}$$

where $j \in \mathcal{E}_\gamma$ and $i' \in \mathcal{L}_\gamma$ is randomly chosen

However, as mentioned, these strategies led to an increased number of false positive detections and did also not contribute to increase the posterior inclusion probability.

5.2.3.2 Prediction

Prediction of phenotypes based on new or unseen datasets is another purpose of genome-wide association studies. The estimation of the true regression coefficients is the basis for prediction; that is, to understand the true underlying biological influence of SNPs on a certain trait. All regression coefficients are obtained by computing the regression coefficients for the subset γ in each iteration of the Markov Chain. Finally, after the convergence of the Markov Chain [37] the samples are summarized to obtain, for example, the posterior mean, which is then used as the regression coefficient for the selected variables. SSVS [6] estimates the regression coefficients in every iteration using a least-squares estimate; whereas, other SSVS variations do not estimate regression coefficients at all [23].
However, by using ridge regression estimates in each iteration to obtain the regression coefficients for the current subset X_γ more accurate samples are obtained. Ridge regression [29] imposes a penalty term on the squared sum of the magnitude of the regression coefficients as shown in Equation 5.12 and improves the estimation of β in the presence of multicollinearity.

$$\hat{\beta}_\gamma^{Ridge} = arg\min_\beta((Y - X_\gamma\beta)'(\hat{Y} - X_\gamma\beta) + \lambda\sum_{i=1}^{p^*}|\beta_i|^2) \tag{5.12}$$

To obtain the optimal regression coefficients β_γ of the current subset, the penalty term λ has to be evaluated first. Two common methods are examined in terms of the computation demands and their accuracy for estimating λ:
The first approach to obtain the regression coefficients β is evaluated, where λ is obtained for each subset by a direct approach [50]. Therefore, Gaussian priors for β and ϵ are assumed. λ can be obtained from Equation 5.14.

$$\begin{aligned}\lambda_{Direct} &= \frac{p^*\hat{\sigma^2}}{\beta'_{\gamma_{LS}}X'_\gamma X_\gamma\beta_{\gamma_{LS}}} \\ \hat{\sigma^2} &= \frac{||y - X\beta_{\gamma_{LS}}||^2}{n - p^*}\end{aligned} \tag{5.13}$$

The second method examined is termed generalized cross-validation [29]. It provides an approximation to the leave-one out cross-validation, a method to validate different choices of the shrinkage parameter λ. Leave-one out cross-validation [29] uses one dataset to validate the regression coefficients obtained by fitting the model to the remaining datasets of the training set. This is repeated until every dataset has been used once for validation. Generalized-cross validation provides an estimate with which to compare the different values of λ. The optimal amount of shrinkage for the current subset γ can be computed using generalized cross-validation as in

Equation 5.14.

$$GCV(\hat{f}_\gamma) = \frac{1}{p_\gamma} \sum_{i=1}^{p_\gamma} \left[\frac{Y - (X'_\gamma X_\gamma + \lambda)^{-1} X'_\gamma Y}{1 - \frac{trace(S_\gamma)}{p_\gamma}} \right]^2 \tag{5.14}$$

Whereas, the S_γ is defined as in Equation 5.16

$$\hat{y} = S_\gamma y \tag{5.15}$$

$$S_\gamma = X_\gamma (X'_\gamma X_\gamma + \lambda)^{-1} X'_\gamma \tag{5.16}$$

S_γ is also known as the hat matrix [29].
Since the error function is quadratic, the best choice of λ can be evaluated by finding the minimum of the quadratic function. For this purpose the MATLAB function `fminsearch` is used.

Both methods showed very similar results, but the direct approach in Equation 5.14 is computationally more favourable and hence better suited to be computed in every iteration step. An evaluation of the direct method is referred to Chapter 6

5.3 Bayesian Penalized Regression

Penalized regression methods are a common approach [29] and are applicable to a wide range of regression problems as outlined in Section 5.1. Penalized regression methods considered in the work at hand are estimations of the regression coefficients where constraints are imposed on their magnitude and are similar to the least squares estimates, but yield some important properties.
The estimation of the regression coefficients obtained by least squares [29] are unbiased estimators and have the smallest variance of all unbiased estimators. Nevertheless, a severe problem [5, 29] of the least squares approach is the tendency of large magnitudes of the estimated regression coefficients; hence, these estimations are sensitive to the datasets. Another problem arises from datasets [37] where the number of variables is larger than the number of observations, which is common in genome-wide association studies (as outlined in the introduction in Chapter 5), as well as the presence of high correlations among the variables. Nevertheless, biased estimators exist [29] yielding a smaller mean squared error than the unbiased least squares estimation and are suitable for the application to large-scale regression problems, known as regularized regression or penalized regression. Two common methods [29] are **ridge regression** and **lasso**, among others. These two methods are outlined and subsequently their Bayesian equivalents are explained.
Initially, ridge regression is introduced by Hoerl *et.al.* [33] to overcome the difficulties arising from multicollinearity. Ridge regression restricts the sum of the

quadratic magnitudes of the regression coefficients. Therefore, the residual sum of squares (RSS) is minimized subject to the constraint $\sum_{i=1}^{p} |\beta_i|^2 \leq t$, also known as L_2 norm. The constraint, usually denoted as λ, as in Equation 5.18, represents the amount of shrinkage imposed. The constraint on the coefficients [29] alleviates the problem arising from high correlations between the variables, where the regression coefficients can become very large. Hence, ridge regression has the property of improving prediction in the face of multicollinearity; nevertheless [29, 37], it is not able to perform subset selection by effectively setting regression coefficients to 0.

In contrast to ridge regression, Tibshirani [59] introduced a different penalty on the magnitude of the regression coefficient, by exchanging the L_2 - term by the non-differentiable constraint expressed by the L_1 norm $\sum_{i=1}^{p} |\beta_i| \leq t$. Hence the lasso estimator is given by the following equation

$$\hat{\beta}^{Lasso} = arg\min_{\beta}((\hat{Y} - X\beta)'(\hat{Y} - X\beta) + \lambda \sum_{i=1}^{p} |\beta_i|) \tag{5.17}$$

and correspondingly the ridge regression estimator is given by Equation 5.18

$$\hat{\beta}^{Ridge} = arg\min_{\beta}((\hat{Y} - X\beta)'(\hat{Y} - X\beta) + \lambda \sum_{i=1}^{p} |\beta_i|^2) \tag{5.18}$$

The lasso yields the property of performing continuous shrinkage and simultaneous variable selection. Due to the penalty-term in Equation 5.17 the solution to obtain estimates for the regression coefficient does not longer exist in closed form. In the presence of multicollinearity [30, 40] the lasso tends to select one among the correlated variables and shrinks the remaining highly correlated variables towards 0. The lasso has the property that it can only select n variables at maximum in settings of p$\gg$n.

Yuan [68] and Park and Casella [49] state that penalized regression methods have the drawback of not providing valid standard errors and do not provide [56] probabilities to measure level of certainty in the resulting model.

A fully Bayesian treatment of the lasso is introduced by Park and Casella [49] providing interval estimates of all parameters, thereby supporting variable selection. Kyung *et al.* [37] adapts the Bayesian treatment for lasso to fit a more general model of the lasso in order to represent other penalized regression methods such as ridge regression, Fused lasso, Grouped lasso and Elastic Net. A Gibbs sampler is used in all of the models to explore the posterior distributions. The mean of the samples from the posterior distribution [37] is then used as estimate.

