

This dissertation is submitted for the degree of Doctor of Philosophy

Genetic association of high-dimensional traits

Hannah Verena Meyer

September 2017

University of Cambridge
Jesus College
European Bioinformatics Institute

The source code of the thesis is available at https://github.com/HannahVMeyer/thesis.

Acknowledgements

I am immensely thankful to all the people who have supported me throughout my PhD.

First and foremost, a big thank you to my supervisor Ewan Birney who gave me the opportunity to work in his research group. Ewan's enthusiasm, ideas and support were invaluable to guide me through the past four years. I would also like to thank the other members –current and past– of the Birney research group: Ian Dunham, Sander Timmer, Sandro Morganella, Valentina Iotchkova (thank you for all your advice on the statistics), Helena Kilpinen, Leland Taylor, Nils Kölling, Tom Fitzgerald, Carl Barton, Anat Melamed and the other Hannah (Currant). I really valued the helpful discussions and feedback and enjoyed our times in the Northumbrian wilderness. A special thanks to Tracey Andrews, Stacy Knoop, Debbie Howe and Christina Karikidis for their help with all the administrative tasks and finding time in Ewan's busy schedule to meet with me.

My PhD project would not have been possible without my collaboration partners. Paolo Casale and Oliver Stegle supported me with their experience and constructive advice in the method development part of this thesis. Konrad Rudolph helped with the development of the R package. The heart project was based on a close collaboration with Stuart Cook's research group in London, in particular with Antonio De Marvao and Declan O'Regan. My thesis advisory committee with John Marioni, Carl Anderson and Jan Korbel was very helpful with their constructive criticism and suggestions, keeping me and Ewan on track and helped with the timely submission of this thesis. I would also like to thank Sylvia Richardson and Nicholas Timpson for their time examining my thesis and for the interesting discussions during my viva.

Many thanks to the people who helped with proofreading of this thesis: the two Nils (Eling and Kölling), Jack Monahan, Hannah Currant, Carl Barton, Paolo Casale, Ewan Birney and David Pattinson (finding the biggest typo in a chapter heading on the day of submission).

Aside from the academic support, there are many people who deserve a special thank you for their moral support and friendship over the past years. My experi-

ence would not have been the same without the countless hours of early mornings on the Cam with Jesus College Boat Club, the exciting volleyball matches with the Cambridge University Volleyball Club and the great people I met in both teams. I would like to thank my fellow EMBL and Cambridge PhD students for the fun times during the predocs course, the many Cambridge formals and warm welcomes when coming back to Heidelberg, especially Nils, Jack, Hannah, Michael, Konrad, Silvia, Christina, Julia, Laura, Joran, Kostek, Chris, Paul and my (former) housemates Sarah, Maria, Nils and Dani. To my close friends from afar, Micha, Melanie, Jana, Lissy, Geli, Léonie and Meike (not quite so far): your support and encouragement, at whatever time and place was wunderbar. I would like to thank Dave, for the discussions and input on my research, the outdoors activities to distract from them, but mostly, for everything else.

Finally, I would like to thank my family who supported me in all my decisions, encouraged me in difficult times and shared my happiness in good ones: vielen Dank!

Declaration of Originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

This dissertation does not exceed the limit of 60,000 words as specified by the Biology Degree Committee.

Contents

List of Figures

=

List of Tables

List of Abbreviations

BFGS algorithm	Broyden–Fletcher–Goldfarb–Shanno algorithm. 89, 91
CCA	canonical correlation analysis. 52, 53
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry. 71, 72
CV	coefficient of variation. 159
DRR	dimensionality reduction via regression. 125, 126, 129, 131, 132, 135–137, 140, 142, 145, 148, 164, 166, 167, 193
FD	fractal dimension. 179–181, 183–185, 188, 189, 210
FDR	false discovery rate. 42, 44, 116, 117
FIN	Finnish in Finland. 71, 72
FWER	family-wise error rate. 42, 44
GBR	British in England and Scotland. 71, 72
GWAS	genome-wide association study. 35–37, 42, 43, 45, 47, 50, 53, 54, 66–68, 80–82, 85, 101, 102, 116–118, 153–155, 161, 166, 169, 172, 175, 176, 178, 183, 185, 187, 188, 191–193, 198
IBD	identical by descent. 50, 157
ICA	independent component analysis. 125, 126, 131, 132, 136, 140, 142, 143, 145, 148, 150, 162
kPCA	kernel principal component analysis. 124, 125, 131, 132, 135, 136, 142, 148, 150, 162, 164, 165

LD	linkage disequilibrium. 34, 35, 44–47, 52, 70, 115–118, 120, 170, 171, 184, 186, 188
LiMMBo	linear mixed model with bootstrapping. 87, 89–91, 93–98, 101–105, 111, 115, 119, 121, 151, 191
LLE	Locally linear embedding. 124, 125, 127, 132, 135, 136, 140, 142, 144, 148, 164
LLR	log-likelihood ratio. 41, 97
LMM	linear mixed models. 48–50, 55–57, 80, 85–90, 101, 104, 108, 115, 116, 119, 145, 147, 183
LVM	left ventricular mass. 153, 154, 183
LVNC	left ventricular non-compaction. 178, 187
MAF	minor allele frequency. 157, 158, 200
MAR	missing at random. 104, 105, 108, 112
MCAR	missing completely at random. 104, 105, 108
MDS	classical multi-dimensional scaling. 123–127, 129, 131, 132, 136, 140, 142, 145, 148, 164, 166, 167
MICE	multiple imputation by chain equations. 111, 112, 114
MLE	maximum likelihood estimator. 38, 39, 49
MNAR	missing not at random. 104, 105, 108
MRI	magnetic resonance imaging. 68, 122, 154, 155, 159, 162, 175, 176, 192, 193
mtGWAS	multi-trait genome-wide association study. 95–97, 104, 115–119, 167, 168, 170–172, 175, 183, 184, 188, 203, 209
mvLM	multivariate linear model. 97, 98, 169
mvLMM	multivariate linear mixed model. 89, 90, 95, 97–99, 115, 119
nMDS	non-metric multi-dimensional scaling. 125, 129, 131, 132, 135, 136, 140, 142, 145, 148, 164, 166, 167
PC	principal component. 48, 53, 72, 73, 98, 123, 124, 126, 142, 144, 208

PCA principal component analysis. 25, 52, 53, 123–127, 129–132, 136, 140, 142, 145, 148, 150, 164–168, 170, 172, 173, 192, 193, 208, 209

PEER probabilistic estimation of expression residuals. 124, 125, 131, 132, 135, 140, 142, 145, 148, 164, 193

QC quality control. 155–157, 206, 207

REML restricted maximum likelihood. 39–41, 49, 56, 57, 89–91, 93–96, 101

RFLP restriction fragment length polymorphism. 30, 32

RMSD root mean squared deviation. 94, 95

RRM realised relationship matrix. 50–52, 72, 86, 115, 183

RSS residual sum of squares. 54

SNP single nucleotide polymorphism. 34–36, 42, 43, 45–47, 51, 52, 70, 71, 80, 81, 83, 86, 98–100, 115–120, 139, 142, 144, 145, 147–150, 155–158, 167–176, 184, 186–189, 192, 193, 200, 202, 208, 210

stGWAS single-trait genome-wide association study. 104, 115–117, 183, 210

TSI Toscani in Italia. 71, 72

tSNE t-Distributed stochastic neighbourhood embedding. 125, 128–132, 135–137, 140, 142, 143, 145, 148, 150

uvLMM univariate linear mixed model. 98, 99

Summary

Over the past ten years, more than 4,000 genome-wide association studies (GWAS) have helped to shed light on the genetic architecture of complex traits and diseases. In recent years, phenotyping of the samples has often gone beyond single traits and it has become common to record multi- to high-dimensional phenotypes for individuals. Whilst these rich datasets offer the potential to analyse complex trait structures and pleiotropic effects at a genome-wide level, novel analytic challenges arise. This thesis summarises my research into genetic associations for high-dimensional phenotype data.

First, I developed a novel and computationally efficient approach for multivariate analysis of high-dimensional phenotypes based on linear mixed models, combined with bootstrapping (LiMMBo). Both in simulation studies and on real data, I demonstrate the statistical validity of LiMMBo and that it can scale to hundreds of phenotypes. I show the gain in power of multivariate analyses for high-dimensional phenotypes compared to univariate approaches, and illustrate that LiMMBo allows for detecting pleiotropy in a large number of phenotypic traits.

Aside from their computational challenges in GWAS, the true dimensionality of very high-dimensional phenotypes is often unknown and lies hidden in high-dimensional space. Retaining maximum power for association studies of such phenotype data relies on using an appropriate phenotype representation. I systematically analysed twelve unsupervised dimensionality reduction methods based on their performance in finding a robust phenotype representation in simulated data of different structure and size. I propose a stability criteria for choosing low-dimensional phenotype representations and demonstrate that stable phenotypes can recover genetic associations.

Finally, I analysed genetic variants for associations to high-dimensional cardiac phenotypes based on MRI data from 1,500 healthy individuals. I used an unsupervised approach to extract a low-dimensional representation of cardiac wall thickness and conducted a GWAS on this representation. In addition, I investigated genetic associations to a trabeculation phenotype generated from a supervised feature extraction approach on the cardiac MRI data.

In summary, this thesis highlights and overcomes some of the challenges in performing genetic association studies on high-dimensional phenotypes. It describes new approaches for phenotype processing, and genotype to phenotype mapping for high-dimensional datasets, as well as providing new insights in the genetic structure of cardiac morphology in humans.

1

Introduction

The field of quantitative genetics has come far since Fisher's initial studies on human growth traits in 1918. Although the concept of inheritance existed at this time, little was known about the molecule responsible. The discovery of the DNA structure in the 1950s and technical break-throughs in analysing its sequence in the following decades have allowed to investigate genetic variance on a detailed scale, moving from whole chromosomes and linkage studies to the analysis of DNA variation on a single-base pair level.

The developments in genotyping and sequencing technologies in recent years have made large scale studies on genetic variation feasible. With the sinking costs of genotyping techniques, the number of samples has risen and studies investigating the effects of single DNA bases often comprise thousands of individuals, in particular in the field of human genetics. Together with the increased number of samples, the number of phenotypes that are measured for each individual has grown from a few measurements to tens, hundreds or even thousands. The availability of these rich datasets provides great opportunities when studying the influence of genetic variation on phenotypic variance. However, it also poses technical challenges when analysing these datasets.

In this thesis, I identified some of these challenges and propose new methods for the genetic analysis of high-dimensional datasets. These new methods are first ex-

plored on simulation studies and subsequently applied to real datasets. Specifically, I developed a new approach for the joint genetic association testing of a large number of phenotypic traits and applied this method to a publicly available dataset of yeast growth traits. I explored different dimensionality reduction methods for very high-dimensional datasets and propose a new measure to define the stability of the dimensionality reduction. Finally, I analysed human heart morphology data for genetic associations, applying the methods from the dimensionality reduction study on simulated data. In this introduction, I will first give a general overview of the history and methods in quantitative genetics, followed by the description of statistical models relevant for this thesis. In order to help with an understanding of the genetic association studies on the human heart morphology data, I also introduce basic concepts of cardiac structure and development and their underlying genetics.

1.1. From ancient ideas of inheritance to the birth of modern genetics

The formulation of the concept of human inheritance –the passing on of traits from parents to offspring– can already be found in works of Hippocrates and Aristotle. In addition to their theory of the inheritance of acquired traits, Hippocrates and Democritus also describe a possible mechanism of inheritance [Zirkle, 1935], a concept later formalised as "pangenesis" by the English naturalist Charles Darwin [1868] and others such as the French Comte de Buffon [1749] and Genevan naturalist Charles Bonnet [1779]. The theory of pangenesis – which translates to whole (Greek: pan) origin (Greek: genesis) or birth (Greek: genos) – describes how the entire parental organism participates in passing on traits to the offspring. In this developmental theory of heredity, all cells in an organism were believed to secrete small particles called gemmules, which circulate through the body to congregate in the gonads. While this theory was quickly refuted, Darwin became renowned for his ideas about trait variation and the link to inheritance. In his famous work *On the Origin of Species* [1859], he postulates natural selection as the central concept of evolution, based on his observations of phenotypic variance in a population, differential fitness based on phenotype and the concept of heritability of this fitness [Lewontin, 1970].

The milestones in genetics made since Darwin's work *On the Origin of Species* as well as accompanying statistical models and techniques in molecular biology are depicted in figure 1.1.

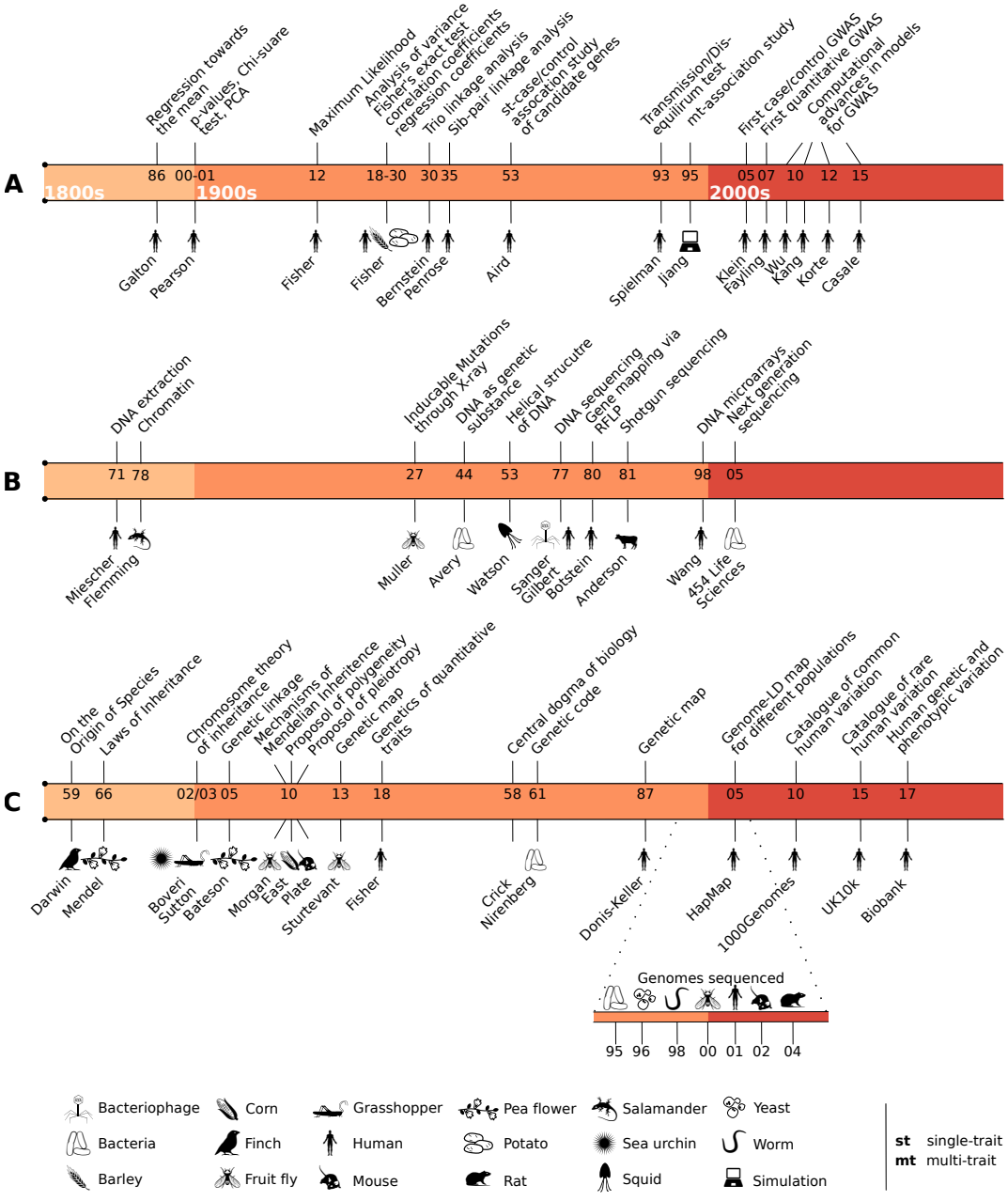

Figure 1.1: Genetics over time. A. Statistical concepts and B. techniques in molecular biology crucial for the advances in genetics. C. The developments in genetics from its birth by Mendel's Laws of Inheritance to large databases cataloguing genetic variation of thousands of individuals. Whilst there are many independent studies in all three areas contributing to the successes in genetics that we observe today, I have attempted to depict all major events that lead to the specific field of human quantitative genetics in the GWAS era. The development of mathematical models is focused towards models used in this thesis. The legend below the timelines specifies the symbols of the organisms used in the respective studies. As references for each entry, the first author of the corresponding publication is shown. Discoveries where multiple authors are named indicate independent studies at the same time making the same discovery/developing techniques.

1.1.1. Mendelian Laws of Inheritance

The Austrian friar Gregor Mendel was the first to systematically study the mechanisms of heritability. By cross-breeding different varieties of pea plants, he was able to follow the inheritance patterns of a number of visually-observable traits such as flower colour, seed shape and plant height. In 1866, he presented his observations in the paper *Versuche über Pflanzenhybride* (experiments on plant hybridisation) where he proposes three general concepts of inheritance which later became known as the Mendelian Laws of Inheritance: i) the Law of Independent Segregation (every individual contains two alleles for each trait which segregate in germ cells leading to a random transmission of alleles to the offspring), ii) the Law of Independent Assortment (traits are inherited independently of each other) and iii) the Law of Dominance (recessive alleles will be masked by dominant alleles and the trait corresponding to the dominant allele will be observed) [Mendel, 1866]. Although his work stayed widely unnoticed during his lifetime, his meticulous studies and documentation ensured his recognition as the father of genetics. In 1900, his work was independently rediscovered by the Dutch botanist Hugo de Vries [De Vries, 1900; Hannah & De Vries, 1950, translation into English] and –although contested by some based on their seeming lack in understanding of Mendel's work [Keynes & Cox, 2008; Monaghan & Corcos, 1986; Monaghan & Corcos, 1987]– the German botanist Carl Correns [Correns, 1900; Piernick & Correns, 1950, translation into English] and the Austrian agronomist Erich Tschermak [1900].

Around the same time, the British geneticist William Bateson set out to make Mendel's work accessible to the scientists not proficient in Mendel's native language German. He translated Mendel's original papers on the Laws of Inheritance [Mendel, 1866] and cross-breeding studies in *Hieracium* [Mendel, 1869] into English and published them in *Mendel's Principles of Heredity: a Defense* [Bateson, 1902]. In [1909], Bateson published an extended version of his original book which allowed Mendel's work to become known in the greater scientific world [Keynes & Cox, 2008], more than 40 years after their original publication. In addition to this work, the book *Recent Progress in the Study of Variation, Heredity, and Evolution* by his former student Robert H. Lock should be mentioned as the first English textbook embracing Mendel's ideas of inheritance [Lock, 1906; Edwards, 2013].

In addition to the rediscovery and translation of Mendel's ideas into English, two other branches of investigations contributed to the understanding of heredity and the identification of the molecular basis of the Laws of Inheritance from 1900 onward: biometrics and molecular biology.

1.1.2. Biometrics

Inspired by Darwin's work on evolution, his half-cousin Francis Galton was interested in mathematically describing and analysing evolutionary concepts. In 1886, he published *Regression Towards Mediocrity in Hereditary Stature*, offering a statistical approach towards understanding inheritance. Based on measurements of height in parents and their children, he observed that the "[t]he height-deviate of the offspring is, on the average, two-thirds of the height-deviate of its mid-parentage". He achieved the quantification of the deviation from the mean by fitting straight lines to the observed heights and finding their slope, thereby developing the technique of linear regression analysis and introducing the concept of correlation [Galton, 1886]. An extension of this work and descriptions of different statistical distributions and processes in heredity were published in his book *Natural inheritance* [Galton, 1889]. Karl Pearson formalised and extended Galton's statistical models for quantifying the effects of inheritance on trait variance by introducing the concept of p-values, the Chi-sqare test and principal component analysis (PCA) [Pearson, 1900; Pearson, 1901]. The marine biologist Walter Weldon applied these statistical concepts to data he had collected on shrimps and crabs [Weldon, 1890; Weldon, 1892], demonstrating selection in natural populations. Together, Galton, Pearson and Weldon are known as the founders of biometrics, the science of applying statistical methods to the study of evolution on quantitative traits, or as Galton described it: "The primary object of Biometry is to afford material that shall be exact enough for the discovery of incipient changes in evolution which are too small to be otherwise apparent." [Galton, 1901, editorial]. Despite the progress in understanding evolution in the light of statistical concepts, their direct study of heredity was impeded by their reluctance to acknowledge the validity of Mendelian genetics [Bulmer, 2003].

1.1.3. Molecular basis of inheritance

Advances in understanding the molecule responsible for inheritance were made by the Swiss physican and biologist Johannes Miescher and the German anatomist Walther Flemming. Miescher [1871] was the first to successfully isolate a substance he called *nuclein* –later known as DNA– from the nucleus . Flemming's experiments on salamander cells lead to the discovery of structures that could easily be stained by basophilic dies and he named them *chromatin* – "coloured material" (greek:khrōmat). He later found chromatin to be originating from the cell nucleus and did further studies into understanding cell division and mitosis [1878]. Al-

though both Miescher's and Flemming's methods and discoveries were crucial in the later identification of DNA as the carrier of inheritance, neither of them made the connection at the time.
With these advances in molecular and statistical techniques and the rediscovery of the Mendelian laws, the new discipline of genetic research attracted much attention.

1.2. The Laws of Inheritance on a cellular level

The first two scientists proposing how Mendel's Laws could work on a cellular level were the German biologist Theodor Boveri and the American Walter Sutton. By experimentally introduced double-fertilisations of sea urchin eggs and subsequent observations of developmental processes in the resulting embryos, Boveri claimed "dass eine bestimmte Kombination von Chromosomen zur normalen Entwicklung notwendig ist, und dieses bedeutet nichts anderes, als dass die einzelnen Chromosomen verschiedene Qualitäten besitzen müssen." i.e. "that a specific combination of chromosomes is necessary for a normal development which in turn means that each chromosome must harbour different qualities" [Boveri, 1902].

At the same time, Sutton described his observations in reduction division (later known as meiosis) and postulated that different chromosomes play different roles in development. Similar to Boveri, he came to the conclusion that "the phenomena of germ cell division and of heredity are seen to have the same essential features [...], with purity of units (chromosomes, characters) and the independent transmission of the same" [1903]. Both studies demonstrated the link between the Mendelian Laws of Inheritance and chromosomes as its carrier and are the basis for the chromosome theory of inheritance, also known as Boveri-Sutton Chromosome theory.

Around the same time, Bateson worked together with Edith Saunders and Reginald Punnett on experiments similar to Mendel's pea hybrids to understand the physiology of heredity. While they confirmed Mendel's original observations, they also discovered traits whose segregation did not follow the Law of Independent Assortment. Although they could not explain the mechanism of these observations, their results lead them to propose the concept of coupling or co-inheritance of traits [Bateson & al., 1905]. The first suggestion that this coupling of traits might result from genes lying on the same chromosome came by Lock [1906] & Edwards [2013].

With the progress in understanding Mendelian Laws on a cellular level came the establishment of terms describing certain entities and properties that are still in use today. In addition to his scientific contributions and his translation of Mendel's

works into English, Bateson became known for coining key terms in the field of genetics, even the term genetics itself [Dunwell, 2007]. He defined the units of inheritance transmission as allelomorphs, which became later abbreviated as alleles and introduced the terms homozygote and heterozygote for individuals carrying the same or different allelomorphs [Bateson, 1902]. The word gene as a term for the Mendelian factors or units of inheritance was introduced by the Danish botanist Johannsen [1911]. He also introduced the terms phenotype as the outward appearance of an individual and genotype as their genetic traits. The terms polygenetic, for traits that are governed by multiple genes [East, 1910] and pleiotropic, for genes that affect multiple, seemingly unrelated phenotypes [Plate, 1910, page 597] also made their first appearance at that time. While these terms are standard in today's field of genetics, their use in that time only rose slowly over time. For simplicity, however, I will from now on refer to any description of Mendelian factors or units as genes.

1.3. Genetic linkage

The American embryologist Thomas Morgan was critical of the ideas of Mendelian inheritance and chromosomes as its carrier [Allen, 1968], yet he would become a crucial figure in establishing the chromosomal theory of heredity and introducing other important concepts of inheritance. In his famous Fly Room at Columbia University, he worked on mutation and breeding experiments in the fruit fly *Drosophila melanogaster* aiming to discover mutations that would lead to the emergence of new species, as described in De Vries' mutation theory [Allen, 1968]. Instead, his experiments on fruit fly mutants for eye color (white instead of red) showed that the pattern of inheritance of the mutant trait followed the Mendelian Law of Dominance. In addition, he discovered that the factor determining eye color was linked to the factor for sex determination [Morgan, 1910; Morgan, 1911a] pointing towards the coupling of traits as observed by Bateson.

In subsequent years, Morgan and his students carried out extensive research on mutant fruit flies which lead to the discovery of crossing over (exchange of paternal and maternal chromosomal material during meiosis) and the formalisation of the concept of genetic linkage [Morgan, 1911b]. Based on the hypothesis that the degree of linkage between phenotypes would be inversely correlated to the linear distance of their genes on a chromosome, they developed the technique of genetic mapping: the localisation of genes underlying phenotypes on the basis of correlation with inheritance patterns [DNA variation], without the need for prior hypotheses about bio-

logical function. Using this technique, where the recombination rate between traits is used to estimate the relative distance of their genes, his student Sturtevant [1913] published the first genetic map[1] describing the relative distances between genes on the X chromosome of *Drosophila melanogaster*. Together with Herman Muller and Calvin Bridges, two other students of Morgan's, they published the book *The Mechanism of Mendelian Heredity* [1915], describing additional genetic maps for chromosome 2 and 3 and list groups of genes that are jointly inherited.

1.4. Towards quantitative genetics

With their development of genetic mapping and cross-breeding of *D. melanogaster* lines, Morgan, Sturtevant, Muller and Bridges conducted the first genotype-phenotype analysis studies. As in Mendel's original experiments and later, similar work by Bateson, Saunders and Punett, the phenotypes they observed were predominantly categorical, such as color of seeds and flowers in pea plants or the white-eyed phenotype in *Drosophila melanogaster*. In contrast, biometricians like Galton and Pearson analysed quantitative traits such as height. Their models fit with the Darwinian model of gradual change through natural selection, but did not explain the mode of inheritance. A great advance in genotype-phenotype mapping allowing for the analysis of quantitative traits came about with the work by the British statistician and biologist Ronald Aylmer Fisher.

An undergraduate student at the University in Cambridge, Fisher [1912] published his first paper *On a absolute criterion for fitting frequency curves* where he outlined the fundamental ideas of maximum likelihood estimation. He later extended on this work and by 1922, he had established the properties of the maximum likelihood estimator such as consistency and minimum variability [Fisher, 1922b] that is still used today [Hald, 1999]. He demonstrated the utility of maximum likelihood estimation in genetics by solving a number of equations to elucidate a genetic map of eight *Drosophila melanogaster* genes based on their crossing over frequencies [Fisher, 1922d]. In the same year and years to follow, he published a series of papers where he derived the distribution and significance testing of regression coefficients, correlation ratios and multiple regression coefficients [Fisher, 1922c; Fisher, 1928], an exact test for two-by-two contingency tables with small expectations (Fisher's exact test) [Fisher, 1922a], partial correlation coefficients [Fisher, 1924b] and the variance ratio,

[1] As opposed to physical maps which are based on exact chromosomal position and were only possible with the development of molecular biology techniques to examine DNA molecules directly [Brown, 2002]

later named after Fisher as the F statistic [Fisher, 1924a].

In 1918, the cornerstone for quantitative genetics was laid with his publication *The correlation between relatives on the supposition of Mendelian inheritance* where he showed that biometrics and Mendelianism are not contradictory but complimentary [Fisher, 1918]. Specifically, by analysing levels of phenotypic correlation between individuals of differing degrees of relatedness, he showed that the observed phenotypic variation can result from Mendelian inheritance. He further distinguished between two different types of genetic components contributing to the phenotype, one simply ascribed to genotypes and the other to "essential genotypes". Today, these components are known as broad-sense and narrow-sense heritability. Broad-sense heritability is the proportion of phenotypic variance explained by the entire genetic variation including additive, dominance (allelic interaction within loci) and epistatic (allelic interaction between loci) genetic effects, while narrow-sense heritability is defined as the ratio of additive genetic variance to total phenotypic variance.

As an additional statistical concept, it was in this work that Fisher defined the term variance as "the square root of the mean squared error". The analysis of variance in biological experiments would be of interest to Fisher in his appointment at Rothamsted Experimental Station where he analysed data from crop experiments with respect to different variance components and developed statistical techniques such as the analysis of variance (ANOVA) [Fisher, 1921; Fisher & Mackenzie, 1923; Eden & Fisher, 1929].

Extending on his 1918 work on trait correlation in light of Mendel's Laws, Fisher published the book *The Genetical Theory of Natural Selection* where he reconciled the long-standing ideas of Darwin's evolutionary theory and Mendelian inheritance. He gives the first, comprehensive quantitative theory of sexual selection, evolution of recombination rates, polymorphism and many more concepts found in today's field of population genetics [Fisher, 1930].

1.5. Progress in deciphering the molecular mechanisms of inheritance

Large steps forward in the molecular understanding of inheritance were the discovery of DNA as the genetic material in 1944 [Avery & al., 1944] and its composition from the four bases adenine, thymine, cytosine and guanine [Vischer & Chargaff, 1948; Chargaff & al., 1949; Chargaff & al., 1952] as well as the resolution of the DNA structure almost a decade later [Watson & Crick, 1953]. These insights brought for-

ward an understanding of other biological concepts such as protein synthesis and enabled Francis Crick to postulate the central dogma of biology: information is transmitted from DNA and RNA to proteins, but information cannot be transmitted from a protein to DNA [Crick, 1958]. The deciphering of the genetic code through Nirenberg and others followed a few years later [Nirenberg & Matthaei, 1961; Crick & al., 1961; Matthaei & al., 1962].

1.5.1. Novel genotype mapping techniques

Three discoveries and novel techniques at the beginning of the 1970s opened the door for the development of new genetic mapping approaches: the discovery of restriction enzymes [Smith & Welcox, 1970; Morrow & Berg, 1972], the ability to clone and amplify specific DNA sequences [Jackson & al., 1972; Cohen, 1973], and the detection of specific DNA sequences from a large pool of DNA fragments (Southern plot) [Southern, 1975]. Based on these techniques, restriction fragment length polymorphism (RFLP) analysis was developed, which allows for the identification of variants from within a specific genomic region using restriction enzyme-digested DNA [Grodzicker & al., 1974; Botstein & al., 1980]. Initially, RFLP analysis was used for genetic linkage maps in model organisms [Goodman & al., 1977; Cameron & al., 1979] and target genes in human [Kan & Dozy, 1978; Jeffreys, 1979; Tuan & al., 1979]. Based on theoretical considerations of using RFLP analysis for a general, target-free genetic mapping in humans [Botstein & al., 1980], the first human genetic map was published in 1987 [Donis-Keller & al., 1987].

1.5.2. Deciphering DNA sequences

While these mapping efforts were underway, the independent development of two different DNA sequencing techniques by two groups, one Frederick Sanger and the other Walter Gilbert together with Allan Maxwell, were a further big leap in understanding the biological basis of genetic variation [Sanger & al., 1977; Maxam & Gilbert, 1977]. Sanger's method of DNA sequencing with chain-terminating inhibitors eventually became the standard for DNA sequencing and subsequent innovations lead to the development of automatic sequencing machines which allowed for sequencing lengths of about one kilobase [Hunkapiller & al., 1991]. For sequencing longer stretches of DNA, a novel strategy named shotgun sequencing was developed [Staden, 1979; Anderson, 1981]. In shotgun sequencing, the long DNA of interest is randomly broken up into shorter DNA fragments which are cloned and

sequenced separately. The occurrence of overlapping DNA fragments given by the random nature of creating the short fragments allows for the *in silico* reconstruction of longer DNA fragments.

In 1995, the first genome of a living organism –the bacteria *H. influenzae*– was sequenced and assembled by shotgun sequencing [Fleischmann & al., 1995]. The genomes of other model organisms were to follow in subsequent years (yeast [Goffeau & al., 1996], *C. elegans* [C. elegans Sequencing Consortium, 1998], *D. melanogaster* [Adams & al., 2000]) until the first draft of the human genome was published in 2001 [International Human Genome Sequencing Consortium, 2001].

The sequence of the human genome, the development of faster, massively-parallel next-generation sequencing techniques (reviewed in [Shendure & Ji, 2008; Heather & Chain, 2016]) and DNA microarrays that allow for the genotyping of hundreds of thousands of genetic markers simultaneously [Wang & al., 1998], started a new era of human genetic and genomic research.

1.6. From genetic linkage analysis to genome-wide association studies

1.6.1. Genotype-phenotype mapping until the 1990s

Genotype-phenotype mapping approaches today can broadly be classified into genetic linkage analyses and population-based association studies. Genetic linkage analysis for human traits had already been applied in the 1930s [Bernstein, 1930; Penrose, 1935], while association studies only became known in the 1950s. For a clearer description of the methods and results, the following sections describe the developments in human quantitative genetics based on study type rather than in their chronological order.

Genetic linkage analysis

Genetic linkage analysis investigates the relationship between a given locus and the trait or disease of interest. As with Morgan's linkage studies in *D. melanogaster*, today's methods are also based on the observation that genetic markers in close physical proximity on a chromosome remain mainly linked during meiosis. By following the segregation of a specific trait in family pedigrees, the recombination rates between genetic markers can be estimated and their relative genomic position determined. To quantify the likelihood of linkage, a variety of measures with different

pedigree requirements have been developed. Some required full parent-offspring trios [Bernstein, 1930; Haldane, 1934], while others showed the possibility of determining genetic linkage based on sib-pairs alone [Penrose, 1935]. A commonly used test allowing for different pedigree structures is the sequential probabilty ratio test for linkage [Morton, 1955; Pulst, 1999]. In this test, the logarithm of the odds that the loci are linked is divided by the logarithm of the odds that the loci are unlinked. This log likelihood of the odds score serves as the measure for the likelihood of linkage. Genetic linkage studies often require strict assumptions about the underlying genetic models such as specification of penetrance and disease gene frequency [Morton, 1955; Pulst, 1999] and have a number of potential confounding variables such as genetic heterogeneity and accurate diagnosis [Bird, 1993]. Nevertheless, linkage studies have been successful in pinpointing genomic loci associated with disease. Initially restricted to known genes or gene products such as haemoglobin (linked to sickle-cell thalassaemia [Ingram & Stretton, 1959]) or haemophilia and colour-blindness [Haldane & Smith, 1947], the development of techniques such as RFLP mapping (section 1.5.1) enabled the detection of genetic markers in candidate genes. With these markers, linkage analysis could be extended to a greater number of candidate genes and led to the discovery of genetic links to diseases such as Huntington's disease [Gusella & al., 1983], cystic fibrosis [Kerem & al., 1989] and bipolar disorder [Baron & al., 1987].

Association studies

In contrast to linkage studies with the association between locus and trait in pedigrees, association studies investigate the relationship of a genetic marker frequency and the trait in a population. The frequencies of the genetic markers in individuals carrying the trait (cases) are compared to those in individuals without the trait (controls). Genetic markers whose frequencies are increased in cases compared to controls are thought to be associated with the risk for diseases. Often, the significance of the association is evaluated via a simple χ^2-test. As with linkage analysis, population association studies were initially limited to known genes or gene products such as in the association for blood antigens and stomach cancer [Aird & al., 1953; Aird & al., 1954]. With the new techniques for determining genetic markers in candidate genes, association studies successfully identified gene-disease associations in for instance diastrophic dysplasia [Hästbacka & al., 1992] and Alzheimers' Disease

[Strittmatter & Roses, 1996][2].

Jiang & Zeng [1995] provided an extension to the population-association model, leaving the strict case-control design and proposing a method to detect association with multiple quantitative traits. In the quantitative association study, an individuals genotype is represented numerically and a model can be fit directly to the genotypes and the continuous trait without relying on case-control status. In the linear model framework introduced by Jiang and colleagues, multiple traits are jointly analysed for genetic association, testing different models such as pleiotropic effects, and gene-environment interaction [Jiang & Zeng, 1995].

1.6.2. "Common disease–common variant" hypothesis

By the mid-1990s, genotype-phenotype mapping in humans was largely focused on candidate gene mapping through either linkage analysis or association studies. Linkage analysis had been very successful in identifying genes linked to Mendelian and monogenetic disorders with 671 genes for which at least one disease-related locus[3] was detected by 1995. Population-based association studies had so far detected about 250 genes associated with disease or dichotomous traits [Hirschhorn & al., 2002]. However, the number of reproducible results was notably lower and showed the difficulties associated with case-control population association studies. Major limitations were seen in the susceptibility to population stratification [Lohmueller & al., 2003] and the low *a priori* probability of the tested gene to be causal. In addition, for illnesses such as heart disease, diabetes or hypertension, the risk of being affected is likely a combination of multiple genetic and environmental factors [Hunter, 2005], which stands in stark contrast to the pattern observed in monogenetic diseases. In monogenetic diseases, the presence of a genetic factor or factors (dominant or recessive) almost completely predicts the presence of diseases such as cystic fibrosis or Huntington's Disease and these factors are generally of low frequency [Sankaranarayanan, 1998]. In the complex diseases, the genetic risk factor may be present in higher frequency and only lead to a small increase in disease risk [Reich & Goldstein, 2001]. Based on these arguments, the "common disease–common variant" hypothesis had been proposed, stating that common polymorphisms may play a role in the susceptibility to common diseases [Risch & Merikangas, 1996;

[2]Spielman & al. [1993] reconciled linkage analysis and case-control association studies, by formally introducing the transmission/disequilibrium test which tests directly for linkage between a disease and marker locus which is known to show population association.

[3]Statistics extracted from Online Mendelian Inheritance in Man: `https://omim.org`; search parameters: "date_updated:1981/1-1995/12"

Lander, 1996; Chakravarti, 1999; Reich & Goldstein, 2001]. For detecting common variants with small or moderate effect sizes, association studies are a more powerful tool than linkage analyses [Ott & al., 2015] and became the method of choice to investigate common disease variants on a genome-wide level. To enable systematic genome-wide screens of common variants, three components were needed: a catalogue of common variation in the human population, experimental techniques to obtain these genotypes in large cohorts, and the computational techniques for the subsequent analyses.

1.6.3. Databases of human variation

The first genome-wide database of common human sequence variation was created within the scope of the International HapMap project which was launched in 2002 [The International HapMap Consortium, 2005; The International HapMap Consortium, 2007; The International HapMap Consortium, 2010]. The HapMap project aimed at characterising the frequencies of single nucleotide polymorphisms (SNPs), i.e. variation on a single base pair level, for different human populations. Based on their genome-wide SNP frequencies, a comprehensive map for linkage disequilibrium (LD) –the non-random association of alleles at different loci [Lewontin & Kojima, 1960]– in different populations was created. By having included parent–offspring trios in the analysis, computational phasing [Stephens & al., 2001] enabled determination of the SNP contribution from each parent and the combination in which they were inherited. This particular combination of SNPs along a chromosome is termed haplotype and was the inspiration for the name of the project. The HapMap collection contains 1.6 million common SNPs in 1,184 reference individuals from 11 global populations. An extension of the work of the HapMap project, the 1000 Genome Project aimed to detect common human genetic variation by whole-genome sequencing of individuals from multiple populations. The project finished in 2015, providing genotypes and haplotypes at more than 88 million variants, including SNPs, short insertions or deletions, and structural variants for 2,504 individuals from 26 populations [1000 Genomes Project Consortium, 2011; 1000 Genomes Project Consortium, 2012; 1000 Genomes Project Consortium, 2015]. The work of the UK10K consortium complemented the work of both previous projects and extends the spectrum of observed genetic variation to rare variants in nearly 10,000 individuals from population-based and disease collections [UK10K Consortium, 2015]. While the major focus of these consortia laid in the collection of comprehensive genotype data, a new resource combining both genotype and phenotype data of more

than 500,000 individuals has recently been published. Phenotypes collected within this resource, the UK Biobank, cover amongst others anthropometric, cardiac and disease phenotypes [Sudlow & al., 2015].

1.6.4. Genotyping of large cohorts

Genotype data of common variants is standardly obtained from DNA microarrays which allow for the genotyping of hundreds of thousands of common SNPs simultaneously [Wang & al., 1998]. Based on the LD structures found in the reference panels (described above), haplotypes of the individuals can be estimated. Comparing the estimated haplotypes of the individuals to haplotype patterns in the reference panel enables imputation of unobserved genotypes in the study cohort. A number of different methods for genotype imputation have been developed including IMPUTE2 [Howie & al., 2009], Beagle [Browning & al., 2007] and MaCH [Li & al., 2010] (reviewed in [Marchini & Howie, 2010]). Via imputation, the number of genotypes per individual can be extended from the hundred thousands on the genotyping array to millions of observed variants in the reference datasets. Using these imputed genotypes for association studies can increase the power of the study and presents a high-resolution view of all SNPs in the associated region [Marchini & Howie, 2010].

1.6.5. Genome-wide association studies

The first successful study to test the "common disease–common variant" hypothesis without gene-based selection of genetic markers was conducted in 2005. Klein & al. [2005] carried out a case-control genome-wide association study (GWAS) for age-related macular degeneration and found a SNP in complement factor H to be associated with an increase in disease risk. Similar to population association studies of candidate genes, the significance of each SNP-disease association was tested via a χ^2-test and the resulting p-values subsequently corrected for multiple testing via Bonferroni correction (see section 1.7.4). Soon after, the Wellcome-Trust case-control consortium published large case-control GWAS for seven common diseases, including bipolar disorder, coronary heart disease and type I and II diabetes [Burton & al., 2007]. In the same year, the first GWAS on quantitative traits followed. Two research groups investigated the genetic effects on body mass index and found links to the FTO gene. In addition, these BMI-associated SNPs also showed strong association to type II diabetes [Frayling & al., 2007] and other SNPs within the FTO gene were also associated to weight and hip-circumference [Scuteri & al., 2007]. Both studies used

a simple linear model (see section 1.7) to find the association of the genetic marker as the explanatory variable and BMI as response variable.

In the following years, the methods for GWAS were extended to enable the genotype-phenotype mapping for sets of SNPs [Wu & al., 2010; Casale & al., 2015], the joint mapping of multiple traits [Korte & al., 2012; Yang & al., 2011; Bottolo & al., 2013; Casale & al., 2015] and the use of more complex models to account for population stratification such as mixed model approaches [Kang & al., 2010; Lippert & al., 2011; Zhang & al., 2010; Svishcheva & al., 2012] and general estimating equations [Cupples & al., 2007].

Based on these methods, thousands of GWAS have been conducted covering common diseases (e.g. asthma [Noguchi & al., 2011; Pickrell & al., 2016], coronary heart disease [Wild & al., 2011; Takeuchi & al., 2012; Lu & al., 2012], migraine [Pickrell & al., 2016; Gormley & al., 2016], blood pressure [Kato & al., 2011; Franceschini & al., 2013]), anthropometric traits (e.g. height [Lango & al., 2010; Wood & al., 2014], weight [Willer & al., 2009], BMI [Speliotes & al., 2010; Yang & al., 2012], waist-hip ratio [Lindgren & al., 2009; Heid & al., 2010]) and other non-disease related quantitative phenotypes (e.g. eye color [Eriksson & al., 2010; Candille & al., 2012; Zhang & al., 2013], freckling [Sulem & al., 2008], facial morphology [Paternoster & al., 2012], hair greying [Adhikari & al., 2016]). The results of these studies are collected in the GWAS catalogue, which currently contains 3,092 publications and 49,769 unique SNP-trait associations [MacArthur & al., 2017, accessed 10.09.2017].

In GWAS, the genetic variants associated with the traits of interest are often not directly informative with respect to finding the target gene and causal mechanism. However, bioinformatics fine-mapping approaches and molecular follow up studies have been successful in identifying target genes and proposed mechanisms for many GWAS discoveries. For some of these GWAS results, the mechanistic insights have triggered drug development and drug repurposing studies. With the increasing sample sizes such as in the UK Biobank resource ==[Sudlow & al., 2015], many new genetic variants are likely to be discovered in the years to come. They will help accounting for more genetic variation and likely yield more accurate genetic predictors (reviewed in [Visscher & al., 2017]).

1.7. Linear models for genome-wide association studies

Simple linear models and linear mixed models are widely applied in genetic association analysis. They offer great control for confounding factors and allow for the

joint analysis of multiple traits. In the following sections, I will describe the general model specifications and parameters, their estimation and application to genetic studies. I will outline the challenges for linear models in GWAS and the approaches developed to overcome these challenges.

For mathematical model descriptions throughout this thesis, I used the following notation: bold, small letters symbolise one-dimensional column vectors e.g. $\mathbf{v}$ and bold capitalised letters matrices e.g. $\mathbf{M}$. A normal distribution is specified by $\mathcal{N}(\text{mean}, \text{variance})$, a multivariate normal by $\mathcal{N}_{\text{r}\times\text{c}}(\text{mean}, \text{variance})$ and a matrix-variate normal by $\mathcal{MN}_{\text{r,c}}(\text{mean}, \text{variance}_{\text{rows}}, \text{variance}_{\text{columns}})$, where r and c are the row and column dimensions, respectively.

1.7.1. Linear regression

In the linear model, the continuous response variable (e.g. phenotype) is described as a linear function of one or more explanatory variables (e.g. genotype and covariates). With N representing the number of samples, y_i the response variable for sample i, $\{x_{i1}, x_{i2}, \dots, x_{iF}\}$ the F explanatory variables for sample i and β_f their corresponding weights, the linear model can be cast as

$$y_i = \sum_{f=1}^{F} x_{if}\beta_f + \psi_i, \quad \text{with } \psi_i \sim \mathcal{N}(0, \sigma_e^2). \tag{1.1}$$

In this model, the residual term ψ_i captures measurement noise and other unaccounted factors that influence the response variable. ψ_i is modelled to follow a normal distribution with mean 0 and variance σ_e^2 and to be independent across samples, i.e. with covariance equals to zero: $\text{cov}(\psi_i, \psi_j) = 0$.

Equivalently, equation (1.1) can be written in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\psi}, \quad \text{with } \boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I}_N), \tag{1.2}$$

where the $N \times N$ identity matrix $\mathbf{I}_N$, the response vector $\mathbf{y}$, the matrix of explanatory variables $\mathbf{X}$, the weight vector $\boldsymbol{\beta}$ and the vector of residuals $\boldsymbol{\psi}$ are defined as:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1F} \\ x_{21} & x_{22} & \cdots & x_{2F} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NF} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix} \text{ and } \boldsymbol{\psi} = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_N \end{bmatrix}. \tag{1.3}$$

The model in equation (1.2) describes the probability distribution of the response variable, given the explanatory variables and corresponding parameter estimates $\boldsymbol{\beta}$ and σ_e^2. This probability is also known as the likelihood function or likelihood $\mathcal{L}$ and plays a key role in statistical inference of the model parameters. Casting equation (1.2) as the likelihood of the model parameters $\boldsymbol{\beta}$ and σ_e^2 yields

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right) = p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma_e^2\right) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta},\ \sigma_e^2 \mathbf{I}_N\right) \tag{1.4}$$

or directly expressed in terms of the response variable

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta},\ \sigma_e^2 \mathbf{I}_N\right). \tag{1.5}$$

The parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_e^2$ that maximise the likelihood function are the maximum likelihood estimators (MLEs) of $\boldsymbol{\beta}$ and σ_e^2. In order to improve numerical stability, the log likelihood is commonly used instead of the likelihood[4]. The full log-likelihood is expressed as

$$\begin{aligned}
\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right) &= \log p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma_e^2\right) && (1.6)\\
&= \log \prod_{i=1}^{N} p\left(y_i \mid \mathbf{X}, \boldsymbol{\beta}, \sigma_e^2\right) && (1.7)\\
&= -\frac{N}{2}\log\left(2\pi\right) - \frac{N}{2}\log\sigma_e^2 \frac{1}{2\sigma_e^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right). && (1.8)
\end{aligned}$$

and the MLE

$$\hat{\boldsymbol{\beta}}, \hat{\sigma}_e^2 = \operatorname{argmax}_{\boldsymbol{\beta}, \sigma_e^2} \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right). \tag{1.9}$$

The MLE of $\boldsymbol{\beta}$ and σ_e^2 are found by finding the maxima of the partial derivates of equation (1.8)

$$\begin{aligned}
\left(\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right)}{\partial \boldsymbol{\beta}}\right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \sigma_e^2=\hat{\sigma}_e^2} &= 0 && (1.10)\\
\left(\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right)}{\partial \sigma_e^2}\right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \sigma_e^2=\hat{\sigma}_e^2} &= 0, && (1.11)
\end{aligned}$$

[4]Since the logarithm is monotonically increasing, maximisation of the log-likelihood is equivalent to maximising the likelihood itself, but offers mathematically convenient properties.

yielding

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{1.12}$$

$$\hat{\sigma}_e^2 = \frac{1}{N}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \tag{1.13}$$

$$= \frac{1}{N}\left(\mathbf{y} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right)^T\left(\mathbf{y} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right). \tag{1.14}$$

Restricted maximum likelihood

In Gaussian models as in equation (1.5), the MLE of the mean estimate $\hat{\boldsymbol{\beta}}$ is unbiased whereas the MLE of the variance component $\hat{\sigma}_e^2$ suffers from a downward bias. The bias of $\hat{\sigma}_e^2$ originates from the loss in the degrees of freedom as a consequence of estimating $\hat{\boldsymbol{\beta}}$ from the data. Patterson & Thompson [1971] proposed a solution for a $\boldsymbol{\beta}$-free estimation of $\hat{\sigma}_e^2$ via restricted maximum likelihood (REML). In short, for a linear regression model with

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\phi}, \quad \text{with } \boldsymbol{\phi} \sim \mathcal{N}\left(\mathbf{0}\,,\; H\left(\theta\right)\right), \tag{1.15}$$

where the covariance term is now described as a general covariance matrix $H\left(\theta\right)$ parameterised by θ, the REML is based on the projection $\mathbf{w}$ of $\mathbf{y}$ by a matrix $\mathbf{A}$ with:

$$\mathbf{A}\mathbf{X} = 0. \tag{1.16}$$

Using equation (1.16) and rewriting equation (1.15) in terms of the projection $\mathbf{w}$

$$\mathbf{w} = \mathbf{A}\mathbf{y} = \mathbf{A}\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\phi}\right) = \mathbf{A}\boldsymbol{\phi} \tag{1.17}$$

yields an expression of $\mathbf{y}$ that is free of $\boldsymbol{\beta}$. By directly estimating $\mathcal{L}\left(\theta \mid \mathbf{A}\mathbf{y}\right)$, the unbiased estimate for θ can be found. In case of the linear regression in equation (1.5) with $H\left(\theta\right) = \sigma_e^2\mathbf{I}_N$, the REML estimate of variance component σ_e^2 is

$$\hat{\sigma}_e^2 = \frac{1}{N-F}\left(\mathbf{y} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right)^T\left(\mathbf{y} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right). \tag{1.18}$$

Comparing equation (1.14) and equation (1.18), it becomes evident that the MLE and REML for the variance component only differ in the denominator where N is replaced by $N - F$, reflecting the loss in F degrees of freedom (number of explanatory variables in the model).

In more complex linear models such as linear mixed models (section 1.7.6), the

estimation of the variance component is equally more complex depending on the covariance structure of the residual effects. The detailed derivation of the REML estimators of parameters from the linear model framework used throughout this thesis can be found in [Casale & al., 2015, Supplementary material].

1.7.2. Simple linear model for genotype associations with a single trait

In genetic association studies, the simple linear model describes the phenotype of interest as the sum of the genetic effect and often additional covariate effects such as height or sex:

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{F}\boldsymbol{\alpha} + \boldsymbol{\psi}, \text{ with } \boldsymbol{\psi} \sim \mathcal{N}\left(0\,,\, \sigma_e^2 \mathbf{I}_N\right) \tag{1.19}$$

and

the phenotype vector for N samples $\mathbf{y} \in \mathcal{R}^{N,\,1}$,
the genetic profile of the SNP being tested $\mathbf{x} \in \mathcal{R}^{N,\,1}$,
the effect size of the SNP $\beta \in \mathcal{R}^{1,\,1}$
the matrix of K covariates $\mathbf{F} \in \mathcal{R}^{N,\,K}$ and
the effect of covariates $\boldsymbol{\alpha} \in \mathcal{R}^{K,\,1}$.

The residual noise $\boldsymbol{\psi}$ is assumed to follow a normal distribution that is independent across the N samples.

In order to model the genotypes quantitatively, they have to be encoded numerically. For genetic association studies in diploid organisms, there are different inheritance models based on the combination of parental alleles a and b (for bi-allelic loci). In a recessive inheritance model (with respect to b), the phenotype is only observed in the presence of two b alleles and the genotypes are encoded as $aa = 0$, $ab = 0$ and $bb = 1$. In the dominant model for b, where only one copy of the allele is necessary to confer the phenotype, the genotypes are $aa = 0$, $ab = 1$ and $bb = 1$. The additive, or allelic dosage, model for b assumes that the effect on the phenotype is proportional to the allele count of b with $aa = 0$, $ab = 1$ and $bb = 2$ [Bush & Moore, 2012]. For association testing without prior knowledge or assumptions about the mode of inheritance, the additive model has been widely adapted and will be used throughout this thesis. It shows reasonable performance across all three models for the majority of effects, however, may suffer from a loss in power for recessive traits with a low causal allele frequency [Lettre & al., 2007].

As equation (1.5) shows, the phenotype is assumed to follow a normal distribution.

In order to avoid model misspecification when testing for genetic association, it is common practice to ensure approximate normality by transforming the observed phenotypes via methods such as Cox-Box [Etzel & al., 2003; Yang & al., 2006] or inverse normal transformation [Scuteri & al., 2007; Guan & Stephens, 2008; Anttila & al., 2010; Casale & al., 2015]. For any association tests conducted throughout this thesis, a rank-based inverse normal transformation is applied to each phenotype.