Since neither the Bayesian lasso nor the Bayesian ridge regression is able to effectively set the regression coefficients of irrelevant variables exactly to 0 a subsequent variable selection is performed [40] by using the **credible interval criterion**.

A variable is excluded if the credible interval of the regression coefficient β_i covers 0. Consequently, a variable is considered relevant if 0 lies outside of the credible interval. More details are presented in Chapter 6. A common choice is a 95% credible interval [40] and will mostly be used in Chapter 6. Li *et.al.* [39] suggests that a 95% invervall leads to significant selections; however, using a the 95% invervall can lead to many exluded variables. Li *et.al.* [40] suggests that using a 50% interval can lead to better variable selection. More details are again discussed in Chapter 6.

5.3.1 Bayesian Lasso

In the first proposal of lasso, Tibshirani [59] noted that the penalty term in Equation 5.18 could be obtained *as Bayes posterior mode of an independent double exponential prior for the βs.* According to Park and Casella [49] the Bayesian lasso appears to be a compromise between the lasso and ridge regression in terms of the regularization path. The following hierarchical model is adopted from Park and Casella [49] as well as from Kyung *et al.* [37] and defines the hierarchical model. Conveniently, the shrinkage parameter λ is also assigned a hyperprior [37] and thus it is not necessary to estimate the appropriate amount of shrinkage by for example cross-validation or generalized cross-validation.

$$y|\alpha, X, \beta, \sigma^2 \sim \mathcal{N}_n(\alpha 1_n + X\beta, \sigma^2 I_n) \tag{5.19}$$

$$\beta|\sigma^2, \tau_1^2, \tau_2^2, ..., \tau_p^2 \sim \mathcal{N}_p(0_p, \sigma^2 D_\tau) \tag{5.20}$$

$$D_\tau = diag(\tau_1^2, \tau_2^2, ..., \tau_p^2)$$

$$\sigma^2, \tau_1^2, \tau_2^2, ..., \tau_p^2 d\sigma^2 \sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{\frac{-\lambda^2 \tau_j^2}{2}} d\tau^2 \tag{5.21}$$

After integrating out $\tau_1^2, \tau_2^2, ..., \tau_p^2$, β has the desired form of a conditional Laplace prior as suggested by Park and Casella [49]

$$p(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\frac{\lambda|\beta_j}{\sigma}} \tag{5.22}$$

The non-informative scale-invariant marginal prior $p(\sigma^2) = \frac{1}{\sigma^2}$ is used as a prior for σ^2. Shrinkage parameter λ, which is usually estimated, is assigned a hyperprior. Therefore, in the Bayesian treatment a gamma prior on λ^2, as denoted in Equation 5.24 is considered and included in the Gibbs sampler. The full conditional distribution of λ^2 is a Gamma with shape and rate

$$\lambda^2 \sim \text{Gamma}(p + r, \sum_{j=1}^{p} \frac{\tau_j^2}{2 + \delta} \tag{5.23}$$

$$where \quad \lambda^2\, 0, r > 0, \delta > 0$$

The hierarchical model is put into a Gibbs sampler to obtain samples from the posterior distributions.

5.3.2 Bayesian Ridge Regression

The same hierarchical setup as defined in Section 5.3.1 is used to represent the Bayesian ridge regression, as well as some other methods [49] through the modification of the priors on $\tau_1^2, ..., \tau_p^2$ and σ^2. According to Park and Casella [49] the hierarchical Lasso is adapted for Ridge Regression by giving all τ_j^2's a degenerative distribution at the same constant value.

CHAPTER 6

Results

The purpose of this Chapter is to apply the methods outlined in Chapter 5 to simulated and real datasets in order to assess the quality of results obtained in terms of their ability to identify relevant SNPs as well as to predict phenotypes.

For the purpose of evaluation two simulated datasets are considered, forming the basis for comparison and inference of which method is superior. The simulated datasets mimic certain properties of small datasets in genome-wide association studies mainly in terms of their correlation structure. Furthermore, the methods are applied to real datasets. Unfortunately, at the time of writing no real GWAS-dataset was available for evaluation purposes.
This thesis originates in collaboration on the FWF-funded project *Genome wide association study for functional longevity and related traits of dairy cows*[1], however, the data were not granted for use outside the project's scope. Furthermore, due to the strict data protection policies no publically available genome data suitable for this work were found. Most datasets such as the datasets used for the GenABEL tutorial or the demonstration dataset for GEMMA software serve the purpose of demonstrating features of the software and could not be used for this Chapter in a meaningful way.
Thus, to assess the quality of the methods as applied to a real dataset the prostate cancer dataset [55], present in various other works [29, 30, 37, 40], is used. More details are discussed in Section 6.3 and Section 6.4.
Preliminary analysis [2], as addressed for example by Thomas [58] or by Beaumont *et.al.* [4], is not part of this Chapter, since it would greatly exceed the scope of the thesis.
For the remainder of the Chapter the modifications to hybrid correlation-based

[1] https://forschung.boku.ac.at/fis/suchen.projekt_uebersicht?sprache_in=en&menue_id_in=300&id_in=8359

search discussed in Section 5.2.3 are referred to as hCBS* for convenience.

The computational results presented in Section 6.1, 6.2 and 6.4 have been achieved using the computational resources provided by the Vienna Scientific Cluster (VSC).

The Chapter presents the results from the analysis of a block-wise correlated dataset in Section 6.1, a pair-wise correlated dataset in Section 6.2 as well as the application of the methods to a real dataset in Section 6.3 and the same dataset extended to a p$\gg$n dataset in Section 6.4. The Chapter concludes with a discussion of the computational demands in Section 6.5.

6.1 Block-wise correlation

The first dataset used yields a block-wise correlation structure and mimics a block of proximate correlated SNPs influencing the phenotype besides another block of correlated variables having no influence. A similar dataset has been used by Kwon *et.al.* [36]. The dataset contains $p = 5,000$ SNPs (variables) and $n = 500$ phenotypes (observations) to demonstrate a small GWAS dataset.
The phenotypes are generated from a univariate normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. For randomly generating the predictors the following correlation matrix Υ_X is used.

$$\Upsilon_X = \begin{pmatrix} \Upsilon_{11} & \Upsilon_{12} \\ \Upsilon_{21} & \Upsilon_{22} \end{pmatrix} \tag{6.1}$$

In 6.1 Υ_{11} is a $10x10$ matrix, corresponding to the correlation of the predictors associated to the phenotype. Υ_{12} and Υ_{21} denote the correlation between predictors associated with the outcome and the remaining predictors. Consequently Υ_{22} represents a $990x990$ block. For this simulation study $\Upsilon_{11} = 0.85$, $\Upsilon_{12} = \Upsilon_{21} = 0.45$ and $\Upsilon_{22} = 0.55$ is used; furthermore $\beta_i = 0.5$ for $i = 1, ..., 10$ and $\beta_i = 0$ for $i = 11, ..., 5,000$ are the regression coefficients.