1.7.3. Testing the genotype association for significance

The significance of the association between phenotypes and the genetic markers can be assessed by testing that the genetic variant has an effect ($\beta \neq 0$) versus the null hypothesis $\mathcal{H}_0$of not having an effect ($\beta = 0$) on the phenotype. The log-likelihood ratio (LLR) test statistic Λ is a commonly used statistic to compare the likelihood of the full model $\mathcal{H}_1$to the one of the null model $\mathcal{H}_0$:

$$\mathcal{H}_1 : \mathbf{y} \sim \mathcal{N}\left(\mathbf{x}\beta + \mathbf{F}\boldsymbol{\alpha},\ \sigma_e^2\mathbf{I}_N\right), \tag{1.20}$$

$$\mathcal{H}_0 : \mathbf{y} \sim \mathcal{N}\left(\mathbf{F}\boldsymbol{\alpha},\ \sigma_e^2\mathbf{I}_N\right). \tag{1.21}$$

The LLR test statistic Λ is defined as

$$\Lambda = \mathcal{L}(\hat{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\sigma_e}) - \mathcal{L}(0, \bar{\boldsymbol{\alpha}}, \bar{\sigma_e}) \tag{1.22}$$

where $\mathcal{L}(\hat{\beta}, \hat{\alpha}, \hat{\sigma_e})$ are the REML of $\mathcal{H}_1$and $\mathcal{L}(0, \bar{\alpha}, \bar{\sigma_e})$ the REML of $\mathcal{H}_0$. 2Λ follows a χ^2_d-distribution with d degrees of freedom equal to the number of tested parameters [Wilks, 1938] and allows for the calculation of the p-value as :

$$P(\Lambda) = \int_{2\Lambda}^{\infty} \chi^2\left(x; d\right) dx = 1 - F_{\chi^2}\left(2\Lambda; d\right), \tag{1.23}$$

where $F_{\chi^2}\left(2\Lambda; d\right)$ is the cumulative density function of the χ^2-distribution with d degrees of freedom. For a single-variant single-phenotype test, the degrees of freedom are $d = 1$ (figure 1.2A, blue). The p-values derived from the χ^2-distribution can be used to interpret the association. The p-value is defined as the probability of finding the observed, or more extreme, results when $\mathcal{H}_0$is true [Krzywinski & Altman, 2013a], or in other words, it serves as an index measuring the strength of evidence against the null hypothesis [Sterne & al., 2001]. The p-values are compared to a predefined significance level α, which specifies the probability of rejecting a true null hypothesis. If $p < \alpha$, the null hypotheses is rejected. Falsely rejected null hypo-

thesis are classified as Type I errors, or false positives, and depend on the stringency of the α threshold. For instance, with $\alpha = 0.05$, 5% of all rejected null hypotheses might be true. Type II errors, or false negatives, occur when the null hypothesis is falsely accepted, i.e. a true association is not detected. The power in a GWAS is the proportion of true positives associations that can be detected, which corresponds to power $= 1 -$ Type II error rate [Krzywinski & Altman, 2013a; Krzywinski & Altman, 2013b].

In a GWAS, S genome-wide SNPs are assessed under $\mathcal{H}_0$(equation (1.21)). With the assumption that the wide majority of $\mathcal{H}_0$are true and potential confounding has been properly adjusted for (section 1.7.5), the genome-wide p-values follow a uniform distribution in $(0, 1]$ (figure 1.2B). To visually examine the p-value distribution, p-values are often depicted in quantile-quantile (qq) plots where the expected $-\log_{10}$ p-values[5] are plotted against the observed $-\log_{10}$ p-values (both sorted in increasing order). A GWAS is well-calibrated if the expected and observed p-value distribution only show deviations for SNPs associated with the phenotype (figure 1.2C). Deviations of the observed from the expected p-value distribution are commonly observed in GWAS of highly polygenic traits or in studies with confounding factors such as population structure and relatedness which can create spurious associations [Marchini & al., 2004; Balding, 2006; Spielman & al., 1993; Lander & Schork, 1994]. Strategies to adjust for these confounding effects and to tell them apart from the true polygenetic effects are described in (section 1.7.5).

1.7.4. Correcting for multiple hypotheses testing in GWAS

The underlying assumption of a GWAS is that the large majority of SNPs will have no impact on the phenotypes, i.e. for each SNP, one tests the null hypothesis of no effect versus the alternative hypothesis of a SNP effect that is different from zero and expects to accept the vast majority of these null hypotheses. However, when testing a large number of null hypotheses, it is likely to observe results with p-values below the significance level even if all null hypotheses are true. In a well-calibrated test, the number of false positive results depends on the *a priori* specified significance level α. For instance, with $\alpha = 0.05$ and ten million genome-wide SNPs, 5×10^5 tests would be expected to be false positives. Methods to correct for multiple hypotheses testing, i.e. reduce the number of Type I errors are reviewed in detail in [Shaffer, 1995]. The most commonly applied methods based on false discovery rate (FDR) and family-wise error rate (FWER) are described below.

[5]In practice, the expected $-\log_{10}$ p-values are obtained through S equally spaced numbers in $(0, 1]$

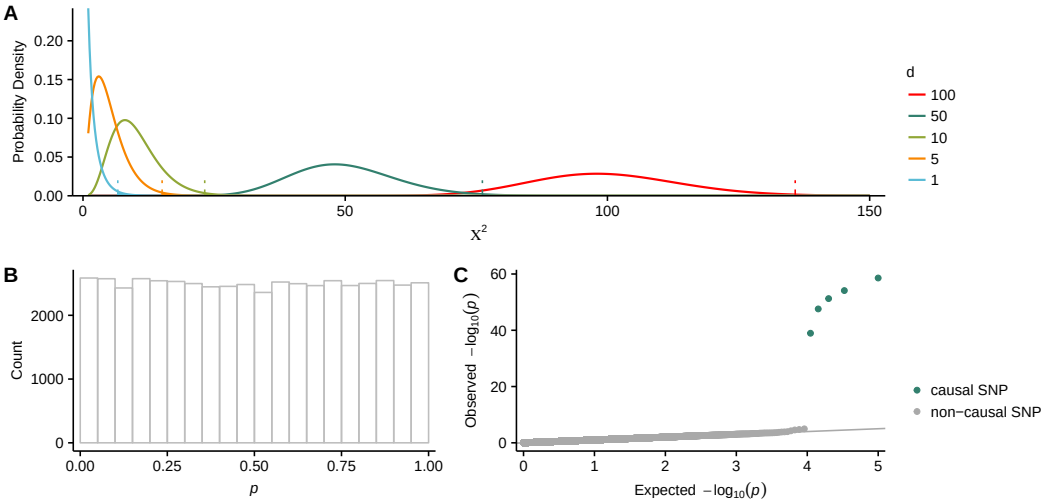

Figure 1.2: Distributions in LLR testing in GWAS. A. Cumulative density functions of χ^2-distributions with different numbers of degrees of freedom (d). The higher the number of degrees of freedom, the higher the χ^2-statistic (x-axis) has to be to obtain p-values regarded as conclusively showing that the null hypothesis is false (indicated as dotted lines and shaded regions under the curves for $\alpha = 0.05$). B. P-value distribution of a well-calibrated GWAS. P-values are derived from the associations of 50,000 bi-allelic SNPs from 1,000 individuals with a single quantitative phenotype. Out of the 50,000 SNPs, five SNPs were simulated to have an effect $\beta \neq 0$. The phenotype was simulated with default parameters as described in chapter 3. C. Quantile-quantile plot of the p-values from the associations in B. The five SNPs with $\beta \neq 0$ are indicated in green.

False discovery rate The FDR corrects for multiple testing based on the expected proportion of false discoveries. The FDR was introduced by Benjamini and Hochberg in [1995] and a number of other FDR-based correction methods were developed thereafter e.g. [Storey, 2002; Donoho & Jin, 2006; Sarkar, 2007]. The original method by Benjamini and Hochberg set out to control the expected values of the FDR based on the ratio of wrongly rejected N and total number of rejected null hypotheses R:

$$\text{FDR} = \mathbb{E}\left[\frac{N}{max\,(R,1)}\right], \tag{1.24}$$

where the maximum in the denominator protects against division by zero. The procedure works as follows: for a total number of m tests, with p-values $p_1, p_2, \ldots, p_m$ ordered in increasing order by their ranks $k_1, k_2, \ldots, k_m$ (smallest p-value p_1 with $k_1 = 1$), the adjusted p-value p'_i is determined as $p'_i = \frac{mp_i}{k_i}$. Choosing to accept all null hypotheses with $p'_i > \alpha$ ensures FDR $< \alpha$.

Family-wise error rate The FWER controls for the probability of observing at least one false positive result within a given experiment (family of tests) [Shaffer, 1995]. Among the FWER-based tests, the most simple procedure to adjust for multiple testing is multiplying all observed p-values $P = p_1, p_2, \ldots, p_m$ by the total number of tests m: $P' = mP$. This method to compute the adjusted p-values P' was proposed by Olive Dunn in 1961, based on properties of Bonferroni's inequalities [Dunn, 1961] and the method is commonly referred to as Bonferroni correction. Accepting all null hypotheses with $p'_i > \alpha$ ensures controlling for FWER $< \alpha$. The main assumption in Bonferroni-based adjusting for multiple testing is the independence of the conducted tests. In genome-wide tests of association, LD structure in the genome induces dependence of tests and correction for multiple testing by a strict multiplication of the total number of tests is too conservative.

Permutation-based adjusting for FWER In order to account for the dependency of the statistical tests in genetic association studies, one can use permutation-based approaches to control the FWER. In these approaches, the link between the parameter of interest i.e. the genotype and the observed phenotype is broken by random permutation of the genotype data across individuals. The association study is conducted T times on T random permutations of the data and the p-values of the permutation experiments $\bar{P}$compared to the observed p-values of association study. For each p_i, p'_i is calculated by recording the number of times p_i is smaller than any $\bar{p}_i$ and sub-

sequently dividing this number by the total number of permutations. Permutation-based approaches have been employed in whole-genome association studies (about 10,000 genotypes) for yeast [Brem & al., 2002; Ehrenreich & al., 2010; Bloom & al., 2013] and human genotype to gene expression association studies for adjusting on gene level [1000 Genomes Project Consortium, 2015]. In these studies the computational burden is moderate, whereas permutation studies for human GWAS with millions of SNPs might become impractical.

LD-corrected genome-wide significance threshold As an alternative to adjusting each p-value individually, a new α' can be specified which controls for the same level of type I errors as α but takes the number of tests that are conducted into account. For the conservative Bonferroni correction, which does not consider the genomic LD structure $\alpha' = \frac{\alpha}{m}$. For human GWAS, the multiplication factor for α has been estimated based on the estimated number of independent variants in the genome. It is is based on an observation of the HapMap project [The International HapMap Consortium, 2005] (section 1.6.3) where about 150 independent, common variants were found per 500 kb region. Extrapolating this number to the human genome size of $\sim$ 3.3 Gb, for $\alpha' = 0.05$ the genome-wide significance threshold was estimated as $\alpha' = 0.05 \times 150 \times (500\text{kb} \times 3.3\text{Gb})^{-1} = 5.05 \times 10^{-8}$. This estimate was later confirmed in a study using different methods for estimating the number of independent variants [Fadista & al., 2016] and is the commonly employed threshold in today's human GWAS. However, this threshold can be different in genetic studies of rare variants (for example [Xu & al., 2014]).

1.7.5. Accounting for population structure and genetic kinship

Confounding of association results based on genotypic differences between cases and controls had been a known challenge before the GWAS era [Spielman & al., 1993] and has remained a critical issue still. If population structure is not taken into account when testing for genotype-phenotype associations, associations might be observed that simply reflect the underlying population structure and lead to an increase in false positive results. Equally, real effects might be masked and genuine associations missed [Marchini & al., 2004]. In the case-control setting, this problem arises easily when the study consists of (undetected) subpopulations which are not evenly distributed among cases and controls. For SNPs where the allele proportions differ between the hidden subgroups, a false positive association will be recorded [Marchini & al., 2004; Balding, 2006]. Quantitative trait association studies can be

subject to similar issues. If the study cohort is comprised of individuals from different ethnicities, spurious associations can be detected that reflect ethnicity rather than causal variation. Campbell & al. [2005] demonstrated in an association study for height in a European-ancestry cohort that association could simply be attributed to differences in SNP frequencies across European ancestry subpopulations. Other studies confirmed allele-frequency differences within cohorts of the same ethnicity [Tian & al., 2008a; Tian & al., 2008b], thus emphasising the need for proper control of population structure. Similar issues arise for a more fine-scaled structure in the cohort induced by samples with different degrees of relatedness. When related individuals are present in the cohort, their genotypes do not reflect random and independent draws from the population frequencies. While this generally does not affect the allele frequency estimates, their variance might be greater than expected, leading to an overdispersed test statistic and increased false positive rate, as demonstrated by Bacenu, Devlin and Roeder for case-control settings and quantitative traits [Devlin & Roeder, 1999; Bacanu & al., 2002]. In addition to population structure and relatedness, spurious associations might arise in studies with recently admixed populations, as described by Lander & Schork [1994] and Ewens & Spielman [1995] where false positive disease associations were found due to allele frequency differences in the parent populations. A number of different methods have been developed to correct for confounding genotype structures.

Post-hoc adjusting

The first methods to adjust for genetic background structure was proposed by Devlin & Roeder [1999]. *Genomic Control* is based on the hypothesis that genetic background structure generates an inflation of the test statistics. Adjustment for population structure is achieved by estimating the inflation factor λ and dividing the test statistic of each association by λ. Extensions to their initial approach for case-control studies included partially modified approaches for estimating λ [Reich & Goldstein, 2001], its application for quantitative traits [Bacanu & al., 2002], and an adjusted approach to take the number of SNPs for the estimation of λ into account [Devlin & al., 2004]. The observation that inflation and sample size seemed to correlate lead Yang & al. [2011] to systematically study different parameters influencing inflation and they found λ to be a function of sample size, LD structure and narrow-sense heritability. Importantly, they showed that λ is also correlated with the number of causal variants, thus studies on traits with polygenetic inheritance can show inflation independent from confounding. Based on this observation, Bulik-Sullivan & al. [2015]

developed LD Score regression, a regression method for distinguishing confounding structures from polygenicity in GWAS. As with *Genomic Control*, the estimated inflation factor from LD Score regression can be used for the post-hoc adjusting of the test statistic.

Adjusting by subsampling

Shortly after the introduction of *Genomic Control*, Pritchard & al. [2000] proposed the concept of *Structured Associations*, where genetic markers unlinked to the phenotype are used to identify subpopulations of samples. Assigning the samples to their respective unstructured subpopulations and testing for association within subpopulations will essentially overcome the problem of population structure present in the overall study population. While useful and employed in association studies for a moderate number of genetic markers and samples [Li & al., 2004; Stein & al., 2009; Kulbrock & al., 2013], it is computationally expensive for large datasets [Price & al., 2006]. In addition, human genetic diversity is better approximated by continuous measures or gradients rather than discrete cluster membership [Serre & Pääbo, 2004; Price & al., 2006].

Relatedness and population estimates as model variables

In contrast to the post-hoc and subsampling approaches, adjusting for population structure and relatedness within the association model is possible by estimating the genetic relationship of the samples and using these estimates as additional variables. Studies on genotype variation in relation to geographical distance have demonstrated that geographic ancestries of individuals can be inferred from genetic markers [Rosenberg & al., 2002; Tang & al., 2005]. Sample clustering based on the genetics is thereby largely correlated with their geographic regions [Rosenberg & al., 2005]. In addition to capturing large scale population structure, genetic markers have also been employed to estimate shared ancestry and relatedness in natural populations [Lynch & Ritland, 1999; Ritland, 2000; Thomas, 2005]. Price & al. [2006] proposed to use genome-wide genetic markers to estimate a genetic sample-by-sample covariance matrix. The SNPs of this genetic covariance matrix represent continuous axes of genetic variation and can be used to adjust for population structure, either by *a priori* regression of the principal components from both the genotype and phenotype data, or by including them as additional covariates in the model. They showed that principal components correctly identified and corrected for population structure based

on geographic differences. However, principal components (PCs) perform poorly in modelling family structure or cryptic relatedness [Yu & al., 2006; Zhao & al., 2007; Kang & al., 2010; Casale & al., 2015]. Yu & al. [2006] have proposed to use a linear mixed model approach to control for population structure and relatedness. The key assumption in this approach is that phenotypic covariance between individuals based on population structure or relatedness is proportional to their relative relatedness. They showed together with Malosetti & al. [2007] and Zhao & al. [2007] that linear mixed models in the analysis of structured samples yield higher power while controlling better for type I errors than *Genomic Control*, *Structured Associations* and –PCs.

1.7.6. Linear mixed models

Linear mixed models (LMM) describe the linear relationship between the response vector and a number of fixed (deterministic) effects and random (unknown) effects. While fixed effects are modelled by estimating the effect sizes of known explanatory variables (equation (1.2)), random effects model a random variable for which distribution parameters are estimated. Specifically, for the response vector of N samples $\mathbf{y} \in \mathcal{R}^{N,\,1}$, the design matrix of F fixed effects $\mathbf{X} \in \mathcal{R}^{N,\,F}$ and their respective effect size vector $\boldsymbol{\beta} \in \mathcal{R}^{F,\,1}$, the design matrix of U random effects $\mathbf{Z} \in \mathcal{R}^{N,\,U}$ and the random effect $\mathbf{b}$, the linear mixed model is cast as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\psi}, \text{ with } \mathbf{b} \sim \mathcal{N}\left(0,\, \sigma_b^2\boldsymbol{\Sigma}\right) \text{ and } \boldsymbol{\psi} \sim \mathcal{N}\left(0,\, \sigma_e^2\mathbf{I}_N\right). \tag{1.25}$$

As in the simple linear model (equation (1.2)), the residual noise is assumed to follow a normal distribution with mean zero and variance parameter σ_b^2. In the formulation considered here, the covariance of the random effect is described by a known covariance matrix $\boldsymbol{\Sigma}$ and its variance parameter σ_b^2. Equation (1.25) can be expressed as the likelihood of the joint probability distribution of $\mathbf{y}$ and $\mathbf{b}$

$$p\left(\mathbf{y}, \mathbf{b} \mid \boldsymbol{\beta}, \sigma_b^2, \sigma_e^2\right) = p\left(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{b}, \sigma_e^2\right) p\left(\mathbf{b} \mid \sigma_b^2\right) \tag{1.26}$$

$$= \mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b},\, \sigma_e^2\mathbf{I}_N\right) \mathcal{N}\left(\mathbf{b} \mid \mathbf{0},\, \sigma_b^2\boldsymbol{\Sigma}\right) \tag{1.27}$$

To find the estimates of the unknown parameters $\boldsymbol{\beta}, \sigma_b^2, \sigma_e^2$, one can first marginalise out $\mathbf{b}$

$$p\left(\mathbf{y} \mid \boldsymbol{\beta}, \sigma_b^2, \sigma_e^2\right) = \int p\left(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{b}, \sigma_e^2\right) p\left(\mathbf{b} \mid \sigma_b^2\right) db \tag{1.28}$$

$$= \mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta},\ \sigma_b^2 \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}_T + \sigma_e^2 \mathbf{I}_N\right) \tag{1.29}$$

and then find the estimates that maximise the marginal likelihood $\mathcal{L}\left(\boldsymbol{\beta}, \sigma_b^2, \sigma_e^2\right) = p\left(\mathbf{y} \mid \boldsymbol{\beta}, \sigma_b^2, \sigma_e^2\right)$. Estimates are usually found by REML instead of MLE to avoid bias in the estimation of the variance components σ_b^2 and σ_e^2. In contrast to the simple linear model (section 1.7.1), the REML of parameters in linear mixed models cannot be solved in closed-form. Different methods for the efficient estimation of the model parameters have been proposed e.g. [Lippert & al., 2011], but will not be described in detail here. In this thesis, the LMM framework LIMIX and accompanying methods (mtSet) were used to build the association models. Within this framework, the REML of the model parameters are found via Broyden's method [Broyden, 1965]. Details of the implementation can be found in [Casale & al., 2015, Supplementary material].

Linear mixed models in genetic association studies

In genetic association studies, LMMs describe the trait of interest as the sum of genetic fixed and random effects, i.e. single variants and background genetic effects, possible additional covariates and residual noise:

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{F}\boldsymbol{\alpha} + \mathbf{g} + \boldsymbol{\psi} \text{ with } \mathbf{g} \sim \mathcal{N}\left(0,\ \sigma_g^2 \mathbf{R}\right) \text{ and } \boldsymbol{\psi} \sim \mathcal{N}\left(0,\ \sigma_e^2 \mathbf{I}_N\right). \tag{1.30}$$

with
- the phenotype vector $\mathbf{y} \in \mathcal{R}^{N,\,1}$,
- the genetic profile of the SNP being tested $\mathbf{x} \in \mathcal{R}^{N,\,1}$,
- the effect size of the SNP $\beta \in \mathcal{R}^{1,\,1}$,
- the matrix of K covariates $\mathbf{F} \in \mathcal{R}^{N,\,K}$,
- the effect of covariates $\boldsymbol{\alpha} \in \mathcal{R}^{K,\,1}$ and
- the genetic relatedeness matrix $\mathbf{R} \in \mathcal{R}^{N,\,N}$.

In analogy to equation (1.29), the random effect $\mathbf{g}$ can be marginalised out, leading to the likelihood expression for equation (1.30) as

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{F}\boldsymbol{\alpha} + \mathbf{x}\beta,\ \sigma_g^2\mathbf{R} + \sigma_e^2\mathbf{I}_N\right). \tag{1.31}$$

In equation (1.31), the genetic covariance structure of the samples, as expressed by the genetic relatedness matrix $\mathbf{R}$, is integrated in the overall covariance structure of the model. As discussed in the next section, the covariance structure introduced by $\mathbf{R}$ captures population structure and polygenic background and leads to well-behaved statistics under the null model [Yu & al., 2006; Kang & al., 2008].

Estimating the kinship between samples

Traditionally, LMMs have been widely used in association studies with pedigrees of known relationship [Eu-Ahsunthornwattana & al., 2014]. The pedigree relationship between two individuals was used to estimate their predicted proportion of the genome that is identical by descent (IBD). The concept of IBD is based on the random Mendelian sampling of chromosomes during successive meiosis from a common ancestor. As such, IBD as a measure is always relative to the founders in the pedigree. IBD estimates can also be generated for a population, where they have to be defined relative to some ancestral population or time point [Browning & Browning, 2010; Glazner & Thompson, 2012]. A matrix of pair-wise IBD estimates is then used as the genetic relatedness matrix $\mathbf{R}$ in the linear mixed model (equation (1.31)).

Alternatively, the genetic relatedness matrix can be estimated from genome-wide genetic marker information. Nejati-Javaremi & al. [1997] showed in simulations that if all loci contributing to a given trait were known, the accuracy of phenotype predictions based on the relatedness matrix estimated from those loci would be higher than for matrices estimated on pedigree information. Similarly, Villanueva & al. [2005] showed that the accuracy of breeding values from relationship matrices computed based on genetic markers is higher than for matrices derived from pedigree information. Extending these simulations, Hayes & al. [2009] showed that the increased prediction accuracy also holds when the relatedness matrix is estimated for a cohort of unknown pedigree using dense genetic markers instead of all true, but unknown causal loci. The use of such a realised relationship matrix (RRM) is now widely employed in GWAS of large cohorts and plant and animal breeding studies as it is able to capture small differences in the proportion of genetic markers that are shared between seemingly unrelated individuals [Lee & al., 2010; Lopes & al., 2013].

A common strategy for the estimation of the RRM, which is used in this thesis, is to compute the average allelic correlation matrix

$$\mathbf{R} = \frac{1}{S}\mathbf{X}\mathbf{X}^T \tag{1.32}$$

where S is the number of SNPs used for the estimation and $\mathbf{X}$ is the $N \times S$ matrix of standardised genotypes of the samples N [Patterson & al., 2006; Yang & al., 2011]. To derive the standardisation of the genotypes based on their allele frequency [Patterson & al., 2006; Yang & al., 2011; Casale & al., 2015], consider the bi-allelic genotype at the ith sample $\mathbf{x}_i$ in Hardy-Weinberg equilibrium i.e. with the allele frequencies of the alleles $p + q = 1$ and the genotype frequencies $p_i^2 + 2p_iq_i + q_i^2 = 1$. Here, p is defined as the reference allele and q as the alternative allele. In the additive genotype model (section 1.7.2), the genotypes can be described in terms of allele dosage d, with $d(p_i, p_i) = 0$, $d(p_i, q_i) = 1$ and $d(q_i, q_i) = 2$. Based on allele dosages, the expected value of the genotype is defined as

$$\begin{aligned} E(x_i) &= d(p_i, p_i) \times p_i^2 + d(p_i, q_i) \times 2p_iq_i + d(q_i, q_i) \times q_i^2 && (1.33)\\ &= 2p_iq_i + 2q_i^2 = 2(1 - q_i)q_i + 2q_i^2 = 2q_i. && (1.34) \end{aligned}$$

With the expected value of the genotype, its variance and standard deviation can be computed

$$\begin{aligned} Var(x_i) &= E(x_i^2) - E(x_i)^2 && (1.35)\\ &= d(p_i, p_i)^2 \times p_i^2 + d(p_i, q_i)^2 \times 2p_iq_i + d(q_i, q_i)^2 \times q_i^2 - (2q_i)^2 && (1.36)\\ &= 2q_i(1 - q_i) && (1.37)\\ \sigma(x_i) &= \sqrt{Var(x_i)} = \sqrt{2q_i(1 - q_i)} && (1.38) \end{aligned}$$

and the genotypes standardised as

$$\bar{x}_i = \frac{x_i - 2q}{\sqrt{2q(1 - q)}}. \tag{1.39}$$

Different strategies have been proposed for the selection of genetic markers in the RRM estimation, including a two-stepped analysis approach allowing for preselection of phenotype-specific variants [Lippert & al., 2013] and grouping SNPs by haplotype [Zhao & al., 2007; Kang & al., 2008]. The latter avoids the bias introduced by the potentially unequal number of experimentally genotyped/imputed SNPs per

haplotype [Speed & al., 2017]. Similarly, choosing only SNPs which are in approximate linkage equilibrium can avoid this bias [Browning, 2008]. As described in [Eu-Ahsunthornwattana & al., 2014], SNPs in approximate linkage equilibrium can be found by strict LD pruning in genomic windows of appropriate size (depending on the organism and study design). Throughout this thesis, RRM estimates are always based on LD-pruned SNP sets.

1.7.7. Joint analysis of multiple phenotypes

Many cohort studies today, ranging from studies in model organism such as yeast and *Arabidopsis thaliana* to human, have rich, high-dimensional datasets including molecular, morphological or imaging derived traits [Bloom & al., 2013; Atwell & al., 2010; Astle & Balding, 2009; Shaffer & al., 2016; Stein & al., 2010]. However, these traits have often been analysed separately, partly for simplicity and partly because of a paucity of models suitable for the analysis of high-dimensional phenotype data. A variety of multi-trait models have been developed which can be broadly grouped into three different classes: i) dimensionality reduction techniques, ii) meta-analysis approaches and iii) multivariate regression models (reviewed in [Shriner, 2012; Yang & Wang, 2012]).

Dimensionality reduction techniques Dimensionality reduction methods in genotype-phenotype mapping seek to find a suitable projection of high-dimensional phenotypes into a lower dimensional space. Two commonly used dimensionality reduction methods are PCA and canonical correlation analysis (CCA). An overview of other methods and a more detailed description of methods in this section will be given in chapter 6.

In PCA , the phenotype data is projected into its principal components - the eigenvectors of the empirical covariance matrix. The amount of variance that each component explains is proportional to its corresponding eigenvalue. The dimensionality reduction is achieved by using all those principal components (in increasing order) until the cumulative sum of the eigenvalues reaches a predefined threshold of total phenotypic variance that should be retained. PCA as a dimensionality reduction technique has for instance been used in studies to find links between genotypes and facial features or obesity phenotypes [Liu & al., 2012; Claes & al., 2014; He & al., 2008]. Recently, Aschard and colleagues [Aschard & al., 2014] demonstrated that simply focusing on the principal components with the highest variance might not exploit the full potential of using PCA for genetic association. They propose a model

of combined PCA where the PCs are grouped based on the level of variance they explain and show a power gain in detecting genetic associations using this approach.

While the PCA dimensionality reduction approach focuses on the phenotype space and subsequent association with the genotypes, CCA seeks to maximise the canonical (ordered) correlation between the transformed phenotypes and genotypes i.e. finding the optimal linear transformation of the phenotypes while simultaneously testing for the association with the genotypes. For a single genetic marker, CCA finds the linear phenotype transformation that explains the maximum amount of covariance between this genotype and all traits by solving the eigendecomposition of a complex phenotype-genotype covariance term [Yang & Wang, 2012]. Ferreira and Purcell [2009] showed in simulations that CCA with multiple traits and one genetic marker controls well for type I errors and has increased power compared to univariate tests. In order to extend CCA to more than one marker, the genotypes also have to undergo a linear transformation and the maximum canonical correlation is found by solving two eigenvalue problems. As the number of genotype markers in GWAS exceeds the number of samples, estimates of the genotype covariance term becomes unreliable [Schäfer & Strimmer, 2005]. Several methods have been developed to circumvent this issue, making use of sparse matrices [Parkhomenko & al., 2009] or *a priori* grouping of the genotypes [Naylor & al., 2010].

Meta-analysis approaches Meta-analysis approaches combine the simplicity of the univariate approaches with the advantages of the multivariate approach. For each phenotype, a univariate association study is conducted and the summary statistics of these tests are combined. Many methods for combining the summary statistics [Xu & al., 2003; Yang & al., 2010; Bolormaa & al., 2014] go back to the work by O'Brien [O'Brien, 1984], who proposed to use a linear combination of the observed test statistics for each univariate test as the new statistics to be evaluated for significance. Subsequent studies proposed different methods for choosing the weights in the linear combination of the univariate test statistic or keeping the same principle computation but re-formulating the alternative hypotheses [Yang & al., 2010; Xu & al., 2003]. These studies showed an increase in power for applying the combined statistic on small marker sets or numbers of phenotypic traits. Bolormaa & al. [2014] showed that the power gains also hold for genotype to phenotype mapping of 32 traits across all genome-wide markers.

Regression models There are a number of different regression models that allow for the multivariate analysis of phenotypes. Among them are graphical models, generalised estimation equations and frailty models, for which a summary of methods and application can be found in [Shriner, 2012; Yang & Wang, 2012]. Here, I will focus on describing the development of multivariate linear regression models for genotype-phenotype mapping.

Before the era of GWAS, Jiang and colleagues [1995] proposed a multi-trait model where the phenotypes are jointly modelled as the sum of the fixed genetic effects of interest, fixed effects for genetic background variation and residual noise. They show that the joint analysis of correlated traits can increase power to detect the underlying genetics and can increase the precision of the parameter estimates. The significance of the association is determined via a likelihood ratio test of the parameter estimates under the null model, where the fixed genetic effect is zero, and the parameter estimates under the alternative model. The alternative model design depends on the underlying biological hypothesis regarding the effect of the genetic variant. Here, Jiang and colleagues differentiate between hypotheses for a simple joint mapping of phenotypes, pleiotropy and gene-environment interactions.

Methods developed thereafter often use the same underlying hypotheses for the mapping, but different techniques for the evaluation of the significance. For instance, two other groups developed methods for the joint analysis of traits based specifically on the residual sum of squares (RSS) matrix of the standard linear model estimated at each locus tested [Knott & Haley, 2000; Korol & al., 2001]. In the model proposed by Knott and Haley, the different properties and descriptors of the RSS are used to determine the significance of the association. To test for pleiotropy for instance, the determinant of the RSS at the test locus is compared to the RSS of the null model of no association. In contrast, Korol and colleagues propose to use the RSS of the multi-trait mapping as a means for trait transformation and dimensionality reduction. The resulting one-dimensional trait per sample is fitted in a single-trait test for significance testing.

While methods described so far have only used fixed genetic effects, Korte and colleagues [2012] were the first to introduce a random genetic effect into the model. Based on the original model by Jiang, they substituted the fixed effect accounting for background genetics by a random effect, turning the multivariate linear model into a multivariate linear mixed model. Since this initial model for multi-trait testing, a number of publicly available linear mixed model frameworks for the genome-wide mapping of a moderate number of traits have been developed [Yang & al., 2014;

Lippert & al., 2014; Zhou & Stephens, 2014; Casale & al., 2015].

Out of the different approaches described above, multivariate linear mixed models have the additional advantage that they can control for complex relatedness and population structure (section 1.7.5).

1.7.8. Linear mixed models for the joint analysis of multiple phenotypes

The multivariate linear mixed model with $P = \{1, 2, ..., p\}$ phenotypes for N samples can be derived as an extension of the univariate model with $P = 1$ phenotype for N samples described in equation (1.31). Consider equation (1.31) as the model description for the ith phenotype (omitting covariates for simplicity) :

$$\mathbf{y}_i \sim \mathcal{N}\left(\mathbf{x}\beta_i\,,\ \sigma^2_{g_i}\mathbf{R} + \sigma^2_{n_i}\mathbf{I}_N\right), \tag{1.40}$$

with $\mathbf{x} \in \mathcal{R}^{N,\,1}$ the genotype, β_i the effect size of the genotype for trait i, $\sigma^2_{g_i}$ and $\sigma^2_{n_i}$ the covariance terms of the genetic and noise random effect for trait i, $\mathbf{R} \in \mathcal{R}^{N,\,N}$ the realised relationship matrix estimated from the genotype data and $\mathbf{I}_N$ the identity matrix. As described by Henderson & Quaas [1976], multivariate LMMs model the covariance between trait i and j as

$$\text{Cov}\left(\mathbf{y}_i, \mathbf{y}_j\right) = \rho_{g_{ij}}\sigma^2_{g_i}\sigma^2_{g_j}\mathbf{R} + \rho_{n_{ij}}\sigma^2_{n_i}\sigma^2_{n_j}\mathbf{I}_N, \tag{1.41}$$

with $\rho_{g_{ij}}$ and $\rho_{n_{ij}}$ the genetic and noise correlation between trait i and j, respectively. Using the multivariate LMM described for trait i in equation (1.40) and the expression of the ij-trait-trait covariance term in equation (1.41), the multivariate LMM for all traits P can be expressed as a matrix-normal distribution:

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta}^T + \mathbf{G} + \boldsymbol{\psi} \tag{1.42}$$

with the phenotype matrix $\mathbf{Y}$ and the effect size vector $\boldsymbol{\beta}$

$$\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 & \cdots & \mathbf{y}_P\end{bmatrix} \in \mathcal{R}^{N,\,P}, \tag{1.43}$$

$$\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_1 & \cdots & \boldsymbol{\beta}_P\end{bmatrix} \in \mathcal{R}^{P,\,1}, \tag{1.44}$$

$$\tag{1.45}$$

the random genetic effect $\mathbf{G}$ and the random noise effect $\boldsymbol{\psi}$ following a matrix-variate normal distribution with row covariance $\mathbf{R}$ and $\mathbf{I}_N$ and column covariance $\mathbf{C}_g$ and

$\mathbf{C}_n$

$$\mathbf{G} = \mathcal{MN}_{N,P}\left(\mathbf{0}\,,\ \mathbf{R}\,,\ \mathbf{C_g}\right), \tag{1.46}$$

$$\boldsymbol{\psi} = \mathcal{MN}_{N,P}\left(\mathbf{0}\,,\ \mathbf{I_N}\,,\ \mathbf{C_n}\right), \tag{1.47}$$

and the genetic and noise trait-by-trait covariance matrices $\mathbf{C}_g$ and $\mathbf{C}_n$

$$\mathbf{C}_g = \begin{bmatrix} \sigma^2_{g_1} & \rho_{g_{12}}\sigma_{g_1}\sigma_{g_2} & \cdots & \rho_{g_{1P}}\sigma_{g_1}\sigma_{g_P} \\ \rho_{g_{12}}\sigma_{g_1}\sigma_{g_2} & \sigma^2_{g_2} & \cdots & \rho_{g_{2P}}\sigma_{g_2}\sigma_{g_P} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{g_{1P}}\sigma_{g_1}\sigma_{g_P} & \rho_{g_{2P}}\sigma_{g_2}\sigma_{g_P} & \cdots & \sigma^2_{g_P} \end{bmatrix}, \tag{1.48}$$

$$\tag{1.49}$$

$$\mathbf{C}_g = \begin{bmatrix} \sigma^2_{n_1} & \rho_{n_{12}}\sigma_{n_1}\sigma_{n_2} & \cdots & \rho_{n_{1P}}\sigma_{n_1}\sigma_{n_P} \\ \rho_{n_{12}}\sigma_{n_1}\sigma_{n_2} & \sigma^2_{n_2} & \cdots & \rho_{n_{2P}}\sigma_{n_2}\sigma_{n_P} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n_{1P}}\sigma_{n_1}\sigma_{n_P} & \rho_{n_{2P}}\sigma_{n_2}\sigma_{n_P} & \cdots & \sigma^2_{n_P} \end{bmatrix}. \tag{1.50}$$

The matrix-variate distribution of the phenotype matrix $\mathbf{Y}$ can be expressed in terms of a multivariate normal distribution (for details refer to equation (C.1) to equation (C.8) in the appendix)

$$\text{vec}\left(\mathbf{Y}\right) \sim \mathcal{N}_{N\times P}\left(\text{vec}\left(\mathbf{x}\boldsymbol{\beta}^{\mathbf{T}}\right),\ \mathbf{C}_g \otimes \mathbf{R} + \mathbf{C}_n \otimes \mathbf{I}_N\right). \tag{1.51}$$

where the Kronecker products $\otimes$ of $\mathbf{C}_g \otimes \mathbf{R}$ and $\mathbf{C}_n \otimes \mathbf{I}_N$ follow the definition of the Kronecker product for any two matrices as:

$$\mathbf{C}_g \otimes R = \begin{bmatrix} \mathbf{C}_{g_{11}}\mathbf{R} & \cdots & \mathbf{C}_{g_{1P}}\mathbf{R} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{g_{P1}}\mathbf{R} & \cdots & \mathbf{C}_{g_{PP}}\mathbf{R} \end{bmatrix} \text{ and } \mathbf{C}_n \otimes \mathbf{I}_N \ = \begin{bmatrix} \mathbf{C}_{n_{11}}\mathbf{I}_N & \cdots & \mathbf{C}_{n_{1P}}\mathbf{I}_N \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{n_{P1}}\mathbf{I}_N & \cdots & \mathbf{C}_{n_{PP}}\mathbf{I}_N \end{bmatrix}.$$

The likelihood of the multivariate linear mixed model is

$$\mathcal{L}\left(\boldsymbol{\beta}^T, \mathbf{C}_g, \mathbf{C}_n\right) = \mathcal{N}\left(\text{vec}\left(\mathbf{Y}\right) \mid \text{vec}\left(\mathbf{x}\boldsymbol{\beta}\right),\ \mathbf{C}_g \otimes \mathbf{R} + \mathbf{C}_n \otimes \mathbf{I}_N\right). \tag{1.52}$$

Maximising $\mathcal{L}\left(\boldsymbol{\beta}, \mathbf{C}_g, \mathbf{C}_n\right)$ requires P parameter estimates for the fixed effect $\boldsymbol{\beta}$ and $\frac{1}{2}P(P+1)$ parameter estimates for both of the $P \times P$ covariance matrices $\mathbf{C}_g$ and $\mathbf{C}_n$. Due to the large number of parameters, REML for multivariate LMM often

relies on gradient-based optimisation methods. Different schemes have been used in LMM for genetic analysis, including average information REML [Gilmour & al., 1995] (used in [Yang & al., 2011]) quasi-Newton methods like Broyden's method [Broyden, 1965] (used in [Casale & al., 2015]), and Brent's algorithm [Brent, 1971] (used in [Lippert & al., 2011; Svishcheva & al., 2012]). The REML implementation of the framework used in this thesis builds on Broyden's method and the detailed derivation can be found in [Casale & al., 2015, Supplementary material]. Commonly used multi-trait association frameworks and their implementation are discussed in detail in the introduction of chapter 4.

Hypothesis testing in multi-trait association studies

As described by Jiang & Zeng [1995] and Korte & al. [2012] (summarised in section 1.7.7, regression models), when testing the association of a genetic marked across multiple phenotypes, different hypotheses for the underlying genetic trait architecture can be formulated. In the most simple case, one can test if the genetic variant has an effect on any of the traits P (any effect test) i.e. the effect size of the fixed effect $\boldsymbol{\beta}$ is unequal to zero for at least one trait : $H_{\mathrm{A}} : \boldsymbol{\beta} \neq \mathbf{0}_P$. In this P-degrees of freedom test, the corresponding null hypothesis of no association is that the effect size of the fixed effect is equal to zero: $H_0 : \boldsymbol{\beta} = \mathbf{0}_P$. In the common effect model, the variant has the same effect size across all traits with $H_{\mathrm{A}} : \boldsymbol{\beta} = \mathbf{1}_P\beta$ and is tested for significance in a one degree of freedom model versus the null hypothesis of no association ($\beta = 0$). A more complicated model allows to test for specific effects of the variant on a given trait p. This can be tested with a one degree of freedom test where a model containing a common effect across all traits and a specific effect for trait p is compared against the common effect model.

2

Cardiac biology

In chapter 7 and chapter 8, I investigate genetic associations of human heart morphology. To aid with an understanding of the relevant biology and key terms, I use this chapter to give a basic overview of human heart morphology, cardiovascular diseases and their underlying genetics.

The human heart is composed of four chambers, the left and right ventricle and the left and right atrium. On the outside it is covered by a tough membranous structure, the pericardium. The innermost layer of the pericardium, the epicardium, is fused to the heart and forms part of the heart wall. It directly connects to the myocardium, the thickest layer of the heart wall which is composed of conductory and contractile cardiomyocytes. On the inside, the myocardium is lined by the endocardium [Betts & al., 2013].

The four chambers of the heart (figure 2.1) are separated through two septal structures, the interventricular and the atrioventricular septum. The blood exchange between the atria and ventricles is enabled through a set of valves embedded in the atrioventricular septum: the mitral valve between left atrium, and ventricle and the tricupsid valve between the right atrium and ventricle. In addition, each ventricle has a valve at its exit point. In the right ventricle, the pulmonary valve separates the ventricle from the pulmonary artery. Similarly, the aortic valve separates the left ventricle from the aorta. There is no direct blood exchange between the left and the

right side of the heart in healthy individuals.

Diseases of the heart are common and one of the leading health issues worldwide. They include a wide range of disorders from atherosclerosis, diseases of the myocardium and the heart's electrical circuit to congenital heart diseases. To help with an understanding of these disease pathologies, the circulatory and conductory system as well as the development of the heart are described below.

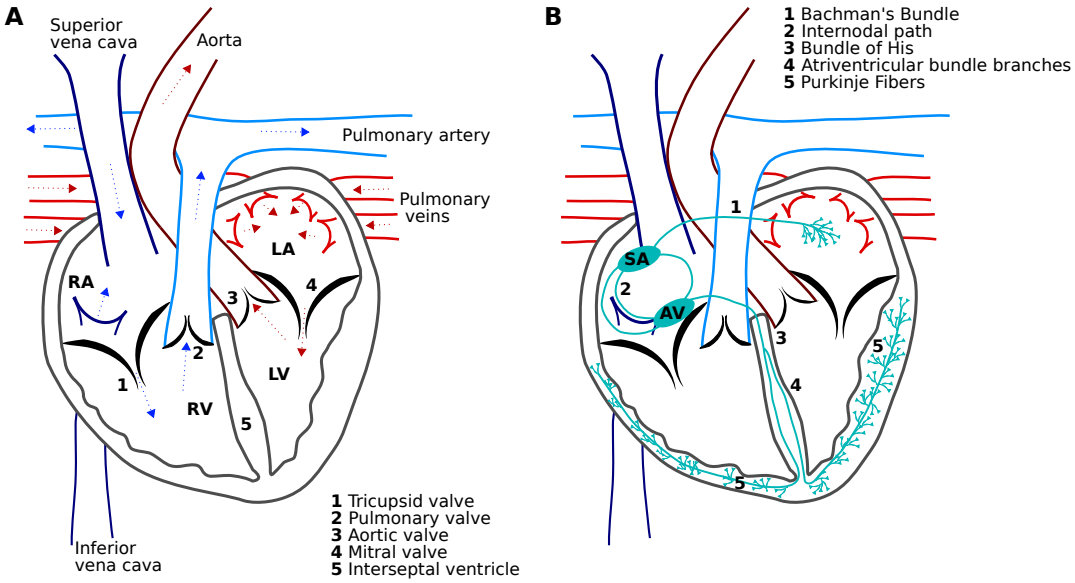

Figure 2.1: Anatomy, circulatory and conductory system of the human heart. A. Circulatory system. Deoxygenated blood (blue arrows) arrives at the right atrium (RA) from the systemic circulation. From the right atrium, it enters the right ventricle (RV) through the tricupsid valve. It leaves the right ventricle through the pulmonary valve into the pulmonary artery entering the pulmonary circuit. Oxygenated in the lung, blood (red arrows) arrives back at the heart at the left atrium (LA) through two branches of the vena cava and enters the passing the mitral valve. It leaves the left ventricle (LV) through the aorta, entering the systemic circulation. Anatomy: The myocardium of the left ventricle is significantly thicker than the right ventricle, as it has to overcome greater pressure of the systemic circuit. The walls of the atria are smooth, whereas the ventricles show protrusions. The atrioventricular septum separating atria and ventricles is not shown for simplicity. It is located at the level of the tricupsid and mitral valves. B. Conductory system. The sinoatrial (SA) node initiates the contraction of the heart by sending an action potential through the atria via cell-cell contact and specific pathways (Bachmann's Bundle and internodal paths). The potential arrives at the atrioventricular (AV) node, where it is delayed to allow for full contraction of the atria before it is passed on to the Purkinje Fibers, through the Bundle of His and the atrioventricular bundle branches. The Purkinje Fibers pass the signal on to the ventricles, leading to their contraction and the pumping of the blood outside of the heart.

2.1. Cardiac cycle

The cardiac cycle begins with the contraction of the atria and ends with the relaxation of the ventricles. During the cycle, the chambers of the heart can be found in two distinct states, systole and diastole. In systole, the chambers contract and pump blood into either the ventricles (atria) or out of the heart (ventricle). In diastole, the chambers are relaxed and fill with blood. Both atria and ventricle cycle through these states, coordinated by impulses sent from the circulatory system. Ventricles are in diastole when atria undergo systole and vice versa. In atrial diastole, the valves separating atria and ventricles are open and facilitate passive blood flow into the ventricles. When the cardiac cycle starts, atria enter systole and pump the remaining blood into ventricles. The amount of blood contained in the ventricles at the end of atrial systole/ventricular diastole is referred to as end diastolic volume. When the ventricle enter systole, the pressure in the ventricles rise compared to the one in the atria which are in diastole and the separating valves are closed as a response to the increased pressure. Once the ventricular pressure overcomes the pressure in aorta and pulmonary arteries, the respective valves open and equivalent amounts of blood are pumped into the systemic and pulmonary cycle. The larger and higher resistance vessels of the systemic circulation compared to the low-pressure vessels of the pulmonary system put a higher demand on the left ventricle which is met by a proportionally higher mass of the left ventricle compared to the right. The amount of blood that each ventricle can pump within one cardiac cycle is defined as the stroke volume. The volume of blood remaining in the ventricle at the end of systole is referred to as end systolic volume. End diastolic volume, stroke volume, end systolic volume are important clinical parameters [Betts & al., 2013].

2.2. Conduction system

The conduction system of the heart establishes the heart rhythm through electrical impulses sent by specialised myocardial conducting cells. The normal cardiac rhythm, called sinus rhythm, is established by the sinoatrial node and is located at the junction of the superior vena cava and the right atrium (figure 2.1B). The sinus node is also called the pacemaker of the heart, since the signal leading to the activation of the myocardial contractile cells and, in consequence, their contraction starts here. Upon initiation of the action potential in the sinus node, the depolarisation spreads through the atria to the atrioventricular node via cell-cell contacts, the internodel

pathways and Bachmann's bundle [Laske & Iaizzo, 2005; Anderson & al., 2009]. The atrioventricular node is located within the atrioventricular septum which prevents the signal to spread directly to the ventricle without being processed. At the atrioventricular node, the signal is delayed to allow the atria to complete their contraction which pumps the blood into the ventricles. From the atrioventricular node, the signal is propagated along the interventricular septum through the bundle of His which divides into the atrioventricular bundle branches. These in turn connect with Purkinje Fibers at the apex of heart, which propagate the impulse to the myocardial contractile cells in the ventricles. The contraction of the ventricles follows the direction of the impulse and travels from the apex towards the base, pumping blood out of the ventricles and into the aorta and pulmonary arteries [Laske & Iaizzo, 2005; Sigg & al., 2010].

2.3. Heart development

The heart is the first functional embryonic organ and already starts to beat by the end of the third week of development [Zambrano & al., 2002]. In the developing heart, three major processes have to be orchestrated: the formation and arrangement of the myocardium into the four-chamber heart, the development of the conduction system, and the heart's circulatory system required for nutrition and oxygen supply to the myocardium. The first two processes happen simultaneously, while the latter can only take place after proper development of the myocardium.

The development of the heart starts in the third week of development, just after gastrulation. In gastrulation, the single-layered sheet of epithelial cells that forms the embryo is re-organised into three germ layers, the ectoderm (external layer), mesoderm (middle layer) and endoderm (internal layer). Each layer will give rise to different tissues and organs in the developing embryo. The heart development begins with the formation of two cardiac crescents from the mesodermal layer (figure 2.2, 1), which are located near the head of the embryo [Christoffels & al., 2000]. Within each cardiac crescent, two structures develop, a plate of myocardial cells and a plexus of endothelial strands. These develop into cardiogenic cords, with the endothelial strands forming a tube structure enveloped by a layer of myocardial cells. By the fusion of the two cardiogenic cords, the early tubular heart is formed (figure 2.2, 2). This early tubular structure already shows peristaltic contraction, despite the lack of valves and conduction system [Goss, 1938; de Jong & al., 1992; Moorman & Lamers, 1994]. The tubular heart then undergoes a right-ward looping where an initial dif-

ferentiation into ventricular myocard, atrial myocard and transitional zones occurs (figure 2.2, 3). The transitional zones will form parts of the septa, valves, conduction system and fibrous heart skeleton [Gittenberger-de Groot & al., 2005]. Through the looping of the heart an inner and an outer curvature is created. The developing atria and ventricle stand out on the outer curvature, whereas transitional zones are brought into proximity on the inner curvature (figure 2.2, 4).

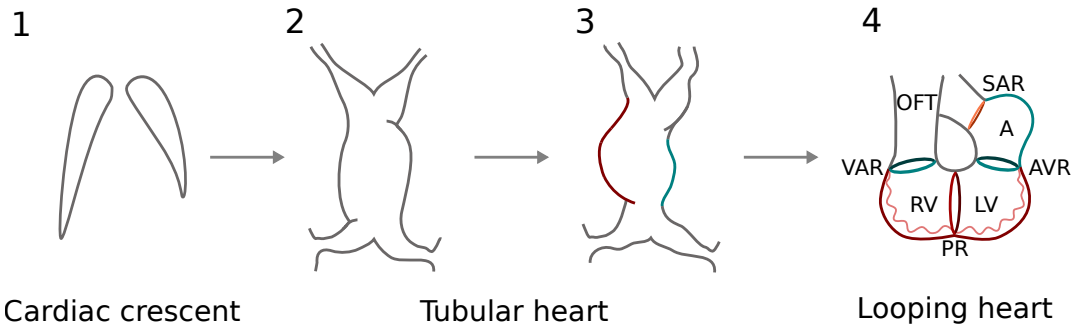

Figure 2.2: Embryonic heart development. 1. The mesoderm gives rise to two cardiac crescents that already show some extend of asymmetry. 2. The cardiac crescent have fused together to from a straight heart tube. 3. The straight heart tube starts a right-ward looping. Parts marked in red will develop into the ventricles, while parts marked in turquoise will become atria. 4. The looping heart with precursors of the atria (A), the left ventricle (LV), the right ventricle (RV) and the outflow tract (OFT). Ring-like structures mark the transitional zones: sinoatrial ring (SAR), atrioventricular ring (AVR), primary ring (PR), ventriculararterial ring (VAR).

Correct looping and positioning ot the transitional zones are critical for the separation of the heart into its functional components. The separation is facilitated through septation at the atria, the ventricles and the arterial pole. For the separation of the ventricles, two processes have to be considered, the inflow and outflow septation. The inflow septation i.e. the septation of the ventricles from one another and from the atria, is mainly achieved through the primary ring. The primary ring gives rise to the ventricular septum that separates left and right ventricle. This process has to be orchestrated with the position of atrioventricular ring, which is pulled towards the right ventricle by a tightening of the inner curvature. The positioning of the atrioventricular ring above the left and right ventricle builds the base for the formation of the mitral and tricupsid valve, respectively, which will separate the atria from the ventricles. The septation controlling the blood flow from ventricles to the arteries (outflow septation) is achieved through the twisting of the ventriculararterial ring into the precursors of the pulmonic and aortic valve and their positioning above the right and left ventricle.

At the end of week nine in development, the heart consists of the four chambers divided by septa with integrated valves. Morphologically, atria and ventricle can be distinguished based on the structure of their myocard. While the myocardium of the atria is thin and has a smooth surface, the ventricles show a much thicker myocardium with protrusions (trabeculations) running along the endocardial surface.

During these rearrangement processes the myocardium also underwent a differentiation into the contracting and conducting myocardium. While many components of the gene regulatory networks that control the differentiation are known today, mechanisms involved in controlling this differentiation on a cellular and region-specific level remain to be discovered [Christoffels & Moorman, 2009; Paige & al., 2015; Park & Fishman, 2017]. Structures important in the development of the conduction system are the sinoatrial ring which will develop into the sinoatrial node, the primary ring which will give rise to the atrioventricular conduction system and the atrioventricular ring developing into Bachmann's Bundles.

2.4. Common cardiovascular diseases

According to the International Statistical Classification of Diseases and Related Health Problems, the classification system of the world health organisation, total cardiovascular diseases include hypertension, hypercholesterolemia, coronary heart disease, cardiac arrhythmias, congenital heart diseases and cardiomyopathies (classification codes I00-I99, Q20-28, version ICD-10 [World Health Organisation, 2016]).

The largest contribution to cardiovascular diseases are coronary heart diseases. Their major clinical manifestations are myocardial infarction (commonly known as heart attack), angina pectoris (chest pain), and sudden coronary death [Wong, 2014]. The common cause of coronary heart diseases is an interrupted blood and consequently oxygen supply to the heart through a blockage of the coronary arteries. Major risk factors are high blood pressure (hypertension) and high blood cholesterol (hypercholesterolemia) [Mackay & al., 2004].

Cardiac arrhythmias are a class of diseases where the observed cardiac rhythm is different from the regular sinus rhythm. They are caused by irregularities of impulse generation and/or conduction. Tachycardia is the condition of an increased heart rate whereas bradicardia describes a lower than normal heart rate. They can cause a reduction in cardiac output and myocardial blood flow and may be life-threatening [Durham & Worthley, 2002].

Congenital heart diseases are diseases with structural abnormalities of the heart

or intrathoraic great vessels that are of functional significance and have been present since birth [Mitchell & al., 1971]. They may be caused by genetic or environmental factors during pregnancy and include ventricular outflow tract obstructions i.e. narrow or blocked arteries and valves and septal defects. Of the latter, interventricular septal defects are the most common [Hoffman, 2005].

Cardiomyopathies describe a class of diseases where the heart muscle fails to function properly. Traditionally, they are classified based on their anatomy and hemodynamics into hypertrophic, dilated, or restrictive cardiomyopathy. The incidence of the latter is rare and no changes in ventricular morphology are observed. This is in stark contrast to hypertrophic and dilated forms, where an increase in ventricular wall thickness or volume are observed, respectively. The increase in wall thickness is caused by a hypertrophy of existing myocytes rather than a hyperplasy as in the developing heart [Lorell & Carabello, 2000]. Dilated cardiomyopathy presents with an increase in cardiac chamber volume and often a modest increase in wall thickness. Both mechanism are in response to cardiac stress and initially improve heart function but in the long run increase myocardial strain and raise metabolic demands [Seidman & Seidman, 2001].