To obtain meaningful results 25 datasets are generated using the same structure. Subsequently, hCBS*, hCBS, SSVS, Bayesian lasso, and Bayesian ridge regression are applied to these datasets. All datasets are normalized and standardized $\sum_{i=1}^{n} x_{ij} = 0$, $\sum_{i=1}^{n} y_i = 0$ and $\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, ..., p$.
In the case of hCBS*, hCBS and SSVS hyperparameters need to be specified and are chosen to give a result with as many true positive detections as possible and minimal amount of false positive detections. To make the results comparable, the same hyperparameters are used for hCBS, hCBS* and SSVS: $\omega = \frac{10}{5,000}$, reflecting the prior belief of truly associated variables in the dataset, $v = 3$ and $\lambda = 1$ and $H_\gamma = cI_p$ with $c = 1$ is used. The proportion of hCBS, or hCBS*, to SSVS moves

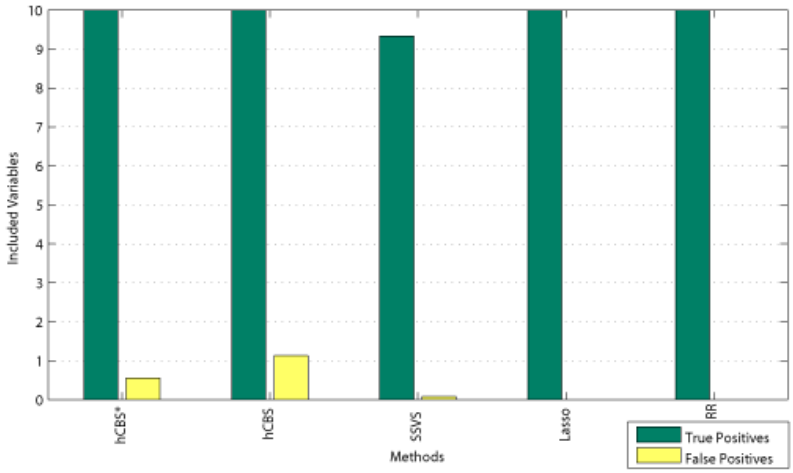

Figure 6.1: True and false positive detections

is set to 0.9, the same as used by Kwon *et.al.* [36]. 0.5 is used as a coin-flip to either use a swap or inclusion/exclusion move.
For SSVS the mixing between hCBS and SSVS is set to 0 to result in only SSVS moves in each iteration.
HCBS, hCBS* and SSVS are run for $1,000,000$ iterations. Of these, $5,000$ iterations are discarded as burn-in period as explained in Section 4.2.
Bayesian lasso and Bayesian ridge regression are run for $15,000$ iterations with $1,000$ iterations are removed from the chain.

Figure 6.1 shows the results using a threshold of 0.5 for the posterior inclusion probability in the case of hCBS, hCBS* and SSVS, as well as for a 95% credible interval in case of Bayesian lasso and Bayesian ridge regression.
All methods except for SSVS are able to identify all relevant variables.
Bayesian lasso and Bayesian ridge regression both perform very well and identify all relevant variables with no false positive detections. hCBS* also performs well by identifying all relevant variables in contrast to only 0.8 false positive detections on average, while hCBS identifies 10 true positive variables and 1.6 non-associated variables. Thus, hCBS* reduces the number of false positive detections in comparison to hCBS. SSVS identifies on average 9.33 out of the 10 associated variables and also yields the lowest number of false positives in comparison to hCBS and hCBS*.
Figure 6.2 shows the regression coefficients obtained by Bayesian lasso along with the 95% confidence intervalls, which are used for variable selection. Figure 6.3 depicts the same obtained by Bayesian ridge regression.

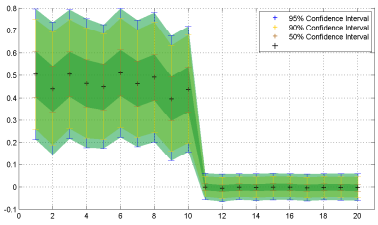

Figure 6.2: Regression coefficients with varying credible interval criterion for variable selection obtained by Bayesian lasso

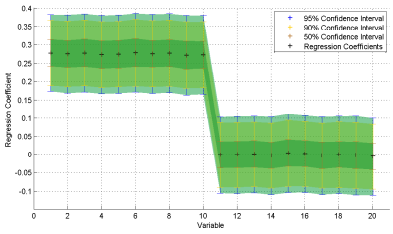

Figure 6.3: Regression coefficients with varying credible interval criterion for variable selection obtained by Bayesian ridge regression

By using the regression coefficients from the selected variables to predict unseen datasets hCBS and hCBS* have a similar mean squared error (MSE) of 1.074 and 1.0753; whereas, SSVS has a MSE of 1.0904. Bayesian lasso and Bayesian ridge regression have a MSE of 1.1534 and 1.1364, respectively; thus producing slightly higher prediction errors.
Relaxing the variable selection criteria to 0.4 in the case of SSVS-based methods and to a 90% credible interval criterion leads to false positive rate of 1.7 and 3.7 for hCBS* and hCBS, respectively. For the remaining methods the number of false positive detections remains unchanged.
If the variable selection criterion for Bayesian penalized regression is again lowered to 50% then 427 and 677 false positives are selected, leading to the conclusion that, in this example, a 50% credible interval criterion is not restrictive enough to be used for variable selection purposes.

As mentioned above the computations were carried out on the Vienna Scientific Cluster using one eight-core node for each method and dataset. Computation of hCBS* and SSVS took on average 160 minutes and hCBS took slightly more than 167 minutes. $15,000$ iterations of Bayesian lasso and Bayesian ridge regression took significantly longer. The former terminated after 17.75 hours and the latter required 22.14 hours on average.

In sum, all methods performed well with Bayesian lasso and Bayesian ridge regression slightly outperforming SSVS-based methods in terms of false positive detections, but requiring a significantly longer computational time. However, hCBS, hCBS* and SSVS predict unseen datasets more accurately as reflected in a lower MSE than the Bayesian penalized regression methods.
A more detailed discussion about the computational demands is referred to Section 6.5.

are

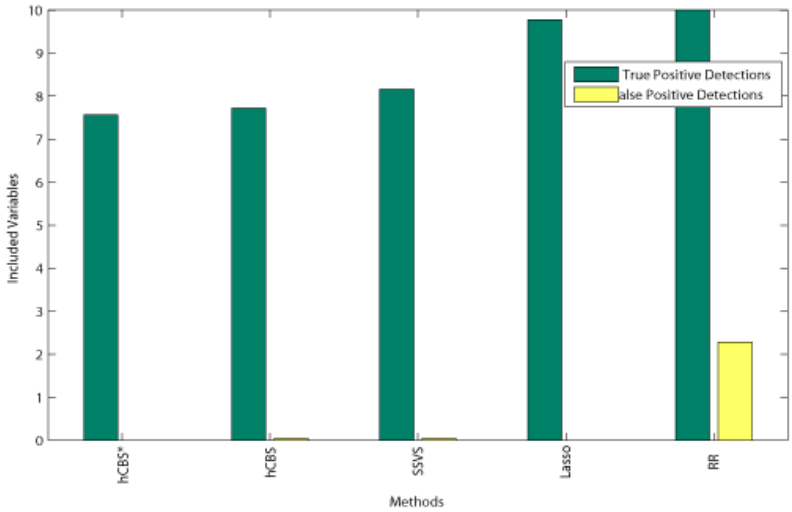

Figure 6.4: True and false positive detections in a pair-wise correlated dataset

6.2 Pair-wise correlation

The second simulated dataset consists of high pair-wise correlation between the variables, mimicking the correlation present in the genome decreasing with distance. Again, $p = 5,000$ SNPs and $n = 500$ phenotypes are used to simulate a small GWAS dataset. The phenotypes are again draws from a univariate normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.
The correlation structure ρ_{ij} is shown in Equation 6.2. Similar simulated datasets are used by Li *et.al.* [40] and Hastie *et.al.* [30].

$$\rho_{ij} = 0.9^{|i-j|} \tag{6.2}$$

HCBS, hCBS* and SSVS are run with $\omega = \frac{10}{5,000}$, specifying the hyperparameter for the number of expected true positive variables, $v = 3$ and $\lambda = 1$ and $H_\gamma = cI_p$ with $c = 0.05$, which can be seen as a penalty term to facilitate inclusion of variables. $1,000,000$ iterations are carried out. Convergence diagnosis indicated that sufficient samples have been collected.
The Bayesian penalized regression methods are run for $15,000$ iterations until convergence is approximately reached.
Again, 25 datasets were generated and computed to obtain average results.