Cardiovascular diseases are caused by a combination of environmental and genetic risk factors. Amongst the environmental risk factors one can distinguish between modifiable risks governed by the individual itself and exposure to risk factors which are often beyond the influence of the individual. The latter include exposure to solvents, pesticides or extremes in noise and temperature [Bhatnagar, 2004; Brook & al., 2010; Babisch, 2014]. Modifiable risk behaviour such as smoking, physical inactivity and a poor diet have been shown to be highly correlated with the incidence of cardiovascular diseases (reviewed in [O'Toole & al., 2008; Cosselman & al., 2015]). Meta-studies examining behavioural change in the English,Welsh and American populations over a period of 20 years, have shown a decline in coronary heart disease mortality due to a reduction in smoking, increased physical activity and other behavioural factors [Unal & al., 2004; Ford & al., 2007]. Genetic risk factors for cardiovascular diseases are described in the next section.

2.5. Genetics of cardiovascular diseases

The genetics of cardiovascular diseases point both to simple Mendelian and complex inheritance patterns. In multiple linkage analyses studies of familial myocardial hypertrophy, several genes have been discovered where mutations segregate in a Men-

delian fashion. These include mutations in *cardiac myosin heavy chain* [Geisterfer-Lowrance & al., 1990], α *tropomyosin*, *cardiac troponin T* and *C*, [Thierfelder & al., 1994; Kimura & al., 1997] and *cardiac mysosin binding protein* [Carrier & al., 1993; Bonne & al., 1995]. Another group of familial cardiovascular diseases, familial hypertension, has been linked to mutations in epithelial sodium channels *SCNN-2* and *SCNN3-3* [Boyden & al., 2012; Glover & al., 2014] as well as *KLH3-CUL3*, genes coding for proteins building a complex involved in Sodium-chloride reabsorption in the kidney [Hansson & al., 1995]. Linkage studies have also pinpointed genes for atrial and ventricular septal defects. They are linked to mutations in the transcription factors, *GATA4* [Schott & al., 1998] and *NKX2-5* [Garg & al., 2003], respectively.

In contrast, the majority of cardiovascular traits follow complex inheritance pattern with interaction between multiple genes and non-genetic factors [Kathiresan & Srivastava, 2012]. GWAS have been successful in finding genetic loci associated with a large number of cardiovascular diseases. Out of the 4,148 studies in the GWAS catalogue (accessed 11.08.2017), 159 contain phenotype descriptions relating to cardiovascular diseases (list of query terms in table A.1 in the appendix).

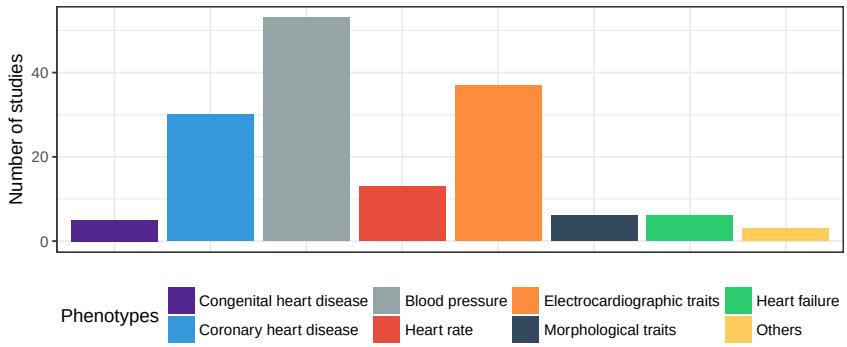

Figure 2.3: GWAS on heart-related phenotypes. Overview of 153 GWAS studies with 59 unique heart-related phenotypes (obtained from the GWAS catalogue [MacArthur & al., 2017, accessed on 11.08.2017]). Phenotypes were grouped into eight phenotype classes. The list of query terms and their grouping can be found in table A.1 in the appendix.

The highest number of studies has been conducted on blood pressure phenotypes, followed by electrocardiographic traits and coronary heart diseases (figure 2.3). Early GWAS on these traits were conducted on samples of the Framingham heart study, a community-based cohort study founded in 1948 to examine the epidemiology of cardiovascular disease [Dawber & al., 1951; Kannel & McGee, 1979]. The Framingham Heart Study 100K SNP genome-wide association study resource was published in

2007 [Cupples & al., 2007] and its 1,345 participants built the basis for 17 GWAS on traits like echochardiographic dimension [Vasan & al., 2007], blood pressure [Levy & al., 2007] and heart rate [Newton-Cheh & al., 2007]. Later studies often contained larger sample sizes or re-analysed previously published studies in meta-analysis. For instance, the international consortium for blood pressure conducted a meta-analysis of 29 previously published GWAS on systolic and diastolic blood pressure phenotypes and discovered 16 novel loci, ten of which were associated with known blood pressure-related genes [Ehret & al., 2011]. Similarly, the large consortium for coronary heart diseases (CARDIoGRAM) conducted a case-control meta-analysis and identified ten novel loci [Nikpay & al., 2015]. The other classes of phenotypes are smaller and more heterogeneous, comprising different congenital heart diseases e.g. congenital left-sided heart lesion [Mitchell & al., 2015; Hanchard & al., 2016] and conotruncal heart defects (i.e. malformations of the cardiac outflow tracts) [Agopian & al., 2014] or morphological traits including cardiomyopathies [Villard & al., 2011] and cardiac wall thickness [Vasan & al., 2009; Arnett & al., 2011].

2.6. Thesis outline

In the following chapters, I describe new methods and applications for the genetic analysis of high-dimensional datasets.

In chapter 3, I introduce the R package that I developed for the simulation of complex phenotype structures. Simulated phenotypes serve as an approximation for observed biological phenotypes and are invaluable for model development. All phenotypes simulated in this thesis are generated based on the strategies described in this chapter. The simulation strategy and applications have been published in [Meyer & Birney, 2018].

Chapter 4 presents LiMMBo, a new approach for finding genetic associations in high-dimensional phenotypes using linear mixed models. I first demonstrate model calibration and power on simulated datasets before I apply LiMMBo to a publicly available dataset of yeast growth traits in chapter 5. A manuscript of LiMMBo and its application is currently under revision and already available in pre-print [Meyer & al., 2018].

In chapter 6, I systematically analysed twelve unsupervised dimensionality reduction methods for their ability to find robust phenotype representations of simulated data with different structure and size. I introduce a new stability measure for choosing the low-dimensional representations and demonstrate that the selected repres-

entation can recover genetic associations.

Finally, I investigate genetic associations for human heart morphology based on magnetic resonance imaging (MRI) data of 1,500 healthy individuals. In chapter 7, I apply the methods and measures described in chapter 6 to obtain a low-dimensional representation of the heart morphology and conduct a GWAS based on this representation. Chapter 8 describes the GWAS on a cardiac trabeculation phenotype derived from a supervised feature extraction approach on the MRI data. The work in these chapters was done in collaboration with Antonio De Marvao, Jiashen Cai, Pawel Tokarczuk Declan O'Regan and Stuart Cook from Imperial College London. Specifically, phenotype acquisition and feature extraction was done by my collaborators, while I was responsible for all remaining analyses, including genotype quality control and imputation. An initial paper using the imputed genotypes was recently published [Biffi & al., 2017].

3
PhenotypeSimulator

For method development in quantitative genetics, one often needs a set of well-characterised genotypes and phenotypes to know the ground truth based on which comparisons of the model performance can be made. In the context of this thesis, genotype and phenotype simulations were crucial for the development of a new method for multi-trait mapping of high-dimensional datasets (chapter 4) and the evaluation of different dimensionality reduction techniques (chapter 6).

The complexity of the simulated phenotype components will depend on the specifics of the model that is being developed. With the detailed whole-genome genotype data available through standard techniques such as genotyping arrays and subsequent imputation and the measurement of multiple traits per sample, the complexity of the hypotheses for testing the underlying genetics of the observed phenotypes have increased. Models range from simple linear models with a few fixed effects on a single trait to complex linear mixed models with fixed and random effect components on multiple traits [Stephens, 2013; Marigorta & Gibson, 2014; Zhou & Stephens, 2014; Loh & al., 2014]. With the increase in analysis complexity, sophisticated approaches for modelling realistic genotype and phenotype structures are needed. These simulated genotypes and phenotypes reflect our perceived understanding of the true phenotype structure and do not guarantee the biologically correctness of real phenotypes. However, they are invaluable in model design, as any

model showing flawed statistics on the possibly simplified biological model will suffer from at least the same flaws on the true biological data.

In this chapter, I will first describe simulation strategies for genotypes with different levels of population structure and relatedness. Following that, I introduce the phenotype simulation strategy used for all simulated datasets within this thesis. In order to broadly distribute this simulation framework, I have developed *PhenotypeSimulator*, an R package for phenotype simulation that allows for a flexible and customisable simulation set-up. *PhenotypeSimulator* can be installed from the Comprehensive R Archive Network [Meyer, 2017] and its code is available on github: `https://github.com/HannahVMeyer/PhenotypeSimulator`. *PhenotypeSimulator* is published as: Meyer, H. & Birney E. (2018) PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships, *Bioinformatics*, bty197.

3.1. Genotype simulation

There are a number of different strategies to generate genotype data for genetic association studies. In the most simple case and assuming bi-allelic SNPs, each SNP is simulated from a binomial distribution with two trials and probability equal to the given allele frequencies (e.g in [Lippert & al., 2013]). This simple approach, however, does not simulate any dependency between the genotypes as is observed with LD structure in the genome. In order to mimic genomic LD structure and allele frequency distributions in the simulated dataset, three general approaches exist: i) backward-time or coalescent simulation, ii) forward time and iii) resampling approaches. The coalescent [Hudson, 2002; Ewing & Hermisson, 2010; Kelleher & al., 2016] and forward-time approaches [Peng & al., 2007; Hoggart & al., 2007; Carvajal-Rodríguez, 2008] use population genetic models to simulate genotypes and are particularly useful for studying evolution and demography. However, they often suffer from computational demands for diploid genome-wide SNP data [Liu & al., 2008; Yuan & al., 2012]. Resampling approaches [Wright & al., 2007; Su & al., 2011; Loh & al., 2014; Casale & al., 2015] offer a practical solution that can be used to efficiently generate genetic data with different relatedness and population structures, which is particularly useful in genetic association studies. They combine existing genotype data into the genotypes of the simulated samples, thereby retaining allele frequency and LD patterns.

I choose to follow the resampling strategies described in [Loh & al., 2014; Casale

& al., 2015] where each diploid individual is simulated as the mosaic of real genotypes from different populations. Depending on the simulation set-up, cohorts with differing levels of population structure and relatedness can be simulated. Cohorts with different degrees of genetic structure will be valuable for evaluating the performance of genetic association models with respect to their adjustment for genetic relatedness and population structure. As far as I am aware, these structures cannot be realised with the publicly available tools described in [Wright & al., 2007; Su & al., 2011].

I used the genotype data from 365 individuals of four European ancestry populations from the 1000 Genomes Project [1000 Genomes Project Consortium, 2015], Utah Residents (CEPH) with Northern and Western European Ancestry (CEU) and Finnish in Finland (FIN) and British in England and Scotland (GBR) and Toscani in Italia (TSI), as the sampling dataset. The resampling strategy works as follows:

1. each individual is randomly assigned a predefined number of unique original genotypes which will serve as its ancestors;

2. the ancestors' genome-wide genotypes are split into blocks of 1,000 SNPs;

3. for each SNP block, one of the ancestor is chosen at random and its genotype is copied to the individual's genome.

The number and the sub-population of ancestors that are chosen for simulating the genomes of a new cohort are critical for controlling the level of structure within the cohort.

The number of ancestors sets the level of relatedness within the cohort. Low numbers of N introduce relatedness among individuals, while high numbers of N lead to low levels of structure and relatedness. For instance, with $N = 2$, each individual in the newly synthesised cohort is composed of genotypes from only two out of the 365 individuals. Consider individual g_1, whose genotypes are drawn from ancestors a_1 and a_2. For Individual g_2, with a chance of $p = 1 - \frac{\binom{363}{2}}{\binom{365}{2}} \approx 0.01$ it shares at least one ancestor with g_1. For exactly one shared ancestor, each SNP block would have a 25% probability of being the same between g_1 and g_2 . With $N = 10$, the probability for at least one common ancestor increases ($p = 1 - \frac{\binom{355}{10}}{\binom{365}{10}} \approx 0.25$). However, for exactly one shared ancestor, the sharing of SNP blocks decreases to 1%.

The choice of sub-population determines the level of population structure in the simulated genotypes: allowing for random selection of ancestors independent from

the four subpopulations in the 1000 Genomes datasets yields low levels of population structure, as this leads to a random sampling of the individuals' genotypes across ancestors ethnicities. Including an *a priori* selection of individuals from one of the four sub-population and subsequently restricting ancestor selection to these individuals will restrict an individuals genotypes to a single sub-population. As all individuals in the cohort are now comprised of distinct genotype subsets, this will give rise to population structure in the simulated cohort.

I simulated three genotype sets, each with 1,000 samples, that differed i) in the number of ancestors N from which the genotypes were chosen and ii) the sub-populations the ancestors were chosen from:

A. unrelatedPopStructure: unrelated individuals with prior assignment of ancestral population ($N = 10$, i.e. only CEU or only FIN or only GBR or only TSI)

B. unrelatedNoPopStructure: unrelated individuals with mixed ancestral population ($N = 10$, i.e. CEU and FIN and GBR and TSI)

C. relatedNoPopStructure: related individuals with mixed ancestral population ($N = 2$, i.e. CEU and FIN and GBR and TSI))

The level of structure and relatedness introduced by this simulation strategy can be visualised by examining the genetic relationship matrix and the PCs of the genotypes. The genetic relationship matrix is estimated as a RRM via equation (1.32) and serves as a measure for relatedness between the individuals, while PCs reflect the genotypic variance in the data (section 1.7.5). The hierarchical clustering of the genetic relationship estimates and scatter plots of the first two PCs for each genotype set are shown in figure 3.1. Samples cluster tightly based on their ancestral subpopulations (figure 3.1A), while there is no clustering and an even spread in the PC plot for the cohort of unrelated individuals with ancestors sampled across all subpopulations (figure 3.1B). The cohort of related individuals shows less spread in the second principal component and higher individual genetic relationship estimates (figure 3.1C).

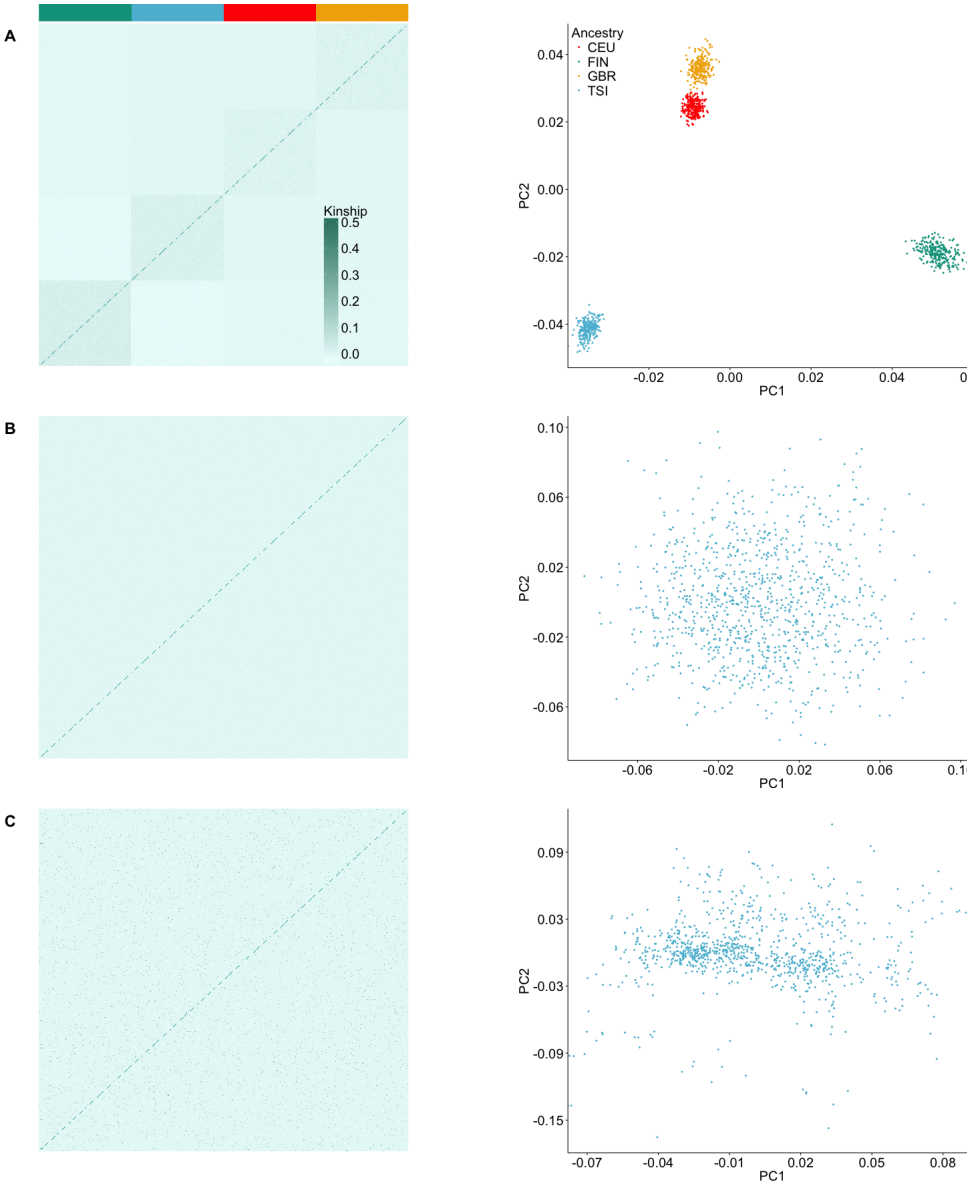

Figure 3.1: Genetic relationship matrices and principal components of three simulated European ancestry cohorts. The genotypes were simulated based on genotype data from four European ancestry populations (ancestry colour key in panel A). Depending on the choice and number of ancestors for the sampling of chromosomes to simulate an individual's genotype, cohorts with differing levels of population and relatedness structure will arise. The left column depicts the hierarchical clustering of the sample-by-sample genetic relationship coefficients (complete linkage clustering of Euclidean distance between coefficients), the right column the first and second PC of the sample genotypes for the three different cohorts: A. unrelated individuals, with population structure: $N = 10$, prior assignment to ancestral population. B. unrelated individuals, no population structure: $N = 10$, no prior assignment to ancestral population. C. related individuals, no population structure: $N = 2$, no prior assignment to ancestral population.

3.2. Phenotype simulation

In this section, I introduce *PhenotypeSimulator*, an R/CRAN package for the flexible simulation of phenotypes with different genetic and non-genetic variance components. *PhenotypeSimulator* is a framework focusing on the simulation of phenotypes, with a particular emphasis on complexity of both multiple phenotypes and multiple genetic loci and genetic background, which is not provided by other multi-phenotype simulation software ([O'Reilly & al., 2012], [Porter & O'Reilly, 2017]). I have written *PhenotypeSimulator* to be easily integrated with external genotype simulation software (such as coalescent and forward time simulation and re-sampling approaches) and it can generate output suitable as input for a number of standard genetic association tools (such as PLINK [Chang & al., 2015], GEMMA [Zhou & Stephens, 2014] or SNPTEST [Marchini & al., 2007]). In the following, I will describe the simulation strategy of the different phenotype components, and will demonstrate the usage and application of *PhenotypeSimulator* by simulating phenotypes to evaluate the power of different linear mixed model designs in a genetic association study.

Phenotypes are typically generated as the sum of genetic effects, effects from non-genetic factors and observational noise. Genetic effects can represent i) genetic variants that are associated with a phenotype and ii) infinitesimal genetic effects that reflect underlying population structure and relatedness in a cohort. Non-genetic effects are used to simulate environmental, experimental or other unexplained variance in the data. Although in many genetic association studies the sources of non-genetic correlation are often combined, I have found it valuable to separate these components to explore the impact of different correlation structures from these sources (see chapter 6). When simulating non-genetic factors, assumptions about their distribution have to be made and this choice depends on the specific biological effects that should be modelled. Common distributions are binomial (e.g. sex), normal or uniform distributions (e.g. weight, height) or categorical variables (e.g disease status). Correlated non-genetic effects can be used to simulate a phenotype component with a defined level of correlation between traits. For instance, such effects can reflect correlation structure decreasing in phenotypes with ordered or spatial components e.g. in imaging data. Observational noise captures any non-specified effects that arise due to, for instance, experimental measurement error. However, *PhenotypeSimulator* can also be used with a combined non-genetic covariance model, similar to more standard linear mixed models [O'Reilly & al., 2012; Zhou & Steph-

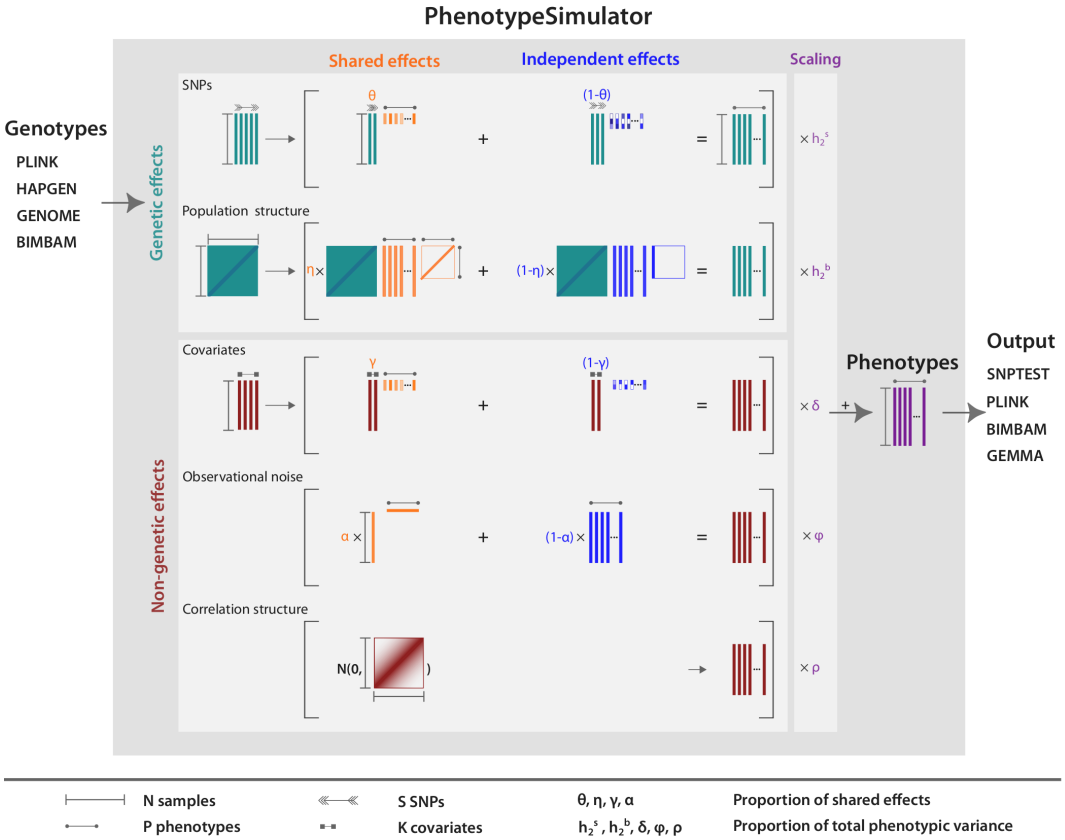

Figure 3.2: Phenotype simulation scheme. *PhenotypeSimulator* can take genotypes from a number of different input formats and uses these as the basis for the simulation of the genetic effects. In addition to the genetic effects, non-genetic covariates, observational noise and non-genetic correlation structure can be simulated. The effect structure of the upper four components can be divided into a shared effect across traits or an independent effect for a number of traits, allowing for complex phenotype structures such as the simulation of pleiotropy. Before combining the phenotype components, they are scaled to a user-defined proportion of the total phenotypic variance. Finally, the simulated phenotype and its components can be saved into a number of different genetic output formats. Arrows, lines and rectangle mark the dimensions of each component.

ens, 2014; Porter & O'Reilly, 2017]

The proportion of variance assigned to each component will differ depending on the biological understanding of the simulated phenotype. *PhenotypeSimulator* allows for the specification of these variance proportions and, in addition, provides the option to divide the explained variance into two components, one that is shared across phenotypes and a second component that acts independently on certain phenotypes. For instance, the level of shared and independent effects for a genetic variant allows for the simulation of different levels of pleiotropy.

There are many ways to simulate these phenotype components depending on the scope and the model to be tested. Typically, it is assumed that the overall phenotype structure is well represented by an additive linear combination of individual components [Stephens, 2013; Marigorta & Gibson, 2014; Zhou & Stephens, 2014; Loh & al., 2014]. For *PhenotypeSimulator*, I assume this phenotype structure and sum the individual phenotype components to generate the final phenotypes.

3.2.1. Phenotype components

In *PhenotypeSimulator*, the phenotypes $\mathbf{Y} \in \mathcal{R}^{N,\,P}$ of N samples and P traits are generated as the sum of i) genetic variant effects $\mathbf{U} \in \mathcal{R}^{N,\,P}$, ii) infinitesimal genetic effects $\mathbf{G} \in \mathcal{R}^{N,\,P}$, iii) non-genetic effects $\mathbf{C} \in \mathcal{R}^{N,\,P}$, iv) correlated non-genetic effects $\mathbf{T} \in \mathcal{R}^{N,\,P}$ and v) observational noise effects $\mathbf{\Psi} \in \mathcal{R}^{N,\,P}$ (figure 3.2). For component i-iv, a certain percentage of their variance is shared across all traits (shared) and the remainder is independent (ind) across traits. The option to divide the variance into shared and independent allows for the simulation of phenotypes with additional complexity. For instance, the level of shared and independent fixed genetic effects allows for the simulation of different levels of pleiotropy.

1. *Genetic variant effects:* For the SNP genetic effects, S random SNPs for N samples are drawn from the (simulated) genotypes. From the S random SNPs, a proportion $\boldsymbol{\theta}$ is selected to be causal across all traits. The shared genetic variant effect is simulated as the matrix product of this shared causal SNP matrix $\mathbf{X}^{\text{shared}} \in \mathcal{R}^{N,\,\theta \times S}$ and the shared effect size matrix $\mathbf{B}^{\text{shared}} \in \mathcal{R}^{\theta \times S,\,P}$. The columns of the shared effect size matrix are simulated to be perfectly correlated, i.e. the effect of a SNP genetic effect is proportionally the same for all affected traits. The effect sizes for $\mathbf{B}^{\text{shared}}$ can either be simulated to have normal or uniform properties. The is implemented as follows in *PhenotypeSimulator*: $\mathbf{B}^{\text{shared}}$ is the matrix product of the two vectors $b_s \in \mathcal{R}^{\theta \times S,\,1}$ and

$b_p^T \in \mathcal{R}^{1,\,P}$. To simulate effect sizes with approximately normal properties [Oliveira & Seijas-Macias, 2012, Eq 31-33], b_s and b_p are drawn from two normal distributions, where $\mu_{b_p} = 0$ and $\sigma_{b_p} = 1$ and μ_{b_s} and σ_{b_s} specified by the user. For the simulation of uniformly distributed effect sizes, b_s and b_p^T are drawn from two exponential distributions whose negative normalised log product yields an approximate uniform distribution [Song, 2005] across the user defined range. The remaining $(1-\theta) \times S$ SNPs are simulated to have an independent effect across a specified number of traits P^{ind}. To realise this structure, $\mathbf{B}^{\text{ind}} \in \mathcal{R}^{(1-\theta)\times S,\,P}$ is initialised with either normally or uniformly distributed entries, with μ_B and σ_B as specified by the user (same as for shared effect). Subsequently, $P - P^{\text{ind}}$ traits are randomly selected and the row entries for $\mathbf{B}^{\text{ind}}$ at these traits set to zero. The independent genetic variant effect is the matrix product of $\mathbf{X}^{\text{ind}} \in \mathcal{R}^{N,\,(1-\theta)\times S}$ and $\mathbf{B}^{\text{ind}}$.