Figure 6.4 summarizes the average number of true and false positive detections. All methods are able to detect a fair amount of true positive variables. Bayesian ridge regression performed best in terms of true positive detections, but also yields

hCBS	1.191
hCBS*	1.233
SSVS	1.139
Lasso	1.1185
RR	1.1011

Table 6.1: Mean squared error predicting unseen datasets

the largest number of false positive detections. Bayesian lasso detects 9.78 true positive variables on average; hCBS*, hCBS and SSVS give similar results, but only hCBS* detects no false positive variables.
SSVS performed better than hCBS and hCBS* in terms of true positive detections, which is a rather unexpected result, since the correlation in the dataset is up to 0.9. HCBS* and Bayesian lasso are the only methods not including any false positives. Figure 6.5 and 6.6 show the regression coefficients as well as the 95%, the 90% and the 50%-confidence interval which is used as credible interval for variable selection.

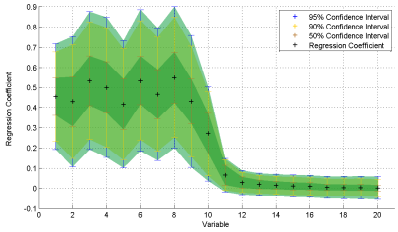

Figure 6.5: Regression coefficients and confidence intervals obtained by Bayesian lasso

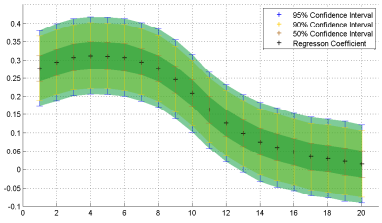

Figure 6.6: Regression coefficients and confidence intervals obtained by Bayesian ridge regression

Table 6.1 shows the average prediction errors over the remaining 24 datasets obtained by the different methods. As shown in the table, all methods perform rather well with a similar prediction error. Obviously, the selected variables by each method influence the MSE.

Computational demands are similar as in Section 6.1. HCBS, hCBS* and SSVS required less than 167 minutes, Bayesian lasso terminated after on average 20.8 hours and Bayesian ridge regression took on average 22.1 hours.

Summarizing the findings, both Bayesian penalized regression methods perform better than the SSVS-based methods in terms of true positive detections. However, Bayesian ridge regression yields the most false positive detections; whereas, Bayesian lasso include no false positives in contrast to 9.76 true positives. HCBS,

hCBS* and SSVS identify a similar number of true positives, but only hCBS* includes no false positives.

6.3 Prostate cancer dataset

As previously mentioned, the prostate cancer dataset comes from a study performed by Stamey *et.al.* [55] which examines the associations between the level of a prostate-specific antigen `lpsa` and eight different clinical measures in men prior to a radical prostatectomy[2]. The dataset contains measurements from 97 men.
The eight different measurements are:

- `lcavol`: logarithmic cancer volume
- `lweight`: logarithmic prostate weight
- `age`: age of the patient
- `lbph`: logarithmic amount of benign prostatic hyperplasia
- `svi`: seminal vesicle invasion
- `lcp`: logarithmic capsular penetration
- `gleason`: Gleason score
- `pgg45`: percentage of Gleason scores 4 or 5

As done by Hastie *et.al.* [29] the dataset is randomly split into a training set of size 67 and a validation set of size 30 to assess the mean squared prediction error and to compare variable selection performed by the different methods. The dataset is normalized $\sum_{i=1}^{n} x_{ij} = 0$, $\sum_{i=1}^{n} y_i = 0$ and $\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, ..., p$ to remove any effects arising from different scales.
In the case of the hCBS, hCBS* and SSVS $\omega = \frac{1}{2}$, the prior expectation of the number of relevant variables, is chosen since it is unknown and 0.5 is an impartial choice. Other hyperparameters such as $v = 3$ and $\lambda = 1$, as well as $H_\gamma = cI_p$ with $c = 1$ as an independent prior, are chosen to ensure a proper acceptance ratio of the Metropolis/Metropolis-Hastings algorithm.
$35,000$ iterations are carried out and in the case of the penalized regression methods $10,000$ iterations are carried out.

The computation took around 46 seconds for the hCBS methods and slightly less than 60 seconds for the hCBS* method. The increased computation time stems from the calculation of the shrinkage coefficient λ as outlined in Section 5.2.3.

[2] A radical prostatectomy is a surgery for removing all parts of the prostate gland.

Methods	Selected variables	MSE	MPIP
hCBS	`lcavol`, `lweight`, `lbph`	0.7472	0.7916
hCBS*	`lcavol`, `lweight`, `lbph`	0.7477	0.7814
SSVS	`lvacol`, `lweight`	0.6952	0.8867
Lasso	`lcavol`, `lweight`	0.7476	-
RR	`lcavol`, `lweight`	0.7765	-

Table 6.2: Selected variables by the different methods

Bayesian lasso took less than 9 seconds, whereas Bayesian ridge regression took less that 6 seconds[3].
HCBS has a mean acceptance rate of 0.41, SSVS has 0.19; whereas the acceptance ratio of hCBS* is 0.38, all indicating a well-mixing chain, although the acceptance ratio of SSVS is at the lower end. The samples are checked for convergence after the computation. The scale reduction factor is below 1.02 and the Geweke tests passed for all variables; hence, strong evidence in favor of convergence is obtained.

Table 6.2 shows the variables selected by the different methods. Both hCBS and hCBS* selected the three out of eight variables with a posterior inclusion probability higher than 0.5 which is a reasonable choice [23] to use as a threshold. SSVS selected two out of the eight variables namely `lcavol` and `lweight`.
The penalized regression methods selected two out of eight using a 95% credible interval [40] as explained in Section 5.3. When applying a 90% credible interval the same two variables are selected again.
`lcavol` and `lweight` are identified as strongly influencing the response `lpsa` by Hastie *et.al.* [29]. All methods applied to the dataset by Li *et.al.* [40] and the methods performing variable selection in Hastie *et.al.* [30] identified `lcavol`, `lweight` and `lbph`, besides others depending on the method. The same three variables are also identified by hCBS and hCBS*.
If the threshold for the posterior inclusion probability is set to 0.4 hCBS* additionally includes `gleason` as well, which also has been identified as significant by three of the four methods by Li *et.al.* [40].

To assess the MSE for each method the coefficients of the variables selected are applied to the test dataset and are shown in Table 6.2. Interestingly, the MSEs of hCBS, hCBS* and Bayesian lasso are very similar; whereas the MSE of SSVS is lower and the MSE of ridge regression is slightly higher than that of other methods.
The last column shows the mean posterior inclusion probability (MPIP) for hCBS,

[3]The computation was carried out in MATLAB on an ASUS N61Jv notebook deploying an Intel i5 M450 Quad-Core and 6GB RAM running Windows 7. Computation time was averaged over 10 computations for each method.