2. *Non-genetic covariate effects:* The non-genetic covariate effects are based on K non-genetic covariates $\mathbf{W} \in \mathcal{R}^{N,\,K}$, with a proportion γ being shared across all traits yielding the shared covariates matrix $\mathbf{W}^{\text{shared}} \in \mathcal{R}^{N,\,\gamma\times K}$. The proportion of $1-\gamma$ non-genetic covariates that are independent make up the independent covariates matrix $\mathbf{W}^{\text{ind}} \in \mathcal{R}^{N,\,(1-\gamma)\times K}$. The distributions for each of the K non-genetic covariates are independent and can be either normal, uniform, binomial or categorical. The distribution and respective parameters are chosen by the user. The effect size matrices $\mathbf{A}^{\text{shared}} \in \mathcal{R}^{\gamma\times K,\,P}$ and $\mathbf{A}^{\text{ind}} \in \mathcal{R}^{(1-\gamma)\times K,\,P}$ were designed as described for the genetic effects. The final non-genetic covariate effects are the matrix product of the covariate matrices and their effect size matrices: $\mathbf{W}^{\text{ind}}\mathbf{A}^{\text{ind}}$ and $\mathbf{W}^{\text{shared}}\mathbf{A}^{\text{shared}}$.

3. *Infinitesimal genetic effects:* The basis of the infinitesimal genetic effect $\mathbf{U}$ is the $N \times N$ genetic relationship matrix $\mathbf{K}$, either estimated from the genotypes of the simulated samples as $\frac{1}{m}\mathbf{X}\mathbf{X}^T$, where m is the mean value of the diagonal elements of $\mathbf{X}\mathbf{X}^T$ or provided by the user. A suitable model for simulating the infinitesimal genetic effect $\mathbf{U} \in \mathcal{R}^{N,\,P}$ with the known $N \times N$ sample covariance $\mathbf{K}$ and trait covariance $\mathbf{C}$ is a multivariate normal distribution (as for instance in [Zhou & Stephens, 2014; Casale & al., 2015]) where

$$\text{vec}(\mathbf{U}) \sim \mathcal{N}_{N\times P}(\,\text{vec}(\mathbf{0})\,,\ \mathbf{C}\otimes\mathbf{K}\,) \tag{3.1}$$

The structure of $\mathbf{C}$ depends on the desired design of the covariance effect, which can be either shared or independent across traits. This distribution can

be realised by simulation a random variable $\mathbf{Z} \in \mathcal{R}^{M,\,L}$ as iid $\mathcal{N}\,(\,0\,,\,1\,)$ and setting

$$\text{vec}(\mathbf{U}) = \mathbf{BZA}^T \tag{3.2}$$

where $\mathbf{B} \in \mathcal{R}^{N,\,M}$ reflects the genetic relationship i.e. sample covariance with $\mathbf{K} = \mathbf{BB}^T$ and $\mathbf{A} \in \mathcal{R}^{P,\,L}$ the trait covariance with $\mathbf{C} = \mathbf{AA}^T$, respectively ($M$ and L depend on the rank of K and C, hence are bound by N and P). A detailed derivation for equation (3.2) from equation (3.1) can be found in chapter C and has similarly been applied in [Casale & al., 2015].

By recasting Equation 3.1 as Equation 3.2, the infinitesimal genetic effect $\mathbf{U}$ described by a multivariate-normal distribution is effectively modelled as the product of three matrices, representing the sample covariance ($\mathbf{B}$), a normally distributed variable ($\mathbf{Z}$) and the trait covariance ($\mathbf{A}$). Different designs of $\mathbf{A}$ will allow for the simulation of shared and independent genetic random effects. For the independent effect, $\mathbf{A}^{\text{ind}}$ is a diagonal matrix with normally distributed entries: $(\mathbf{A}^{\text{ind}})^T = \text{diag}(a_1, a_2, \ldots, a_P) \sim \mathcal{N}\,(\,0\,,\,1\,)$, such that $\mathbf{U}^{\text{ind}} = \text{vec}(\mathbf{BZ}(\mathbf{A}^{\text{ind}})^T)$. $\mathbf{A}^{\text{shared}}$ of the shared effect is simulated as a matrix of column rank one, with normally distributed entries in column one and zeros elsewhere: $a_{i,1} \sim \mathcal{N}\,(\,0\,,\,1\,)$ and $a_{i,j\neq 1} = 0$ such that $\mathbf{U}^{\text{shared}} = \text{vec}(\mathbf{BZ}(\mathbf{A}^{\text{shared}})^T)$.

4. *Correlated non-genetic effects:* Correlated non-genetic effects are simulated as a multivariate normal distribution with a covariance matrix described by a defined trait-by-trait correlation. Any correlation structure between the phenotypes can be simulated with this effect component, as the desired correlation matrix $\mathbf{C}$ can be supplied by the user. In addition, as a simple approximation for spatially correlated phenotypes as they might occur for instance in image-based phenotypes, *PhenotypeSimulator* provides the construction of $\mathbf{C}$ as follows: traits of distance $d = 1$ (adjacent trait columns) will have the highest specified correlation r, traits with $d = 2$ have a correlation of r^2, up to traits with $d = (P-1)$ with a correlation of $r^{(P-1)}$) , such that the correlation is highest at the first off-diagonal element and decreases exponentially by distance from the diagonal. The correlated non-genetic effect matrix is simulated as $\mathbf{T} \sim \mathcal{N}_{N\times P}(\,\mathbf{0}\,,\,\mathbf{C}\,)$.

5. *Observational noise:* The observational noise effects $\mathbf{\Psi}$ are simulated as the sum of a shared and an independent observational noise effect. Both effect components are simulated by the matrix product of $\mathbf{B} \in \mathcal{R}^{N,\,P} \sim \mathcal{N}\,(\,0\,,\,1\,)$ with $\mathbf{A} \in \mathcal{R}^{P,\,P}$. To realise the shared effect $\mathbf{\Psi}^{\text{shared}}$, $\mathbf{A}^{\text{shared}}$ is simulated as a matrix

of row rank one, with normally distributed entries in row one and zeros elsewhere: $a_{1,j} \sim \mathcal{N}(0,\ 1)$ and $a_{i\neq 1,j} = 0$. $\mathbf{A}$ of the independent component is a diagonal matrix with normally distributed entries:
$(\mathbf{A}^{\text{ind}})^T = \text{diag}(a_1, a_2, \dots, a_P) \sim \mathcal{N}(0,\ 1)$.

3.2.2. Scaling and phenotype construction

PhenotypeSimulator requires at least one phenotype component to simulate the phenotypes. Components can be combined as specified by the user and the correlation they introduce in the trait structure can be controlled by the specified levels of independent and shared effects (at the extremes, components can be simulated to either only have shared or independent effects). If desired, a simple phenotype structure following a model as cast for instance in the multi-variate normal model by [Zhou & Stephens, 2014] can be achieved by specifying only genetic variant effects, non-genetic covariate effects, infinitesimal genetic effects and observational noise. I have designed *PhenotypeSimulator* such that the amount of variance that each component should contribute to the total phenotypic variance can be specified by the user. Every component is thereby scaled by a factor a such that its average column variance explains x percent of the total variance. The scale factor a is derived as follows: Let X be a random variable with expected value $E[X] = \mu_x$ and variance $V[X] = E[(X-\mu_x)^2]$ and let $Y = aX$. Then

$$\begin{aligned} E[Y] &= a\mu_x \\ V[Y] &= E[(Y-\mu_y)^2] \\ V[Y] &= E[(aX - a\mu_x)^2] \\ &= a^2 E[(X-\mu_x)^2]. \end{aligned} \tag{3.3}$$

Hence, the scaling of a random variable by a leads to the scaling of its variance by a^2. To scale the phenotype components such that their average column variance $\overline{V}_{col} = \frac{V_1+\dots+V_p}{p}$ explains a specified percentage x of the total variance, choose the scaling factor a such that:

$$\begin{aligned} x &= a^2 \times \overline{V}_{col} \\ a &= \sqrt{x\overline{V}_{col}^{-1}} \end{aligned} \tag{3.4}$$

The final simulated phenotype $\mathbf{Y}$ is expressed as the sum of the scaled genetic variant effects, the non-genetic covariates, the correlated non-genetic effects and obser-

vational noise effects:

$$\begin{aligned}\mathbf{Y} = \mathbf{X}^{\text{shared}}\mathbf{B}^{\text{shared}} + \mathbf{X}^{\text{ind}}\mathbf{B}^{\text{ind}} + \mathbf{W}^{\text{shared}}\mathbf{A}^{\text{shared}} + \mathbf{W}^{\text{ind}}\mathbf{A}^{\text{ind}} \\ + \mathbf{U}^{\text{shared}} + \mathbf{U}^{\text{ind}} + \mathbf{T} + \mathbf{\Psi}^{\text{shared}} + \mathbf{\Psi}^{\text{ind}}.\end{aligned} \tag{3.5}$$

3.2.3. Case study

To demonstrate the usage and application of *PhenotypeSimulator*, I simulated a set of phenotypes and used them to evaluate the power of different linear mixed model designs in GWAS. In order to demonstrate the integration of *PhenotypeSimulator* with already established simulation and GWAS tools, I choose Hapgen2 [Su & al., 2011] for genotype simulation, used *PhenotypeSimulator* for phenotype simulation based thereon and applied GEMMA (version 0.96) [Zhou & Stephens, 2014] for the GWAS. The analysis code and parameters of this case study, from the data simulation to the genome-wide association study are supplied as a vignette to the R package.

I simulated genotype data for 1,000 individuals via Hagen2, mimicking population structure from four populations in the 1000Genomes project [1000 Genomes Project Consortium, 2012] (similar to the genotype structure described in section 3.1). The simulated genotypes of this cohort served as the basis for the genetic variant and infinitesimal genetic effects. I generated a phenotype set consisting of three traits with ten genetic variant effects and four non-genetic covariates. For the ten genetic variant effects, I randomly selected ten variants from the genotypes and simulated shared genetic variant effects across all phenotypes. I introduced additional correlation structure by including an infinitesimal genetic effect based on the individuals' kinship estimates as well as a non-genetic correlated (correlation: 0.8) and an observational noise effects. The total genetic variance accounts for 60% of the variance leaving 40% of variance explained by the noise terms. Figure 3.3 shows the trait-to-trait correlations of the final phenotype and each of its components.

The final phenotypes served as the response variable in the GWAS based on LMMs with the simulated SNPs and non-genetic covariates as fixed effects and the kinship estimated from the genotypes as part of the genetic random effect [Zhou & Stephens, 2014] (see section 1.7.6). I analysed the power of jointly modelling all three phenotypes (multi-trait) and the power of single-trait models where the association of each phenotype is analysed separately. The single-trait GWAS was run for all three traits. All GWAS were conducted with GEMMA (version 0.96) [Zhou & Stephens, 2014]. In both, the multi-trait and single-trait GWAS, the phenotypes (-p flag) were modelled as the sum of genetic (simulated SNPs; -g flag) and non-genetic (simulated covari-

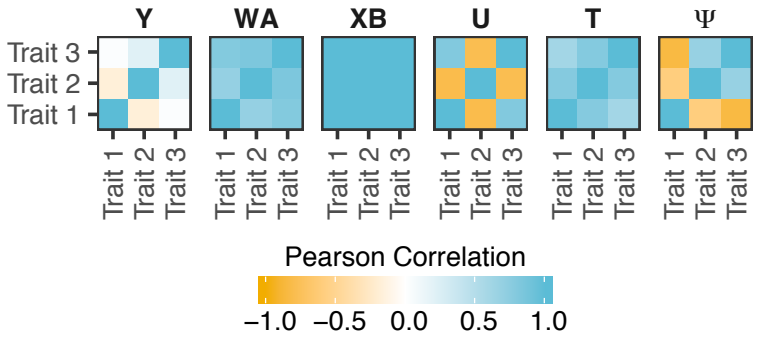

Figure 3.3: Phenotype simulation. Heatmaps of the trait-by-trait correlation (Pearson correlation) of a simulated phenotype ($\mathbf{Y}$) and its five phenotype components: genetic variant effects $\mathbf{XB}$, infinitesimal genetic effects $\mathbf{U}$, non-genetic covariates $\mathbf{WA}$, correlated non-genetic effects $\mathbf{T}$ and observational noise $\mathbf{\Psi}$. The non-genetic covariates consist of four independent components, two following a binomial and two following a normal distribution. The genetic variant effect of ten causal SNPs with shared effect across all traits, yielding the strong correlation structure observed above. The highest correlation for the correlated non-genetic effect was set at 0.8.

ates; -c flag) fixed effects, a random genetic effect (with the eigenvectors and values of the kinship matrix, -u and -d flag) and observational noise (linear mixed model with likelihood ratio test using the -lmm 2 flag). For a comparison of the number of causal SNPs recovered in the multi-trait and single-trait GWAS, the p-values of the single-trait GWAS were adjusted by the number of test conducted (Bonferroni adjustment for three tests).

For the simulated phenotypes with shared genetic variant effects only, the multi-trait GWAS shows a greater power compared to any of the single trait analyses (figure 3.4. The multi-trait GWAS detected four out of the ten SNPs for which a phenotype effect was modelled that pass the commonly used genome-wide significant threshold of 5×10^{-8} [Fadista & al., 2016]. The single-trait GWAS only recovered three of these SNPs. The ability of linear (mixed) models to detect the SNPs for which a phenotype effect was modelled depends on the allele frequencies of these SNPs and the effect size [Cohen, 1992; Halsey & al., 2015]: the higher the effect size and/or the allele frequencies the better the power to detect the SNP effects. The p-values of all SNPs with simulated effect on the phenotypes in relation to their allele frequencies and simulated effect sizes is depicted in figure 3.5. It shows a strong trend for SNPs with high allele frequencies and large simulated effect sizes to have low p-values.

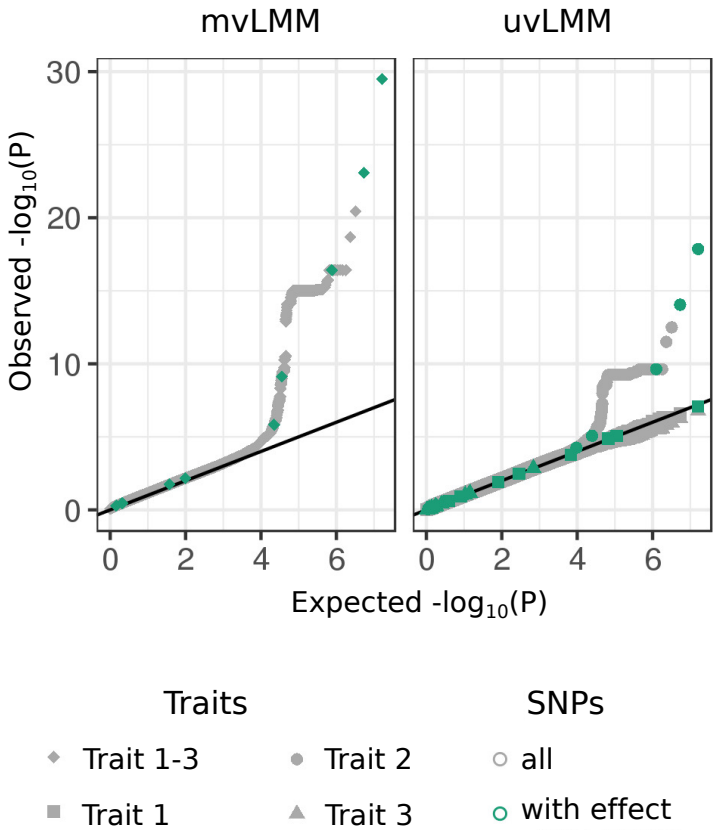

Figure 3.4: Comparison of multi-trait to single-trait GWAS. Quantile-quantile plots of p-values observed from the multi-trait GWAS (via multivariate linear mixed model; mvLMM) to single-trait GWAS (via univariate linear mixed models; uvLMM) fitted to each of the about eight million genome-wide SNPs (grey), including the ten SNPs for which a phenotype effect was modelled (green)

3.3. Conclusion

PhenotypeSimulator offers a framework for complex multi-trait, multi-locus phenotype simulations in quantitative genetics packaged in an easy to use manner for statistical geneticists. *PhenotypeSimulator* it is the only simulation package that I know that can simulate complex multi-trait phenotypes with complex multi-locus genetics, including a population structure term with phenotypic correlation. It can create realistic covariate structures with similar properties (e.g. categorical covariates or covariates drawn from different distributions) to real covariates. The different phenotype components can be independently extracted and scaled, for example having the"true" variance components and covariance matrices from the simulation readily

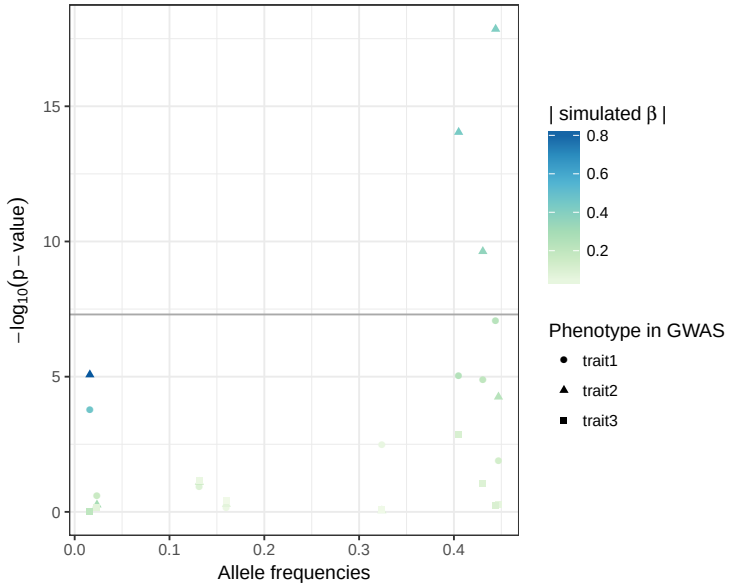

Figure 3.5: Relationship between p-values, allele frequencies and simulated effect sizes. The p-values of all SNPs with a simulated effect on the phenotypes are depicted in relation to their allele frequencies and simulated effect sizes. SNPs with low-allele frequencies and/or small simulated effect sizes do not pass the genome-wide significance threshold (horizontal line).

available for comparison to inference schemes.

The underlying model for *PhenotypeSimulator* corresponds to the common place linear mixed model framework. As such, it is limited in its use for benchmarking between methods, where linear mixed models methods are likely to perform best. However, the need for an underlying model is true for any simulation package.

I have developed *PhenotypeSimulator* as a flexible component in the standard genetics pipeline, with the ability to both read genetic formats from well used tools and output phenotypes compatible with many tools. It is freely available as R/CRAN package and its code is present on github (`https://github.com/HannahVMeyer/PhenotypeSimulator`). This allows easy large scale deployment for comprehensive simulation across many parameter settings.

In this thesis, phenotypes simulated with *PhenotypeSimulator* built the basis for the method development in chapter 4 and chapter 6.

4

Extending linear mixed models to high-dimensional phenotypes

Different strategies and challenges for multi-trait GWAS of high-dimensional phenotypes have been discussed in section 1.7.7. Phenotypes can either be transformed into a lower dimensional space prior to the association study or the summary statistics from single-trait GWAS can be combined post-hoc to obtain quasi multi-trait association results. In contrast, multivariate LMMs can directly model the genotypic association across a moderate number of phenotypes. In the following chapter, I will describe the challenges of multivariate LMMs for high-dimensional phenotypes and present LiMMBo, a new method for the genotype-phenotype mapping of high-dimensional datasets.

LMMs have become a workhorse in genetic association studies as they allow to control for complex sample-by-sample covariance structures that can reflect population structure and relatedness (discussed in detail in section 1.7.6). In summary, LMMs commonly describe the phenotype as a linear combination of fixed effects – experimental and/or technical covariates and the genotype marker of interest, and a random genetic effect and residual noise which capture the genetic and residual covariances between traits. The association of the genetic marker is evaluated by comparing the alternative hypothesis that the genotype has an effect on the pheno-

type which is unequal to zero to the null model of no effect (section 1.7.8). In practice, this means estimating the effect size of the fixed genetic effects and the random effect covariance terms for the alternative model and the random effect covariance terms for the null model where the effect size of the genetic marker is zero.

The first LMM implementations estimated all variance components (genotype effect size and random effect covariance terms, equation (1.52)) anew for each SNP-phenotype association. However, in human genetics effect sizes are generally assumed to be small compared to the overall phenotypic variance [Kang & al., 2010; Zhang & al., 2010]. Consequently, estimates of the random effect covariance terms under the null model can serve as a good approximation. Based on these differences in the estimation of the random effect covariance terms, LMMs can broadly be grouped into two categories. The exact methods with covariance estimates under the alternative model and approximate methods, where the random effect covariance terms are only estimated once under the null model of no fixed genetic effect and are then used as predefined random effects in the alternative models for all genome-wide associations.

Within these two categories, one can further distinguish between methods only applicable as univariate tests or methods that allow for multivariate testing. Table 4.1 summarises commonly used frameworks and describes their computational complexity[1].

Among the exact methods, FaST-LMM-select reduces the complexity best in terms of sample size by selecting the number of SNPs to use for the estimation of the RRM. However, it can only be applied in univariate analyses while MTMM and GEMMA extend to multivariate cases. BOLT-LMM scales best with increasing samples sizes in the group of approximate tests, by directly using the genotypes and not computing or storing the RRM. All other methods have an upfront $O(N^3)$ operation for the eigendecomposition of the RRM. TASSEL reduces this complexity based on grouping of the samples and thereby effectively reducing the size of the RRM.

With the generation of ever-increasing cohort sizes in genetic association studies, most LMM frameworks are optimised for the number of samples as described above for BOLT-LMM and TASSEL. While the remaining methods still have the upfront cubic computation of the RRM's eigendecomposition, subsequent steps have been adapted to scale linearly or quadratically with the number of samples for the majority of the applications.

[1]The computational complexity and algorithms for the GCTA implementations [Yang & al., 2011] of multivariate genetic variance estimation [Lee & al., 2012] and LMM for association testing [Yang & al., 2014] could not be found in the original publications and are therefore not listed

Table 4.1: Linear mixed model frameworks for genetic association studies. A list of popular LMM frameworks, grouped by their usage of covariance estimates when fitting the alternative model (first column: E: exact, A: approximate). The complexity describes the complexity for fitting a single LMM as specified in the original publication or summarised elsewhere, as indicated by the footnotes. P indicates the trait size that the model was designed for (according to the original publication). Models with specific parameters are described in more detail in the text (FaST-LMM-select and TASSEL). N: number of samples; s_c: number of SNPs used for singular value decomposition; c: compression factor with $c = \frac{N}{g}$ for g individuals per group; t, t_1 andt_2: average number of iterations needed to find parameter estimates. GRAMMAR-Gamma, FaST-LMM-select: t steps of the Brent's algorithm; GEMMA, MTMM: t_1 steps of the EM algorithm, t_2 steps of the NR algorithm; BOLT-LMM: t steps of the variational Bayes and conjugate gradients; TASSEL: t steps of the ProcMixed algorithm in SAS; mtSet: t steps of the L-FBGS.

	Framework	Complexity O	P	Reference
E	FastLMM-select	$Ns_c^2 + N^2 + tN$	1	[Lippert & al., 2011]
	GEMMA	$N^3 + N^2P +$ $t_1NP^2 + t_2NP^6$	10	[Zhou & Stephens, 2014] [Zhou & Stephens, 2014]
A	MTMM	$t_1N^3P^3 + t_2N^3P^7$ $+N^2P^2$	2	[Korte & al., 2012][2]
	EMMAX	$N^3 + tN + N^2$	1	[Kang & al., 2010]
	TASSEL	$\frac{1}{c^3}N^3$	1	[Zhang & al., 2010]
	GRAMMAR-Gamma	$N^3 + tN + N$	1	[Svishcheva & al., 2012]
	BOLT-LMM	tN	1	[Loh & al., 2014]
	mtSet	$N^3 + t(NP^4 + P^5)$	10	[Casale & al., 2015]

The reduced complexity in the sample term comes as a trade-off with the number of traits that can be analysed. Specifically, computations become prohibitive as soon as a few tens of traits (table 4.1, column P) are considered, with computational complexities ranging from $O(P^5)$ to up to $O(P^7)$ for existing methods [Casale & al., 2015; Korte & al., 2012]. In practice, this limits these models to moderate trait numbers.

To overcome this limitation, I developed a simple, but surprisingly effective heuristic to efficiently estimate large trait covariance matrices in linear mixed model with bootstrapping (LiMMBo), thereby allowing for the analysis of datasets with a large number of phenotypic traits. LiMMBo and its application (chapter 5) is currently under revision and available in pre-print [Meyer & al., 2018]. I conducted all simulations and analyses and generated all results. I provide LiMMBo as an open source Python package (`https://pypi.org/project/limmbo/`) with command line interface and its source code is available on github: `https://github.com/HannahVMeyer/limmbo`.

[2] Listed in [Zhou & Stephens, 2014]

4.1. LiMMBo: Linear mixed modeling with bootstapping

To extend the range of LMMs for high-dimensional phenotype sets, I chose to build on an approximate model in order to avoid the repeated estimation of the trait-by-trait covariance matrices. In that respect, the multivariate LMM developed by Lippert, Casale and colleagues [Lippert & al., 2014; Casale & al., 2015] harboured many advantages. It is computationally efficient for a moderate number of traits, has successfully been used in multi-trait studies [Cannavò & al., 2016; Schor & al., 2017] and collaboration with its developers was easily realisable. Their model is cast as

$$\mathbf{Y} = \mathbf{G} + \mathbf{\Psi}, \tag{4.1}$$

where the $N \times P$ phenotype matrix $\mathbf{Y}$ for N individuals and P traits is modelled as the sum of a genetic (or polygenic) component $\mathbf{G}$ and a noise component $\mathbf{\Psi}$ (I have omitted additional fixed effects for notational brevity). Here, $\mathbf{G}$ and $\mathbf{\Psi}$ are random effects following matrix normal distributions:

$$\begin{aligned} \mathbf{G} &\sim \mathcal{MN}_{N,P}\left(0, \mathbf{R}, \mathbf{C}_g\right) \\ \mathbf{\Psi} &\sim \mathcal{MN}_{N,P}\left(0, \mathbf{I}_N, \mathbf{C}_n\right), \end{aligned} \tag{4.2}$$

where $\mathbf{R}$ denotes the $N \times N$ genetic relationship matrix, $\mathbf{I}_N$ is the $N \times N$ identity matrix and $\mathbf{C}_g$ and $\mathbf{C}_n$ are the genetic and the residual $P \times P$ trait covariance matrices, respectively. The marginal likelihood of the model in equation (4.1) can be expressed in terms of a multivariate normal distribution of the form

$$p\left(\mathbf{Y}|\mathbf{C}_g, \mathbf{R}_N, \mathbf{C}_g\right) = \mathcal{N}\left(\mathrm{vec}\left(\mathbf{Y}\right)|0, \mathbf{C}_g \otimes \mathbf{R}_N + \mathbf{C}_n \otimes \mathbf{I}_N\right), \tag{4.3}$$

where the covariance structure of the phenotypes (in shape of the $N \times P$ phenotype vector $\mathrm{vec}\left(\mathbf{Y}\right)$ through stacking the columns of the phenotype matrix) is described by the sum of the Kronecker products $\otimes$ of the sample and trait covariance terms. This model enables efficient inference schemes by exploiting Kronecker identities for the eigendecomposition of the full covariance matrix [Lippert & al., 2014; Rakitsch & al., 2013; Zhou & Stephens, 2014; Casale & al., 2015]. In particular, it allows for decoupling the decomposition of $\mathbf{C}_g$ and $\mathbf{R}_N$, which greatly increase the efficiency of the inference as $\mathbf{R}_N$ is constant. The model in equation (4.1) also corresponds to the null model when using the multi-trait LMM for genetic association mapping.