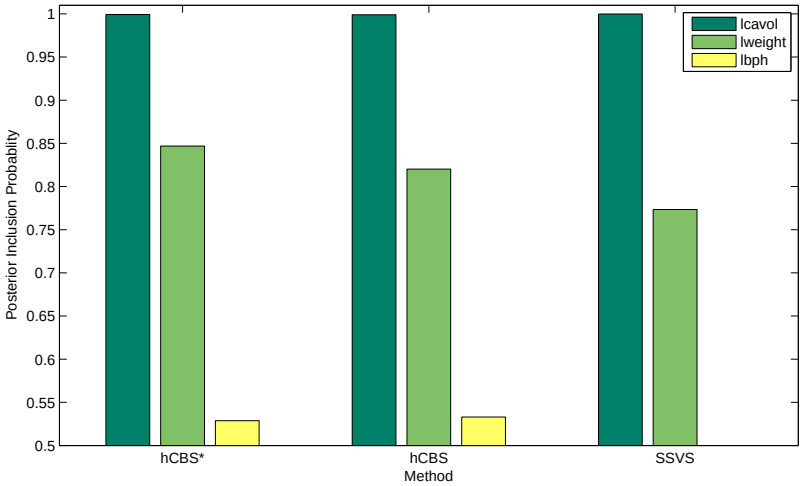

Figure 6.7: Posterior inclusion probability of the selected variables

hCBS* and SSVS and depicts the ratio of the times a variable is included in the subset to the total number of iterations. The inclusion probability of the selected variables is shown in Figure 6.7.

6.4 Extended prostate cancer dataset

Due to the absence of an available GWAS datasets at the time of writing the prostate cancer dataset is transformed into a p$\gg$n dataset to assess the quality of results on a real dataset. The same dataset as in Section 6.3 is used; additionally, white-noise variables are added to bring the total number to 200 variables with 67 measurements in the training set. The additional predictors are highly correlated having an average correlation coefficient of 0.95.
The same approach is used by Hans [25]. This yields the advantage of comparability to the original dataset.
Again the training set has a size of 67 and the remaining 30 measurements are used to assess the quality of prediction and variable selection.
For hCBS and hCBS* the expected ratio of associated variables is set to $\omega = 0.02$, to maintain the ratio used in the example in Section 6.3. The remaining hyperparameters are set to $v = 3$ and $\lambda = 1$ as well as $H_\gamma = cI_p$ with $c = 0.5$ as an independent prior, again, to ensure a proper acceptance ratio. HCBS and hCBS* are run for $250,000$ iterations with $5,000$ burn-in samples. In the case of Bayesian lasso and Bayesian ridge regression $20,000$ iterations are carried out where the first $1,000$ samples are removed as a burn-in period. All results are checked for convergence and the maximum $\hat{R}$-score was 1.0022 for variable `gleason` in hCBS.

Methods	Selected variables	MSE	MPIP
hCBS	`lcavol`, `lweight`, `age`, `lbph`, `gleason`, 53	0.73	0.68
hCBS*	`lcavol`, `lweight`, `lbph`, `svi`, `lcp`, `pgg45`, 37	0.69	0.73
SSVS	`lcavol`	0.73	0.99
Lasso	`lcavol`	0.85	-
RR	`lcavol`	0.94	-

Table 6.3: Selected variables by the different methods

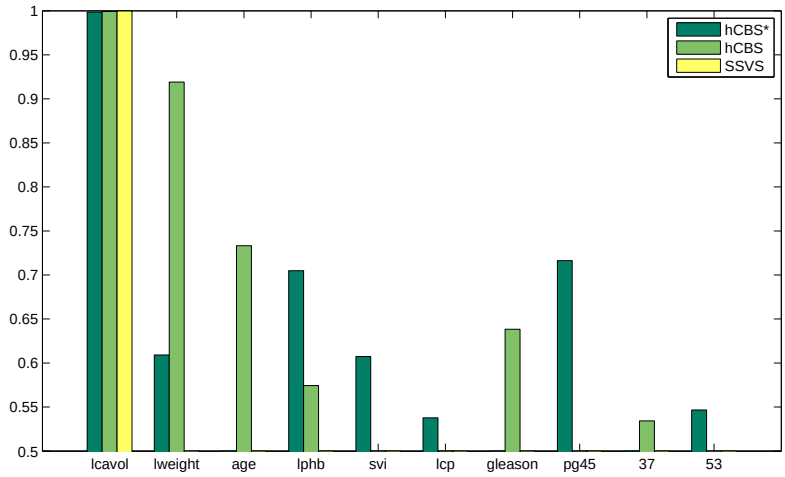

Figure 6.8: Posterior inclusion probability of the selected variables

Table 6.3 shows the variables selected by the different methods, which are above 0.5 for the posterior inclusion probability in the case of hCBS, hCBS* and SSVS and selected by a 95% credible interval criterion in the case of Bayesian lasso and Bayesian ridge regression. HCBS selects 5 out of the 8 real variables and identifies the same variables as in the original dataset (`lcavol`, `lweight`, `lbph`) along with `age` and `gleason`. `age` is only identified by Bayesian lasso in Li *et.al.* [40] and is also not considered relevant by hCBS*, Bayesian lasso and Bayesian ridge regression. `gleason`, is only identified by the methods applied by Li *et.al.* [40] and considered as insignificant by Hastie *et.al.* [29] and Haste *et.al.* [30].
HCBS* includes, besides the same variables included in Section 6.3, `svi`, `lcp` and `pgg45`; one more variable than hCBS. `svi` is identified as relevant by Kyung *et.al.* [37], Hans [26] and Hastie *et.al.* [29] and all methods except for Elastic Net in [30]. `lcp` and `pgg45` are identified by Hastie *et.al.* [30]. HCBS and hCBS*

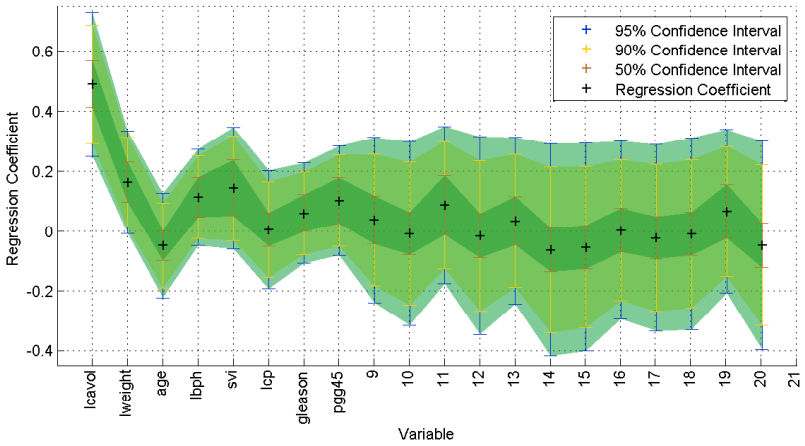

Figure 6.9: Regression coefficients and confidence intervals obtained by Bayesian lasso

additionally include one white-noise variable.
Figure 6.8 shows the posterior inclusion probabilities of the selected variables, which represents the confidence of the variables selected.
Bayesian lasso and Bayesian ridge regression only identify `lcavol` as relevant and consider all other variables as insignificant. The selection of `lcavol` agrees with previous studies using various methods, as `lcavol` is always identified as relevant. To depict the variable selection in Bayesian lasso Figure 6.9 shows the regression coefficients of the first 20 variables along with the 95%, 90% and 50% confidence intervals used for variable selection. The regression coefficients with the confidence intervals for Bayesian ridge regression is depicted in Figure 6.10.
The mean squared errors using the selected variables are also shown in Table 6.3 using the training dataset for evaluating the prediction error. HCBS, hCBS* and SSVS show similar MSEs; whereas Bayesian lasso and Bayesian ridge regression show slightly larger MSEs. HCBS* yields the lowest prediction error.