The complexity of this multivariate LMM implementation (from now referred to

as "standard REML") is $O(N^2 + t(NP^4 + P^5))$ with N the number of samples, P the number of traits, and t the number of iterations of Broyden's method, which uses an approximation of the second derivative for optimising the REML of the parameter estimates. From this equation, it becomes evident that as the number of traits increases, the complexity increases steeply and explains why this LMM set-up is not feasible for large trait sets (as is the case for other inference schemes table 4.1). To overcome the bottleneck of estimating the trait-by-trait covariance matrices, I developed a simple method that efficiently uses a subsampling approach to estimate $\mathbf{C}_g$ and $\mathbf{C}_n$.

4.2. Covariance estimation via bootstrapping

The key innovation of LiMMBo is to perform the variance decomposition on b bootstrap samples of s traits instead of on the whole dataset, and use those bootstrap samples to reconstruct the full $\mathbf{C}_g$ and $\mathbf{C}_n$ matrices (figure 4.1). In detail, from the total phenotype set with P traits, b subset of s traits are randomly selected. b depends on the overall trait number P and the sampling size s and is chosen such that each two traits are drawn together at least c times (default: $c = 3$). For each subset, the variance decomposition is estimated via the null model of the multivariate linear mixed model (mvLMM)), i.e. without the genetic variant effect $\mathbf{x}$ (equation (4.3)) and the $s \times s$ covariance matrices $\mathbf{C}_g^s$ and$\mathbf{C}_n^s$ recorded. For each trait pair, their covariance estimate is averaged over the number of times they were drawn. The challenge lies in combining the bootstrap results in such a way that the resulting $\mathbf{C}_g$ and $\mathbf{C}_n$ matrices are true covariance matrices i.e. positive semi-definite and serve as good estimators of the true covariance matrices. This is achieved by fitting (least-squares estimate) the covariance estimates of the b subsets to the closest positive-semidefinite matrices via a limited-memory version of the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS algorithm), which uses approximations of the Hessian matrix for finding the parameter estimates [Byrd & al., 1995]). The average estimates of the parameters are used to initiate the matrices.

4.3. Data simulation

Using *PhenotypeSimulator* (chapter 3), I simulated a number of different phenotype datasets to evaluate LiMMBo in terms of scalability, model calibration and power. The datasets differed in their overall trait size P, the percentage of variance explained

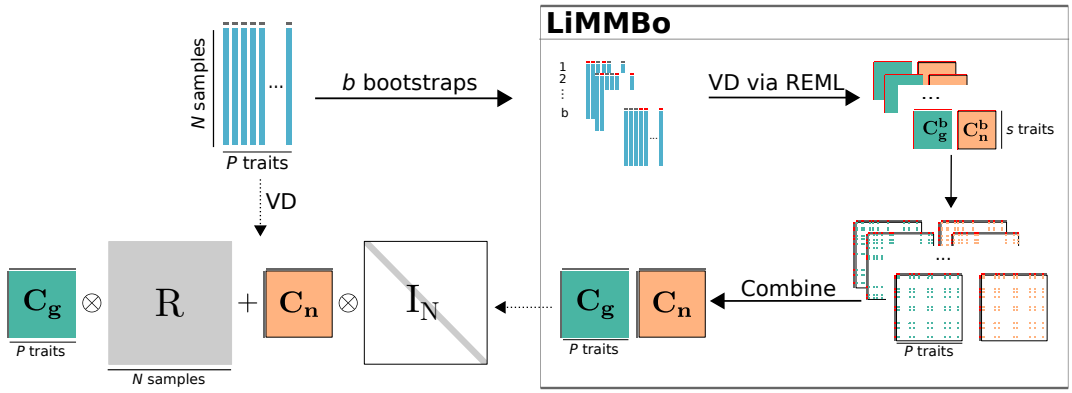

Figure 4.1: Variance decomposition. On the left-hand side, the phenotype set of P traits and N samples is decomposed into its $P \times P$ trait-to-trait covariances $\mathbf{C}_g$ and $\mathbf{C}_n$, based on the provided genetic sample-to-sample kinship estimate matrix $\mathbf{R}$. The noise sample-to-sample matrix $\mathbf{I}$ is assumed to be constant (identity matrix). Standardly, this is done by restricted maximum likelihood estimation of the null model of the mvLMM (Eq. 4.3). However, this direct variance decomposition (VD) via the standard REML implementation only works for moderate number of phenotype sizes. For higher trait-set sizes, LiMMBo serves as an alternative to the standardREML (right-hand side). Here, the phenotypes' variance components are estimated on b s-sized subsets of P which are subsequently combined into the overall $P \times P$ covariance matrices $\mathbf{C}_g$ and $\mathbf{C}_n$.

by genetics h_2 (sum of genetic variant and infinitesimal genetic effects) and the number of different phenotype components simulated to create the final phenotype. The phenotypes were simulated as described in section 3.2, based on the parameters and parameter values described in table 4.2 and table 4.3. Parameter values were generally chosen to cover a wide range a possible combinations and trait sizes. Parameters for levels of variance explained by the genetic and noise components were set to test their effect on the variance decomposition algorithm of the underlying LMM framework [Casale & al., 2015]. The variance decomposition is initiated by allocating an even split of variance explained to the genetic and random noise effects. The levels of variance explained were thus set to 0.5 each and deviations from this equal split into either direction (0.2, 0.8).

4.4. Scalability of LiMMBo

The complexity of the variance decomposition of the LMM framework that LiMMBo builds on is $O(N^2 + t(NP^4 + P^5))$. The second term depends on the overall trait size and describes the complexity of estimating the trait-by-trait covariance matrices $\mathbf{C}_g$ and $\mathbf{C}_n$. By bootstrapping s-sized samples from the overall trait size, this complexity

term changes to $bt(Ns^4 + s^5$, with the covariance estimation carried out for b bootstraps. In addition to the estimation of the covariance terms, the overall complexity of LiMMBo also depends on the fitting the BFGS algorithm n times to the full trait-set of size P. LiMMBo makes use of a Cholesky decomposition of the matrices to be fitted, resulting in $\frac{1}{2}P(P+1)$ model parameters to be fitted for both $\mathbf{C}_g$ and $\mathbf{C}_n$. Thus, the overall complexity of LiMMBo is $O(N^2 + bt(Ns^4 + s^5) + nP^2)$, which is the sum of the complexity of the bootstrap variance decompositions and the complexity of fitting the BFGS algorithm.

In order to assess and compare how LiMMBo scales, I performed variance decomposition both with LiMMBo and the standard REML approach on phenotypes with trait sizes ranging from 10 to 100 traits (parameters for phenotype simulation as described in table 4.3, total of ten simulated datasets per setup). For $P = 10$, the sampling datasize s was set to $s = 5$, otherwise $s = 10$. Figure 4.2 shows the overall time taken by the standard REML approach, LiMMBo and its two main components, the bootstrapping and the combination of the bootstrap results.

Table 4.2: Parameters for phenotype simulation. The total variance for the genetic and noise effects is the sum of the variance of their effect components and has to add to 1. Each component has a certain percentage of its variance that is shared across traits, while the rest is independent.

		variance	shared	independent
genetic effects	total	h_2		
	genetic variant effect	h_2^s	θ	1-θ
	infinitesimal genetic effects	h_2^g	η	1-η
noise effects	total	(1-h_2)		
	covariate effect	(1-h_2)δ	γ	1-γ
	observational noise	(1-h_2)(1-δ)	α	1-α

Table 4.3: Parameter values of simulated phenotypes for assessing scalability, calibration and power. The "genotype" parameter specifies the simulated genotype cohort which was used to simulate genetic effects (described in section 3.1). P are the different traitset sizes that were simulated. The parameters that follow are described in table 4.2 and specify the variance explained by each of the phenotype components. A variance explained equals zero means that this component was not simulated and corresponding non-applicable variance terms are designated with "-".

	Parameter values	
Parameter	Power	Calibration
Genotypes	relatedNoPopstructure	relatedNoPopstructure unrelatedNoPopstructure unrelatedPopstructure
P	10, 50, 100	10, 20, ..., 100
h_2^s	0.05, 0.2, 0.0125	0
h_2^g	0.95, 0.98, 0.9875	1
h_2	0.8, 0.5, 0.2	0.8, 0.5, 0.2
(1-h_2)δ	0.4	0
(1-h_2)(1-δ)	0.6	1
(1-h_2)	0.2, 0.5, 0.8	0.2, 0.5, 0.8
θ	0.6	-
η	0.8	0.8
γ	0.6	-
α	0.8	0.8

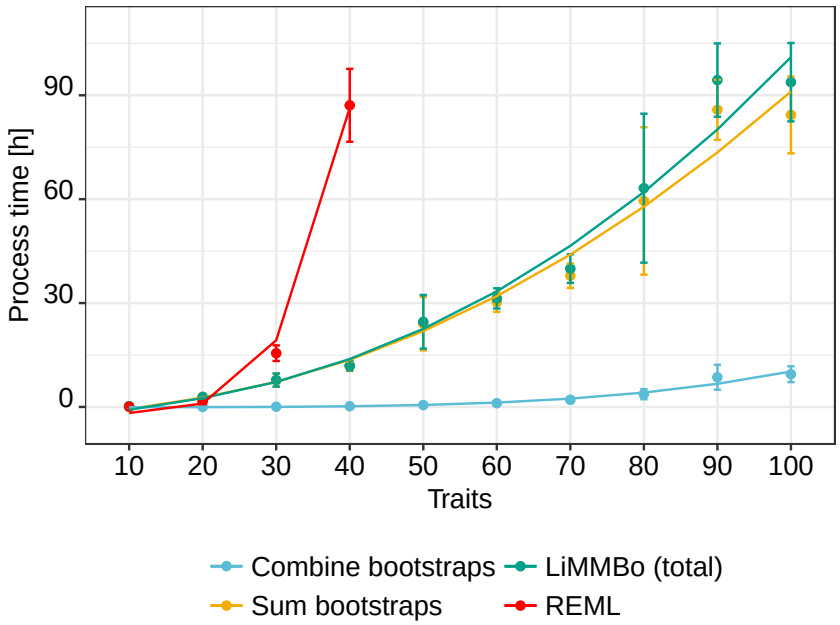

Figure 4.2: Scalability of LiMMBo compared to standard REML. Empirical run times for LiMMBo and the standard REML approach on three simulated datasets per phenotype size, with N =1,000 individuals each and different amount of variance explained by the genetic background signal (0.2, 0.5, 0.8). Points mark the mean run time across the different setups, error bars indicate their standard deviation. Lines were fitted for the bootstrapping step (orange): $n(Ns^4 + s^5)$; the combination of the bootstrapping (blue): $\frac{1}{2}P(P+1)$ and their combined run time (turquoise): $n(Ns^4 + s^5) + \frac{1}{2}P(P+1)$. b: number of bootstraps, s: bootstrap size, P: phenotype size. The majority of the run time is required for the bootstrapping. The run time for the standard REML results (red) are only depicted up to $P = 40$ when they already exceed the run times for $P = 100$ in the LiMMBo approach (REML: $O(N^2 + t(NP^4 + P^5))$).

The majority of the run time of LiMMBo is taken by the variance decomposition of the bootstrapped subsets, which accounts for at least 85% (70 traits) and on average 97% of the total run time. As a comparison, the time taken by the standard REML approach quickly exceeds the time of LiMMBo and becomes unfeasible for more than 30 traits.

While the bootstrapping keeps the complexity of LiMMBo effectively at $O(P^2)$, it has the major advantage of allowing for parallelisation of the covariance estimation step. Thus, LiMMBo computes the variance decomposition of each bootstrap independently and enables the use of multiple cores, allowing for an additional speed up of the process.

The role of the bootstrap size s, the number of bootstraps b and the co-sampling

of traits c on complexity has not been evaluated yet. Different combinations of these parameters will potentially yield different run times and might influence the covariance estimates and model calibration, which are described in the next sections. For the remainder of this chapter, the bootstrap size $s = 10$ and co-sampling of traits $c = 3$, which were used for the estimation of run time differences, are adapted for all further analyses. The influence of s, b and c and additional experiments for evaluating their role in the model are discussed in section 4.8.

4.5. LiMMBo yields covariance estimates consistent with REML estimates for moderate trait numbers

I evaluated the suitability of LiMMBo for covariance estimation of $\mathbf{C}_g$ and $\mathbf{C}_n$ on simulated datasets with different strength of infinitesimal genetic effects. I simulated phenotype sets composed of infinitesimal genetic effects $\mathbf{G}$ and observational noise effects $\mathbf{\Psi}$ only, omitting any genetic variant effects (additional parameters as described in table 4.3) and estimated these variance components subsequently with LiMMBo and standard REML. Variance estimation on simulated datasets allows for the comparison of the estimated covariance matrices to the true covariance matrices based on which the phenotypes were simulated. By computing the root mean squared deviation (RMSD) between the true and estimated covariance matrices from both methods, I obtain a measure that is directly comparable and independent of the trait set:

$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(C_{\text{true}} - C_{\text{estimate}})^2}{n}} \tag{4.4}$$

Figure 4.3 shows the comparison of both standard REML and LiMMBo-derived covariance matrices compared to the simulated, true covariance matrices. In the regime where REML is feasible, i.e. moderate trait sizes of up to 30, the RMSD can directly be compared: both methods provide consistent estimates across trait sizes with little difference between the methods. Importantly, the RMSD stays constant for the LiMMBo-derived estimates of the covariances, even for phenotypes of higher sizes.

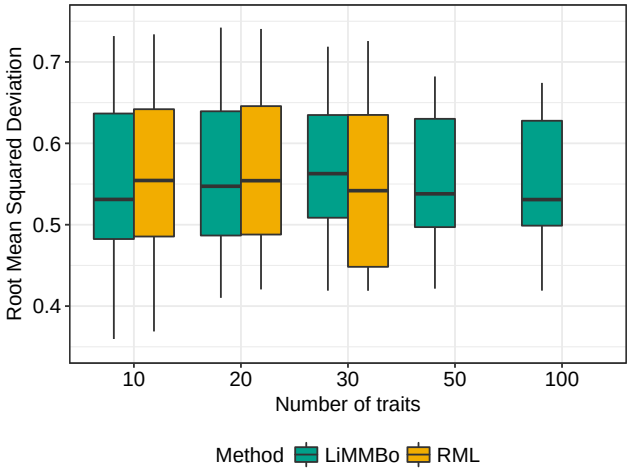

Figure 4.3: Comparison of trait-by-trait covariance estimates derived from standard REML and LiMMBo. Phenotypes with different percentage of variance explained by genetics ($h_2 = 0.2, 0.5, 0.8$) and different trait numbers were simulated. Subsequently, the genetic and noise trait-by-trait covariance matrices $\mathbf{C}_g$ and $\mathbf{C}_n$ were estimated both via LiMMBo and standard REML. These estimates were compared to the true (simulated) covariance matrix by computing their root mean squared deviation (RMSD; equation (4.4)). The boxplots summarise the RMSD across different variance levels for ten independent simulations each. For moderate traitset sizes ranging from 10 to 30 traits, LiMMBo and the REML approach yield consistent covariance estimates. Covariance estimation via LiMMBo stays stable with these observations in the higher trait sizes ($P = 50, 100$).

4.6. mtGWAS with LiMMBo-derived covariance matrices are well calibrated across all phenotype sizes

One key aspect in statistical method development is to ensure that the method is well-calibrated under the null model. Apart from gaining knowledge about the genetic and noise trait-by-trait covariance structure of a phenotype, variance decomposition into different random effect components yields estimates that can be supplied as known parameters to approximate mvLMM methods and multi-trait genome-wide association study (mtGWAS). As introduced by Jiang & Zeng [1995] and adapted by Korte & al. [2012], there are different model designs for mvLMM, depending on the underlying biological hypothesis regarding the effect of the genetic variant. The different models were described in section 1.7.8 and include any effect (effect size is unequal to zero for at least one trait), common effect (same effect size across all traits) and specific effect test (specific effects of the variant on a given trait). In practice, it is common to test for any effect as a means of discovering associated

genotypes and to refine the type of association later. As such, I chose to apply an any effect test for both the calibration and power analysis.

In order to test if LiMMBo-derived covariance estimates yield well calibrated test statistics, I simulated phenotype sets composed of infinitesimal genetic and observational noise effects only with 10, 20, 30, 50 and 100 traits and parameters described in table 4.3. For trait sizes of up to 30 traits, I compared the calibration of mtGWAS for LiMMBo- and standard REML-derived covariance matrices. As shown in figure 4.4, both methods yield p-values following a uniform distribution under the null model (compare figure 1.2C) across all phenotype sizes and variance explained by genetics, thus show appropriate calibration. For higher trait sizes, I also compared the

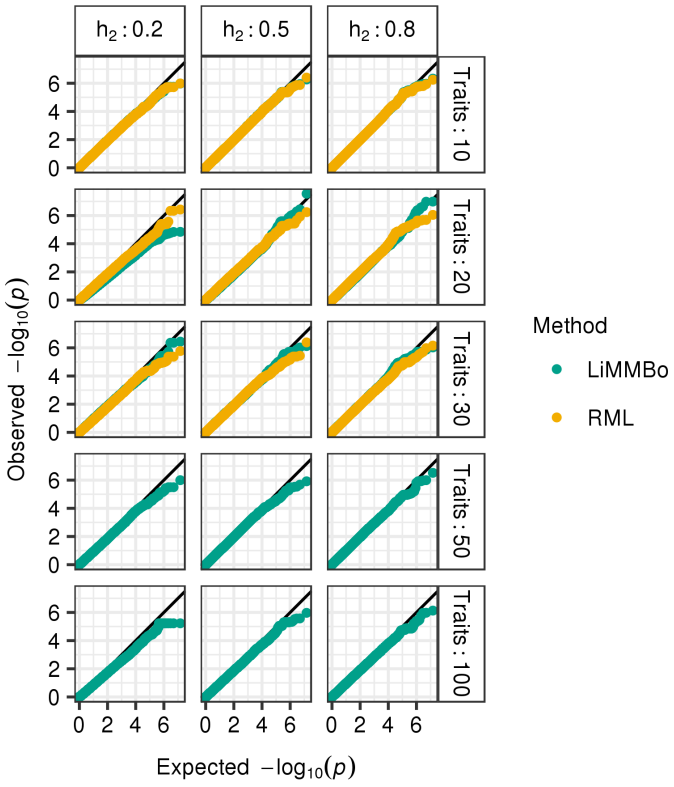

Figure 4.4: Calibration of mtGWAS based on covariance estimates from standard REML and LiMMBo. For moderate trait numbers ranging from 10 to 30 traits, phenotypes with different percentage of variance explained by genetics were simulated. The genetic and noise trait-by-trait covariance matrices $\mathbf{C}_g$ and $\mathbf{C}_n$ were then estimated both via LiMMBo and standard REML. The model calibration i.e. uniform distribution of p-values under the null model was assessed by mtGWAS with covariance estimates derived from either LiMMBo or REML. Quantile-quantile plots show uniform distribution for both methods across all trait sizes and levels of proportion of variance explained by genetics.

calibration of mtGWAS using a mvLMM to using a simple multivariate linear model (mvLM). The mvLM does not require the variance decomposition into different random effects, i.e. avoids the computational bottleneck of estimating the trait-trait covariance matrices, but simply uses principal components of the genotypes as fixed effects to adjust for population structure. For the residual trait-by-trait covariance structure σ_n, I used the empirical phenotypic trait-by-trait covariance. As depicted in figure 4.5, the calibration of the mvLM depends strongly on the population structure. For populations without related individuals, the mvLM shows a uniform p-value distribution and points to the usefulness of this simpler model approach for populations with well-defined structure. However for structured populations, the mvLM is poorly calibrated and clearly demonstrates the difficulty of adjusting for population structure via fixed effects in highly structured populations. In these scenarios, multi-trait mapping of high-dimensional phenotypes is only possible via LiMMBo.

4.7. Multi-trait genotype to phenotype mapping increases power for high-dimensional phenotypes

Multi-trait linear mixed models for low to moderate phenotype sizes have been shown to improve power by leveraging correlated background structure and trait-by-trait correlations resulting thereof [Casale & al., 2015]. For assessing the significance of the genotype-phenotype association via LLR test statistics where the likelihood of the full model is compared to the likelihood of the null model i.e. without the fixed genetic effect, the LLR statistic are translated into p-values via the appropriate χ^2 distribution with P degrees of freedom (section 1.7.3 and figure 1.2A, [Wilks, 1938]). In order to test if there is still a gain in power for a mvLMM with high-dimensional phenotypes, i.e. large number of degrees of freedom, I simulated phenotypes where I varied key parameters whose influence on power I wanted to investigate.

I varied trait numbers ($P = \{10, 50, 100\}$), the contribution of the genetic effects to the phenotypic variance ($h_2 = \{0.2, 0.5, 0.8\}$ and proportion of traits that are affected by the genetic variant effects ($a = \{0.2, 1\}$). Parameters of this phenotype simulation are described in table 4.2 and table 4.3. For each of these phenotype sets, I added 20 genetic variant effects to a subset of traits, creating phenotypes with different proportions of traits affected by the genetic variant effects. For each set-up, I simulated 50 independent phenotypes (a total of $2,250$ phenotypes $=$ 3 $h_2 \times 3$ trait sizes $\times 50$ permutations $\times 5$ subset sizes) and estimated the trait-by-trait

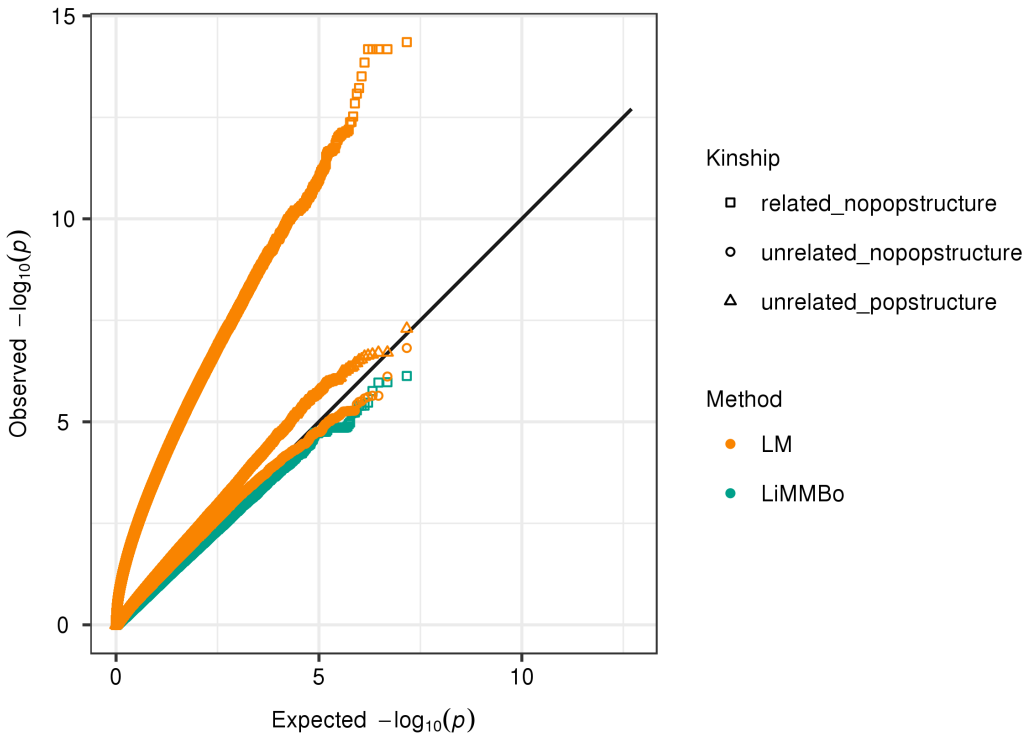

Figure 4.5: Calibration of mtGWAS via a simple linear model and LiMMBo. The three phenotype sets with 100 traits each were modelled as the sum of infinitesimal genetic and observational noise effects. The basis for the infinitesimal genetic effects build the three genotype cohorts simulated in section 3.1. The phenotypic variance explained by genetics was set to $h_2 = 0.8$. For the mvLMM (only shown for the population with related individuals), covariance estimates were derived via LiMMBo. In the mvLM, population structure was adjusted for via the first ten PCs of the genotype data. The mvLM is well calibrated for populations without related individuals. For the populations containing the latter, only the mvLMM is well calibrated.

covariance matrices $\mathbf{C}_g$ and $\mathbf{C}_n$ via LiMMBo. I used these estimates in a mvLMM to test the association between the known causal SNPs (from the simulation) and the phenotypes. In addition, I determined the association of the causal SNPs for each trait independently via univariate linear mixed model (uvLMM). The significance of the associations was assessed by comparing the p-values of these original associations to p-values obtained from mvLMM and uvLMM on 1,000 permutation of the genotypes. For the uvLMM, the p-values were adjusted for multiple testing by the number of traits that were tested and the minimum adjusted p-value across all traits for a given SNP recorded. For each SNP, the number of times the (adjusted)

p-value of the permutation was less or equal to the observed p-value was recorded and divided by the total number of permutations, yielding an empirical p-value per SNP.

I compared the results of the univariate and multivariate models to evaluate two key differences in the models. First, I can test which burden of the multiple association testing weights heavier, the correction for multiple testing in the uvLMM or the increased degrees of freedom in the mvLMM. This effect can be analysed by varying the number of traits in the phenotypes and keeping the other parameters constant. As depicted in figure 4.6A, for the highest number of phenotypes tested, both models are comparable in the number of causal SNPs they detect. For the other trait sizes tested, the multivariate model out-performs the univariate model by far. For these comparisons, an ideal scenario was assumed and all traits were affected by the genetic variant effects ($a = 1$) and the total genetic variance was low ($h_2 = 0.2$).

The influence of the proportion of traits affected by the causal SNPs on the power to detect these is depicted in figure 4.6B. This analysis allows for the evaluation of the second key difference in the models. The multivariate model can exploit correlated background structure and allows for the detection of pleiotropic effects, while the univariate model can only detect simple SNP-trait associations. This advantage becomes clear in figure 4.6B, where the median number of detected true SNPs depending on the proportions of traits affected by the causal SNPs is depicted. Here, the number of traits was kept constant at $P = 50$ and the mean genetic variance across all traits fixed at $h_2 = 0.2$, i.e. with an increase of the number of affected traits the contribution of the genetic component per trait decreases. The univariate model suffers from the weaker genetic components when a large number of traits are affected and loses power. In contrast, the multivariate model can still detect increasing percentages of true causal SNPs. The influence of the proportion of phenotypic variance explained by all genetic, i.e. genetic variant and infinitesimal genetic effects is shown in figure 4.6C. For both models, the number of detected SNPs decreases with increasing h_2, as the effect sizes of the SNPs become negligible compared to the overall genetic variance. However, the multivariate model is still able to exploit the correlation of the variant effects across traits and detects more SNPs in cases of high h_2. An overview of all parameter comparisons can be found in figure B.1 in the appendix.