Table 6.4 shows the results when the variable selection criteria for Bayesian penalized regression methods are relaxed to a 90% credible interval criterion. Both methods include `lweight` if a 90%-confidence interval is used and additionally include `lbph`, `svi`, `gleason` , `pgg45` and three white-noise variables if a 50%-confidence interval is applied as suggested by Li *et.al.* [40]. The difference between the 95%, the 90% and the 50%-credible interval criterion is shown in Figure 6.9 for Bayesian lasso and for Bayesian ridge regression in Figure 6.10. If the variable selection criteria are relaxed to 0.4 for hCBS, hCBS* and SSVS no other variables are additionally included.

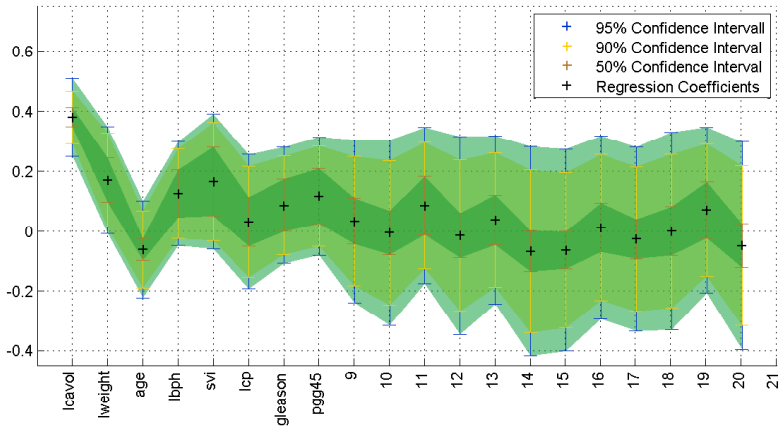

Figure 6.10: Regression coefficients and confidence intervals obtained by Bayesian ridge regression

Methods	Selected variables	MSE
$\text{Lasso}_{90\%}$	`lcavol`, `lweight`	0.7994
$\text{Lasso}_{50\%}$	`lcavol`, `lweight`, `lbph`, `svi`, `gleason`, `pgg45`, 132, 142,168	0.6686
$\text{RR}_{90\%}$	`lcavol`, `lweight`	0.8865
$\text{RR}_{50\%}$	`lcavol`, `lweight`, `lbph`, `svi`, `gleason`, `pgg45`, 132, 142,168	0.7069

Table 6.4: Selected variables by the different methods

The computation was carried out on the VSC on an 8-core node for each method. hCBS took 169 seconds and hCBS* 143 seconds to complete the $250,000$ iterations; whereas Bayesian lasso took 84 seconds and Bayesian ridge regression 42 seconds both computing $20,000$ iterations.

6.5 Computational Analysis

The purpose of this Section is to assess the computation time for all the methods using datasets of various sizes, each having ten variables associated to the outcome.

A direct comparison is rather difficult, since the main feature of the SSVS-based methods is to perform variable selection during computation so that only a subset of the variables $p_\gamma \leq p$ is used in every iteration; whereas, Bayesian lasso

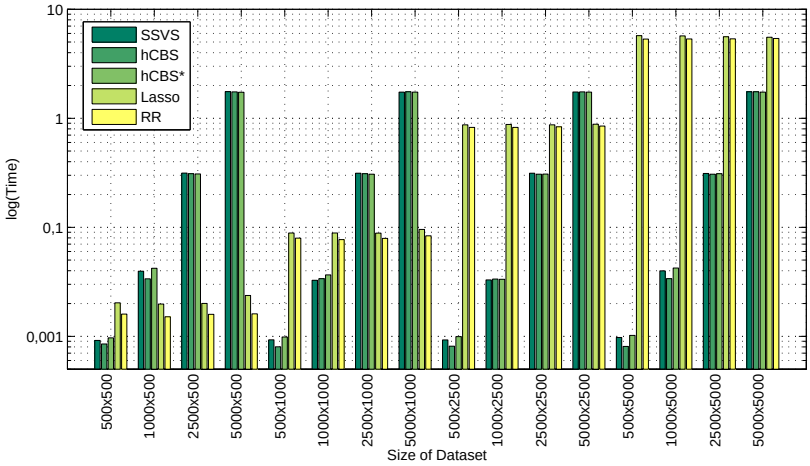

Figure 6.11: Required time for analyzing dataset of various sizes

and Bayesian ridge regression compute all variables in every iteration and perform variable selection subsequent to the computation. As a consequence the performance of SSVS-based methods depends on the number of relevant variables. Since in typical GWAS [23] only a few SNPs influence the phenotype, all datasets have ten associated variables and consequently the hyperparameter ω for SSVS, hCBS, and hCBS* is set to $\frac{10}{\#variables}$.
Moreover, since samples in the Metropolis-Hastings-algorithm are generated and then either accepted or rejected, identical samples are included in chain, leading to an increased autocorrelation. As a consequence, more samples are needed to give meaningful results.
For those reasons the comparison is based on the time required for one iteration and is averaged over 500 iterations and 5 repeats per method.

From the results shown in Figure 6.11, it can be seen, that the computation time of SSVS-based methods scale with increases in the number of phenotypes or more generally with an increase in the number of observations. Note that the y-scale is logarithmic. In contrast Bayesian penalized regression methods scale with the number of SNPs or in general with the number of variables. This stems from the fact, that in every iteration each variable has to be sampled separately which is the most time consuming calculation.
As the number of SNPs increases in a GWA study Bayesian penalized regression methods will require excessive computational time and resources.
Bayesian penalized regression methods require on average 5.5 seconds per iteration for a dataset of size 5,000x5,000. In contrast SSVS-based methods take 1.7

seconds. If the data set consists of 500 phenotypes instead of 5,000, then Bayesian penalized regression methods still require around 5.5 seconds; whereas, computation time of SSVS-based methods decreases to 0.004 seconds.

Since SSVS-based methods are able to compute datasets in less time than Bayesian lasso and Bayesian ridge regression they are computationally better suited for the application in large GWA studies.

CHAPTER 7

Discussion

7.1 Discussion

This thesis discusses modifications to the hybrid correlation-based search method and gives a comparison to Bayesian penalized regression methods such as lasso and ridge regression both on real and simulated datasets. The simulated datasets are partly highly correlated to mimic the structure of real GWAS datasets.

Results show that the modification to the variable selection procedure of correlation-based search explained in Section 5.2.3 leads to a reduced number false positive detections in both datasets. In the example computed in Section 6.1 hCBS* also identifies all relevant variables. Although, the number of true positive detections in the simulated dataset in Section 6.2 is slightly lower than by hCBS, setting the hyperparameter H_γ to a lower value results in an improved true/false positive ratio. On the contrary, using ridge regression, as introduced in Section 5.2.3, to estimate the regression coefficients in every iteration does not lead to improved prediction results, which is reflected in a slightly higher MSE when the same variables are selected, as discussed in the previous Chapter.

Bayesian lasso and Bayesian ridge regression perform well in terms of variable selection on both simulated datasets, identifying all relevant variables. Bayesian lasso detects no false positives in either dataset, while Bayesian ridge regression includes no false positives in the first dataset as discussed in Section 6.1, but a rather high number of non-associated variables in the pair-wise correlated dataset shown in Section 6.2. Note that the regression coefficients in both simulated datasets are set to the same value of 0.5; thus, each associated variable has the same influence. Bayesian lasso and Bayesian ridge regression also perform comparably to SSVS-based methods all preforming well when predictions are made for unseen datasets.

In the real datasets in Section 6.3 and Section 6.4, where the variables have different influences, Bayesian penalized regression methods both identify only the variables having the strongest influence. In contrast, hCBS* identifies the most variables previously found by other studies. SSVS performed inferior than hCBS and hCBS* in both real dataset examples.

A difficulty arising from the use of hCBS, hCBS* or SSVS is the specification of the hyperparameters since the choice influences variable selection. Although the mixing of the Markov chain can be regarded as guideline for hyperparemter specification, guesses for the optimal values are mostly vague. However, if the hyperparameters are set to arbitrary values SSVS-based methods are still able to detect a fair amount of relevant variables. For example, when the dataset used in Section 6.2 is computed using different hyperparameters such as $v = 1$ and $\lambda = 20$, resulting in a broad distribution for the residual error term in Equation 5.1, and $c = 1$ then hCBS identifies 5.7, hCBS* identifies 5.8 and SSVS identifies 6.2 associated variables. All methods detect less true positives, but are still able to at least identify a fair amount.

Computing a 500x$5,000$ dataset using either Bayesian lasso or Bayesian ridge regression requires long computation times as discussed in Section 6.5, which leads to excessive computational demands when real GWAS datasets are computed with ten to hundreds of thousand SNPs to be analyzed.
Computational times of Bayesian penalized regression methods scale with the number of SNPs; whereas, computational time of SSVS-based methods mostly scale with the number of phenotypes. Computing $1,000,000$ iterations using hCBS, hCBS* and SSVS takes noticeably less time than computing $15,000$ iterations of Bayesian penalized regression methods when the dataset is p$\gg$n. Note that only ten of the variables influence the outcome and the average number of variables in the subset of SSVS-based methods varied from ten to thirty. If there are more variables included in the subset is higher or if there are more genomes in the dataset the computation time of SSVS-based methods increases significantly.
However, since it is mostly considered that only a small number of SNPs influence a phenotype and the number of genomes is rather low compared to the number of SNPs, hCBS, hCBS* and SSVS are better suited for the application with very large datasets as is often the case with GWAS.

All methods considered in this thesis are able to perform variable selection with a reasonable amount of true positive detections and a low number of false positive detections. SSVS-based methods outperform Bayesian penalized regression methods in terms of computability of large p$\gg$n datasets, still resulting in useful results and Bayesian penalized regression are superior to hCBS, hCBS* and SSVS

in terms of more true positive detections. All methods are able to make reasonably accurate predictions using the selected variables.

A feasible way to tackle the computational challenges arising in GWAS, which is left to be addressed in future work, would be to consider a two-step strategy where the initial selection of SNPs is performed by hCBS* using hyperparameters that are not to restrictive to variable inclusion (especially setting the hyperparameter for H_γ to a low value since H_γ basically regulates the penalty of variable inclusions). A second step would involve computing the reduced set of SNPs using either Bayesian lasso or Bayesian ridge regression. A similar approach is used by Li *et.al.* [39] first reducing the initial set of SNPs by applying a supervised principle component analysis and subsequently computing the remaining SNPs using Bayesian lasso. A related approach, to the methods considered in this work, is proposed by Hans [25] where the variable selection ability of SSVS is combined with Bayesian lasso to compute each subset.

Bibliography

[1] Christophe Andrieu, N De Freitas, and A Doucet. An introduction to MCMC for machine learning. *Science*, 50(1):5–43, 2003.

[2] David J Balding. A tutorial on statistical methods for population association studies. *Nature reviews. Genetics*, 7(10):781–91, October 2006.

[3] M. Baragatti and D. Pommeret. A study of variable selection using g-prior distribution with ridge parameter. *Computational Statistics & Data Analysis*, 56(6):1920–1934, June 2012.

[4] Mark a Beaumont and Bruce Rannala. The Bayesian revolution in genetics. *Nature reviews. Genetics*, 5(4):251–61, April 2004.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science + Business Media, LLC, 2006.

[6] P. J. Brown, M. Vannucci, and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):519–536, August 2002.

[7] PJ Brown and M Vannucci. Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics*, 182(December 1997):173–182, 1998.

[8] PJ Brown and M. Vannucci. Multivariate Bayesian variable selection and prediction. *Journal of the Royal*, 60(3):627–641, August 1998.

[9] Xiaodong Cai, Anhui Huang, and Shizhong Xu. Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. *BMC bioinformatics*, 12(1):211, January 2011.

[10] Bradley P. Carlin. *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications*. Oxford University Press, Inc, 2006.

[11] Carla Chia-Ming Chen, Holger Schwender, Jonathan Keith, Robin Nunkesser, Kerrie Mengersen, and Paula Macrossan. Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 8(6):1580–91, 2011.

[12] Hugh Chipman. Bayesian variable selection with related predictors *. *Statistics*, 24(1):17–36, 1996.

[13] MA Cleveland, Selma Forni, Nader Deeb, and Christian Maltecca. Genomic breeding value prediction using three Bayesian methods and application to reduced density marker panels. *BMC proceedings*, 4(Suppl 1):1–7, 2010.

[14] Gustavo de los Campos, Daniel Gianola, and David B Allison. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature reviews. Genetics*, 11(12):880–6, December 2010.

[15] B Devlin and N Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–22, September 1995.

[16] Persi Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, November 2008.

[17] Peter Donnelly. Progress and challenges in genome-wide association studies in humans. *Nature*, 456(7223):728–31, December 2008.

[18] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie a. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Pouya Kheradpour, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C. J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura a. L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Peter J. Good, Elise a. Feingold, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Morgan C. Giddings, Thomas R. Gingeras, Roderic

Guigó, Timothy J. Hubbard, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, John a. Stamatoyannopoulos, Scott a. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Yanbao Yu, John Wrobel, Brian a. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Matthew L. Eaton, Alex Dobin, Andrea Tanzer, Julien Lagarde, Wei Lin, Chenghai Xue, Brian a. Williams, Chris Zaleski, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, Yoshihide Hayashizaki, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Yijun Ruan, Piero Carninci, Cricket a. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Linda L. Grasfeder, Paul G. Giresi, Anna Battenhouse, Nathan C. Sheffield, Kimberly a. Showers, Darin London, Akshay a. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Z. Zhang, Piotr a. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Vishwanath R. Iyer, Meizhen Zheng, Ping Wang, Jason Gertz, Jost Vielmetter, E. Christopher Partridge, Katherine E. Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M. Bowling, Michael Anaya, Marie K. Cross, Michael a. Muratet, Kimberly M. Newberry, Kenneth McCue, Amy S. Nesmith, Katherine I. Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Stephanie L. Parker, Sreeram Balasubramanian, Nicholas S. Davis, Sarah K. Meadows, Tracy Eggleston, J. Scott Newberry, Shawn E. Levy, Devin M. Absher, Wing H. Wong, Matthew J. Blow, Axel Visel, Len a. Pennachio, Hanna M. Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Giovanni Bussotti, Claire Davidson, Glo-

ria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David a. Hendrix, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Jing Leng, Michael F. Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Electra Tapanari, Michael L. Tress, Marijke J. van Baren, Stefan Washietl, Laurens Wilming, Amonida Zadissa, Zhengdong Zhang, Michael Brent, David Haussler, Alfonso Valencia, Nick Addleman, Roger P. Alexander, Raymond K. Auerbach, Suganthi Balasubramanian, Keith Bettinger, Nitin Bhardwaj, Alan. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[19] James M Flegal, Murali Haran, and Galin L. Jones. Markov Chain Monte Carlo: Can We Trust the Third Significant Figure? *Statistical Science*, 23(2):250–260, May 2008.

[20] Brooke L Fridley. Bayesian variable and model selection methods for genetic association studies. *Genetic epidemiology*, 33(1):27–37, January 2009.

[21] Edward I George and Robert E. McCulloch. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881, September 1993.

[22] E.I. George and R.E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–374, 1997.

[23] Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815, September 2011.

[24] C. Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, September 2009.

[25] Chris Hans. Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing*, 20(2):221–229, November 2009.

[26] Chris Hans. Elastic Net Regression Modeling With the Orthant Normal Prior. *Journal of the American Statistical Association*, 106(496):1383–1393, December 2011.

[27] R B O Hara and M J Sillanp. A Review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Analysis*, 4(1):85–118, 2009.

[28] BL Harris and FE Creagh. Experiences with the Illumina high density bovine beadchip. *Interbull Bulletin*, (44):3–7, 2011.

[29] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2009.

[30] Trevor Hastie and Hui Zou. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67:301–320, 2005.

[31] WK K Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[32] LA Hindorff, J ((European Bioinformatics Institute) MacArthur, A Wise, HA Jnkins, PN Hall, AK Klemm, and TA Manolio. A Catalog of Published Genome-Wide Association Studies, 2012.

[33] AE Hoerl. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[34] Rolf Knippers. *Molekulare Genetik*. Georg Thieme Verlag, Stuttgart, Stuttgard, 9. edition edition, 2006.

[35] Leonid Kruglyak. The road to genome-wide association studies. *Nature reviews. Genetics*, 9(4):314–318, April 2008.

[36] Deukwoo Kwon, Maria Teresa Landi, Marina Vannucci, Haleem J Issaq, Darue Prieto, and Ruth M Pfeiffer. An Efficient Stochastic Search for Bayesian Variable Selection with High-Dimensional Correlated Predictors. *Computational statistics & data analysis*, 55(10):2807–2818, October 2011.

[37] Minjung Kyung, Jeff Gill, Malay Ghosh, and George Casella. Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis*, 5(2):369–412, 2010.

[38] AL Lehninger, DL Nelson, and MM Cox. *Lehninger principles of biochemistry*. W. H Freeman, New York, 5 (septemb edition, 2005.

[39] Jiahan Li, Kiranmoy Das, Guifang Fu, Runze Li, and Rongling Wu. The Bayesian lasso for genome-wide association studies. *Bioinformatics (Oxford, England)*, 27(4):516–23, February 2011.

[40] Qing Li and Nan Lin. The Bayesian Elastic Net. *Bayesian Analysis*, 5(1):151–170, 2010.

[41] Yulan Liang and Arpad Kelemen. Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. *Statistics Surveys*, 2:43–60, 2008.

[42] M Mangino. Fundamentals of genome-wide association studies. *Heart and Metabolism*, 55:33–37, 2012.

[43] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics*, 11(7):499–511, July 2010.

[44] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John P a Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics*, 9(5):356–69, May 2008.

[45] N. Metropolis, A.W. Rosenbluth, M. N. Rosenbluth, Teller. A.H., and E. Teller. Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[46] Michael L Metzker. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46, January 2010.

[47] Theo H E Meuwissen and Mike E Goddard. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics, selection, evolution : GSE*, 36(3):261–79, 2004.

[48] Jason H Moore, Folkert W Asselbergs, and Scott M Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics (Oxford, England)*, 26(4):445–55, February 2010.

[49] Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, June 2008.

[50] Yngve Selén, Richard Abrahamsson, and Peter Stoica. Automatic robust adaptive beamforming via ridge regression. *Signal Processing*, 88(1):33–49, January 2008.

[51] Fabyano Fonseca Silva, Luis Varona, Marcos Deon V. de Resende, Júlio Sílvio S. Bueno Filho, Guilherme J.M. Rosa, and José Marcelo Soriano Viana. A note on accuracy of Bayesian LASSO regression in GWS. *Livestock Science*, 142(1-3):310–314, December 2011.

[52] Axel Skarman, Mohammad Shariati, Luc Jans, Li Jiang, and Peter Sø rensen. A Bayesian variable selection procedure to rank overlapping gene sets. *BMC bioinformatics*, 13:73, January 2012.

[53] Chris C a Spencer, Zhan Su, Peter Donnelly, and Jonathan Marchini. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS genetics*, 5(5):e1000477, May 2009.

[54] Sudeep Srivastava and Liang Chen. Comparison between the stochastic search variable selection and the least absolute shrinkage and selection operator for genome-wide association studies of rheumatoid arthritis. *BMC Proceedings*, 3(Suppl 7):S21, 2009.

[55] T A Stamey, J N Kabalin, J E McNeal, I M Johnstone, F Freiha, E A Redwine, and N Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *The Journal of urology*, 141(5):1076–83, May 1989.

[56] Matthew Stephens and David J Balding. Bayesian statistical methods for genetic association studies. *Nature reviews. Genetics*, 10(10):681–90, October 2009.

[57] Daniel O Stram. Tag SNP selection for association studies. *Genetic epidemiology*, 27(4):365–74, December 2004.

[58] Ducan Thomas. *Statistical methods in genetic epidemiology*. Oxford University Press, Inc, New York, 1st edition, 2004.

[59] R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (*, 58(1):267–288, 1996.

[60] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, February 2005.

[61] Eric J Topol, Sarah S Murray, and Kelly a Frazer. The genomics gold rush. *JAMA : the journal of the American Medical Association*, 298(2):218–21, July 2007.

[62] B Walsh. Markov Chain Monte Carlo and Gibbs Sampling, 2004.

[63] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics (Oxford, England)*, 25(6):714–21, March 2009.

[64] Ai-Jun Yang and Xin-Yuan Song. Bayesian variable selection for disease classification using gene expression data. *Bioinformatics (Oxford, England)*, 26(2):215–22, January 2010.

[65] Nengjun Yi. A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics*, 167(2):967–75, June 2004.

[66] Nengjun Yi, Varghese George, and David B. Allison. Stochastic Search Variable Selection for Identifying Multiple Quantitative Trait Loci. *Genetics*, 164(3):1129–1138, July 2003.

[67] Nengjun Yi and Shizhong Xu. Bayesian LASSO for quantitative trait loci mapping. *Genetics*, 179(2):1045–55, June 2008.

[68] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, February 2006.

[69] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–8, January 2012.

List of Figures