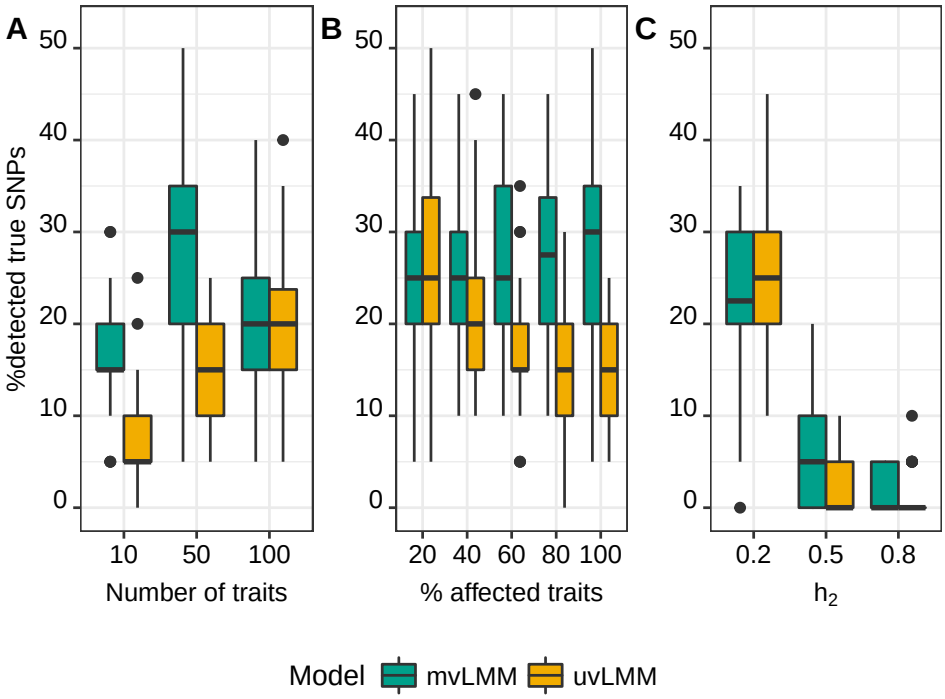

Figure 4.6: Power comparison for mvLMM and uvLMMs of high-dimensional phenotypes. Each panels show the influence of one simulation parameter on the power to detect the causal SNPs. When investigating one parameter, the other parameters were fixed at a certain value. For each set-up, 50 independent datasets were simulated and analysed. A. Influence of the number of traits: proportion of traits affected and the total genetic variance fixed at $a = 1$ and $h_2 = 0.2$, respectively. B. Influence of proportion of traits affected: trait size and total genetic variance fixed to $P = 50$ and $h_2 = 0.2$ respectively. C. Influence of total genetic variance: trait size and proportion of traits affected fixed to $P = 100$ and $a = 0.6$.

4.8. LiMMBo for multi-trait GWAS and beyond

In this chapter, I introduced LiMMBo, a new method for the multivariate analysis of large trait numbers, which uses a bootstrap method to estimate complex trait covariance matrices. The main benefit of LiMMBo is that it scales to 100s of phenotypes, both because of its inherent sub-sampling method and that the most computationally intense part of the method can be parallelised. To take advantage of the parallelisation, I implemented an optional automatic detection for multiple cores which allows for easy realisation of this process via the *Parallel Python Software* [Vanovschi, 2017]. In practice, this means that trait sizes up to 30 or 40 can be in hours, rather than taking several days as for standard REML-based methods. Most notably, complex datasets of 100s of traits, which is out of scope for the REML approaches, are feasible when using LiMMBo. I showed that the covariance matrices estimated via LiMMBo are as good an estimator of the real covariance matrices as the ones of the validated REML approach. Consequently, these covariance matrices produce well calibrated null models when used in LMM for GWAS, showing the validity of the approach. To show the advance of LiMMBo, I demonstrated the power gain for multi-trait GWAS of high-dimensional phenotypes with LiMMBo over standard single-trait models across a wide range of phenotype architectures. I made LiMMBo accessible as an open source, python module at `https://github.com/HannahVMeyer/LiMMBo/tree/master/limmbo`. LiMMBo is compatible with the LIMIX package for linear mixed models [Lippert & al., 2014].

The bootstrapping has proven powerful to reduce the computational complexity for estimating the covariance parameters and made the analysis of complex datasets with high trait numbers possible. However, so far, I only examined the complexity and calibration dependent on the size of the overall phenotype set P. Of additional interest would be understanding the (co-)dependence of the bootstrap size s, the number of bootstraps b and the co-sampling of traits c. Based on already simulated datasets, a systematic comparison of the run times, covariance estimates and calibration of different combinations of s and b could be conducted. For each of these combinations, different thresholds for c could be examined.

Much of the attraction of linear mixed models in genetics has been their ability to model complex genetic relatedness. As described by [Kang & al., 2010] and demonstrated in this chapter, simple linear models are not suitable for analysing phenotypes with complex underlying genetic relatedness, whereas linear mixed models with the covariance matrices estimated by LiMMBo are appropriate and possible up

to 100s of traits. Complex relatedness in populations is wide-spread in plant and animal breeding [Bolormaa & al., 2014; Yang & al., 2014], and increasingly common in human bottleneck populations [Tachmazidou & al., 2013]. Furthermore, as the population numbers increase in human genetics, complex cryptic relationship structures are more prevalent [Reich & Goldstein, 2001], meaning that methods such as LiMMBo will be more applicable in the future in human genetics.

Trait-by-trait covariance matrices are useful for a variety of high dimensional big data problems across genomics, from statistical genetics to single cell analysis. The ability to accurately estimate large trait-by-trait covariance matrices using this bootstrap method may be applicable to more domains than GWAS, e.g. many gene expression studies use covariance matrices. Previous work from Schäfer & Strimmer [2005] showed the large gene dimensions coupled with small(er) sample sets means that empirical covariance matrices could not be accurately estimated; other investigators [Ledoit & Wolf, 2004; Furrer & Bengtsson, 2007; Bickel & Levina, 2008] used shrinkage methods to create valid covariance matrices. The work from Teng & Huang [2009] uses subsampling but with strong shrinkage priors to generate the final covariance matrix. By fitting the average to closest true covariance, LiMMBo ensures positive-semidefiniteness of the covariance while avoiding ill-conditioned matrices, which usually introduces large biases in the final use of these models. Thus, covariance estimation based on the method implemented in LiMMBo might be applicable and useful in other areas of quantitative genetics.

The ability to generate large cohorts of well phenotyped and genotyped individuals has forced the development of many new methods in statistical genetics. With the advent of genotyped human cohorts up to 500,000 individuals with over 2,000 different traits [Sudlow & al., 2015], and plant phenotyping routinely in the 1,000s of individuals from structured crosses with 100s of (image-based) phenotypes [Atwell & al., 2010; Yang & al., 2014], new informative and scaleable methods are needed. LiMMBo extends the reach of linear mixed models into this new regime, allowing for new complex genetic associations to be made.

5

LiMMBo applied to multi-trait GWAS in *Saccharomyces cerevisiae*

In the previous chapter, I introduced LiMMBo and showed its calibration and power on simulated datasets. In this chapter, I will explore its utility on a real dataset. Amongst the publicly available studies, such as flowering, defense and developmental phenotypes in *Arabidopsis thaliana* [Atwell & al., 2010] or human blood metabolites [Shin & al., 2014], I found the dataset of 46 quantitative traits in yeast generated and analysed in the study by Bloom and colleagues [Bloom & al., 2013] most suitable for several reasons. First, they investigated the growth of a yeast F2 cross on several different substrates. The genetic architecture of an F2 cross is highly structured, making it an ideal test scenario for a linear mixed model capable of adjusting and profiting from population structure in the sample. Second, the measured phenotypic traits have a broad spectrum of correlation, with highly related phenotypes for metabolically similar compounds to very low correlation between certain chemicals. At the same time, the phenotypic measurements are all obtained by measuring the growth size of the colonies and hence, the variable type and unit is the same across phenotypes. Lastly, the collection and quality control of the data were well described and the data were easily accessible in a user-friendly format. However, as with many studies where multiple measurements per sample are obtained, not all

samples were fully phenotyped.

In the following chapter, I will first describe the data processing and imputation strategy for the yeast phenotypes. I will then show the results of applying LiMMBo and subsequent mtGWAS to the dataset and compare the results to the association obtained from single-trait genome-wide association study (stGWAS). Finally, I will explore the benefits of jointly modelling large numbers of traits in genetic studies.

Like LMM and methods based thereon, LiMMBo requires samples to be fully phenotyped as the model cannot deal with missing values. In order to understand how to deal with missing values in the dataset, it is important to have an understanding of the underlying process generating the missing data [Rubin, 1976]. In general, one can distinguish between three processes, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [Little & Rubin, 2002]. Their formal definitions are based on the data $\mathbf{X} \in R^{N,P}$, the binary indicator matrix $\mathbf{M} \in R^{N,P}$ and ϕ, the (unknown) parameter of the missing data process, i.e. the parameter of the conditional distribution g_ϕ of $\mathbf{M}$ given $\mathbf{X}$. N is the number of observations and P the number of observed variables. Entries in $\mathbf{M}$ take two values, $m_{ij} = 1$ if an observation is missing or $m_{ij} = 0$ if it is observed. The data $\mathbf{X}$ can formally be grouped into $\mathbf{X} = \mathbf{X}_{\text{obs}} + \mathbf{X}_{\text{miss}}$, where $\mathbf{X}_{\text{obs}}$ and $\mathbf{X}_{\text{miss}}$ are the observed and missing parts of the data, respectively. Data are MAR if the distribution of missingness only depends on $\mathbf{X}_{\text{obs}}$

$$g_\phi(\mathbf{M}|\mathbf{X}, \phi) = g_\phi(\mathbf{M}|\mathbf{X}_{\text{obs}}, \phi) \forall\, \mathbf{X}_{\text{miss}}, \phi. \tag{5.1}$$

If the distribution is also independent of $\mathbf{X}_{obs}$,

$$g_\phi(\mathbf{M}|\mathbf{X}, \phi) = g_\phi(\mathbf{M}|\phi) \forall\, \mathbf{X}, \phi, \tag{5.2}$$

the data is MCAR. If, on the other hand, the distribution of missingness is dependent on $\mathbf{X}_{\text{miss}}$, hence

$$g_\phi(\mathbf{M}|\mathbf{X}, \phi) = g_\phi(\mathbf{M}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{miss}}), \phi) \forall\, \mathbf{X}, \phi, \tag{5.3}$$

the data is classified as MNAR. To illustrate these cases, consider an example where there are N colonies of yeast and one wants to automatically detect the size and the density of each colony with a suitable instrument ($P = 2$). If the instrument fails with a constant probability ϕ for any colony independent of the measurement, then the pattern of missing values in the data is MCAR. If the probability that the density measurement is missing changes with the value of the size measurement, but is not

dependent on the density of colonies with the same size, then the data are MAR. In contrast, data are MNAR if the probability of obtaining a density measurement depends on the density of colonies with the same size.

In practice, detecting the missing data mechanism often proves difficult. Testing for MCAR can be done via statistical tests [Little, 1988], but distinguishing between MAR and MNAR cannot be achieved formally as this would require knowledge of the missing values [Little & Rubin, 2002; van Buuren & Groothuis-Oudshoorn, 2011]. However, there are visualisation tools that provide diagnostic plots and approximate measures which can help make assumptions about the missingness mechanism [Templ & al., 2012; Garson, 2015].

When analysing datasets with missing data, there are four general approaches to choose from: i) methods simply based on the complete data, ii) methods based on complete data with weighting procedures, iii) model-based and iv) imputation-based procedures. In the first class, incompletely recorded samples are simply excluded, which is the most easy to implement method, but is inefficient and can lead to major bias, especially if the data is MNAR [Little & Rubin, 2002]. Weighting procedures also exclude incompletely sampled data, but apply a weighting to the recorded samples, where the weights attempt to adjust for the missing data as if it were part of the sample design. Model-based procedures define a model for the observed data and base inference and parameter estimates on the likelihood or posterior distribution of that model. The last class of methods, imputation-based approaches, estimate the missing values based on the observed values and the completed dataset can be analysed by standard methods (an extensive review of the different methods can be found in [Little & Rubin, 2002]). The precise usage of the methods and underlying assumptions will be dependent on the missing data mechanism.

I found the imputation approach most applicable for dealing with the missing phenotype values in the yeast dataset as they were simple to apply, did not lead to a decreased sample size and possible loss in power (as method i would have) and did not require recasting the model underlying LiMMBo (as would have been required for method iii). There are a vast number of imputation methods available, which can be categorised by both the method for imputation and the number of times the missing values are imputed. Methods include simple mean prediction, where the missing data for a given variable is replaced by the mean of all known values of that variable and derivations thereof such as KNN or FKM, which use the mean of the k-nearest neighbours to replace the missing values [Troyanskaya & al., 2001; Li & al., 2004]. Instead of imputing based on the mean, i.e. the centre of a distribution, other

strategies use random draws from a predictive distribution of plausible values of the missing value, where the predictive distribution is conditioned on the observed data. These techniques can then be used to either impute one value for each missing item (single imputation) or more than one value to account for imputation uncertainty (multiple imputation) [Little & Rubin, 2002]. For complex datasets, multiple imputation has emerged as the method of choice [Rubin, 1987; Schafer, 1997].

5.1. Dataset and imputation

The dataset generated by Bloom & al. [2013] consists of phenotype and genotype data of 1,008 prototrophic haploid *Saccharomyces cerevisiae* segregants derived from a cross between a laboratory strain (BY MATa) and a wine strain strain (RM MATα). In brief, the segregants were generated by mating of the haploid parental strains and subsequent sporulation of the diploid heterozygote. Sporulation resulted in 1,008 four-spore tetrads that showed 2:2 segregation of mating type and drug-resistance markers. From each tetrad one spore was selected for further analyses (figure 5.1).

For phenotyping, these segregants were grown on agar plates in 46 growth conditions. These can broadly be grouped into growth on different carbohydrates or derivatives thereof (lactose, lactate, raffinose, maltose, mannose, sorbitol, trehalose, xylose, galactose), growth on different culture media (YPD, YNB) with different pH (YNB:pH3, YNB:pH8) or in different temperatures (YPD:4C, YPD:15C, YPD:37C), growth on different antibiotics and xenotbiotics (e.g. cadmium chloride, neomycin, zeocin, cis platin). For a full list, see labels in figure 5.4. After incubation for 48h, the colony size of each segregant grown in the different conditions was measured. The final phenotypes were defined as the colony size normalised to colony size growth on control medium. For the remainder of this chapter, a trait is defined as this normalised growth size in one condition. Out of the 1,008 segregants, 303 segregants were phenotyped for all 46 traits.

Segregants were genotyped using Illumina short-read sequencing. After mapping, quality control and filtering for unique genotype markers, all 1,008 segregants were genotyped 11,623 unique genotypic markers.

5.1.1. Missing data mechanism

In order to gain an understanding of the dataset, I first looked at the frequencies and distribution of missing values. There are 135 different combinations of missing values across the samples and the missing phenotypes are not evenly distributed

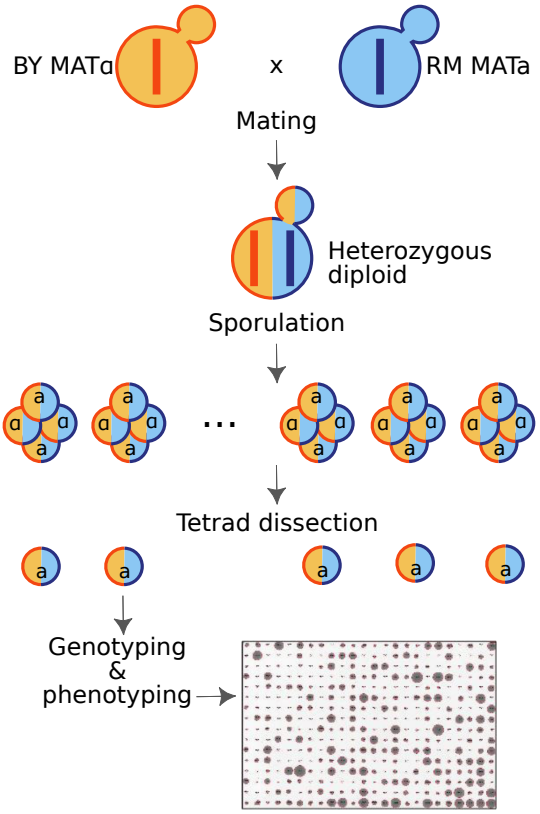

Figure 5.1: Generation of yeast dataset. Haploid parental strains BY MATa andRM MATα were mated to generate diploid heterozygotes. These diploid heterozygotes were sporulated, during which they undergo meiosis and yield tetrads of recombinant haploids. From each tetrad, one spore was selected. For phenotyping, these segregants were grown an agar plates in different conditions. Adapted from [Bloom & al., 2013].

(figure 5.2A). Some traits such as cobalt chloride are present for almost all samples while others such as sorbitol or raffinose are missing in more than a third of the samples. I used Little's global test for MCAR to analyse whether these observed data patterns can be accounted for by a MCAR mechanism. Little's method tests the null hypothesis that the data is MCAR [Little, 1988; Beaujean, 2015], which can in this case be rejected with a p-value of 2×10^{-34} (based on a χ^2 distribution, χ^2 =5,902, df =4,631).

Determining if data is MAR or MNAR cannot be tested for formally and relies on approximate measures and assumptions based on the experimental procedures [Schafer & Graham, 2002; Garson, 2015; Templ & al., 2012]. Garson [2015] suggests to use significance tests of missingness. If it can be demonstrated that one or more variables in the dataset are significantly correlated with missing values, missingness may be predictable, which is the requirement for imputing MAR data. In order to test for predictable missingness, I created an indicator matrix for the phenotype matrix, where observed values were encoded as zero and missing values as one. For each of the 46 traits in the dataset, I correlated the observed values across all samples with each column of the indicator matrix, i.e. the missingness patterns per trait. If all values were observed for a given trait, all values in the indicator matrix in this column were equal to zero and the correlation between the trait and the missingness was set to NA. Figure 5.3 shows the correlation patterns between the phenotypes and the missing values per trait. For traits like cobalt chloride and magnesium sulfate, where little data is missing, many entries are NA. Overall, for a number of traits and missingness patterns, there is sufficient evidence for predictable missingness and MAR assumptions for further analyses were considered valid. Most importantly, for data with MAR, the missing data mechanism is ignorable for maximum likelihood based methods and no further adjustments for the mechanisms have to be made in the modelling [Rubin, 1976; Little, 1988]. Thus, the MAR assumption of missingness in the yeast data allows for imputation via the likelihood-based method of multiple imputation and LMMs.

5.1.2. Imputation via MICE

Imputation of missing values requires an understanding of which missing trait values can be reliably imputed and to find the best parameter settings for the imputation. In order to do this, I needed a fully phenotyped dataset with the same structure as the yeast dataset, where missing values could be introduced, imputed and subsequently compared to the true values. I chose a simple approach using the subset

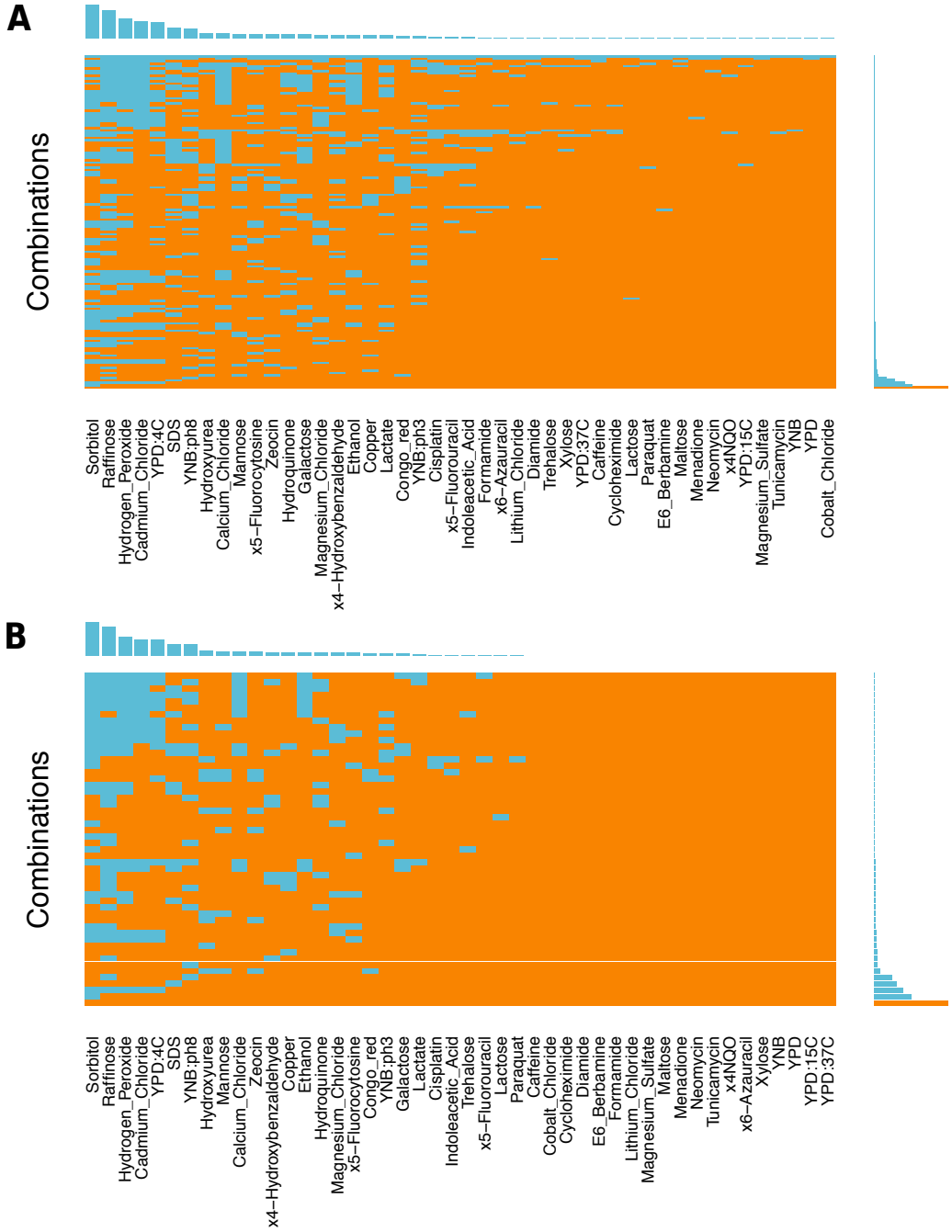

Figure 5.2: Frequencies and distributions of missing values in the yeast phenotype data. In both panels, the aggregation plot (middle) depicts all existing combinations of missing (blue) and non-missing (orange) values in the traits. The bar chart on its right shows the frequencies of occurrence of the different combinations. The histogram on the top shows the frequency of missing values for each trait (R Package: *VIM* [Templ & al., 2012]). A. The full dataset contains normalised colony sizes for growth in 46 different conditions of 1,008 genotyped yeast segregants. 306 segregants are fully genotyped (bar chart, orange bar). B. Fully-phenotyped dataset of 306 segregants with simulated missing values based on the observed missingness pattern for the entire pool of 1,008 segregants. Generated via R function *VIM::aggr*.

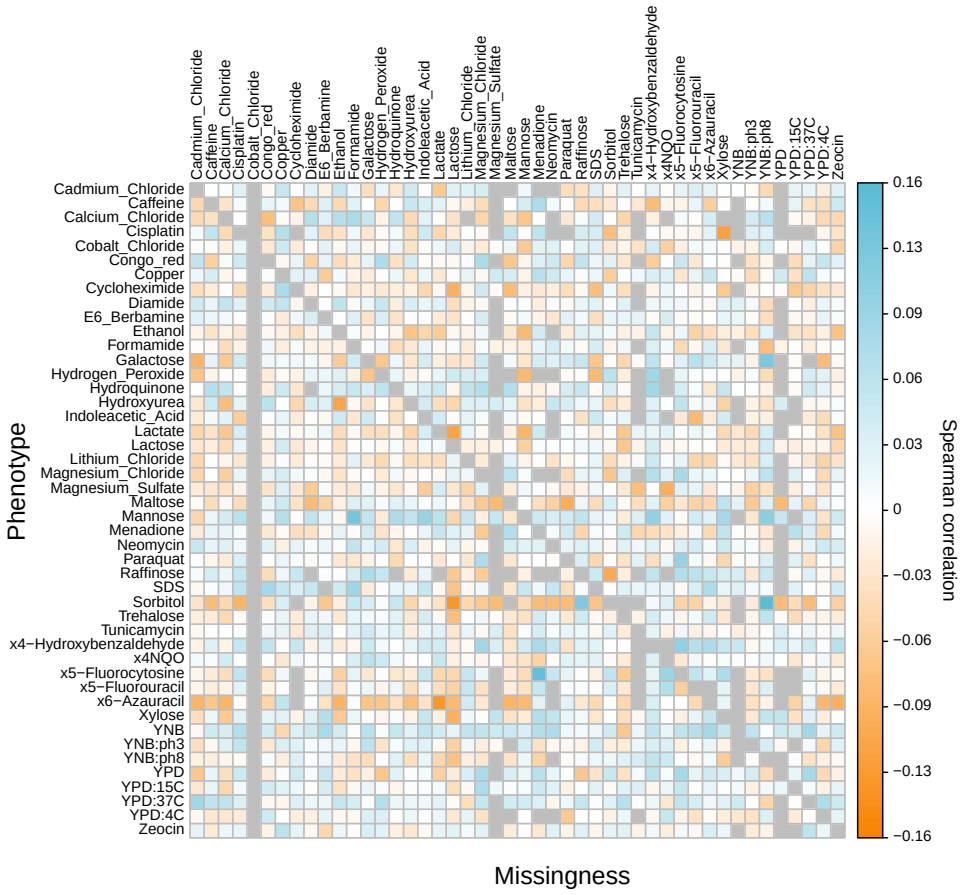

Figure 5.3: Correlations of observed phenotypes with missing data values. For each of the 46 traits, the Spearman's rank correlation coefficient ρ was computed with each column of the indicator matrix of the phenotypes, containing zero for observed values and one for missing values. The strength and the direction of correlations are depicted above, with the original phenotypes in rows and the indicator matrix of the phenotypes across columns. Grey squares indicate NA, i.e. columns in the indicator matrix for which no traits were missing when correlated with the observed values for a given trait. Generated via R function *corrplot::corrplot*.

of the 303 fully phenotyped samples and introducing missing values with a similar pattern of missingness as observed in the original dataset. The results for the real (figure 5.2A) and simulated (figure 5.2B) dataset are similar in terms of frequencies and combinations of missing/non-missing traits. I used this simulated dataset as input to the imputation framework based on multiple imputation by chain equations (MICE) [van Buuren & Groothuis-Oudshoorn, 2011].

MICE belongs to the general class of multiple imputation frameworks, where several imputed versions of the dataset are generated and each variable is imputed separately. The imputed values are chosen from plausible values drawn from a distribution that is specific for each variable, in this case for each trait. This distribution is derived from the dataset $\mathbf{X} \in R^{N,P}$ itself, with X split into missing and observed parts $\mathbf{X} = (\mathbf{X}_{\text{miss}}, \mathbf{X}_{\text{obs}})$, the binary indicator matrix for missingness $\mathbf{M} \in R^{N,P}$ and a set of predictor variables Z. The MICE algorithm is usually divided into four steps [Rubin, 1987; Van Buuren & Oudshoorn, 1999; Pigott, 2001]:

1. Specify the posterior predictive density $p(\mathbf{X}_{\text{miss}}|Z, \mathbf{M})$ given the non-response mechanism $p(\mathbf{M}|\mathbf{X})$ and the complete data model $p(\mathbf{X})$.
2. Draw imputations from this density to produce m complete data sets.
3. Perform m complete-data analyses on each completed data matrix.
4. Pool the m analysis results into final point and variance estimates.

Garson [2015] approach allows me to obtain reliable imputation estimates while having to estimate the variance components via LiMMBo only once. As described in the previous chapter, LiMMBo strongly reduces the computation time for the variance decomposition (section 4.4), but it is still the time consuming factor in the analysis.

The two main choices when applying MICE for imputation have to be made in step one: the type of the imputation model and the choice of predictor variables.

Imputation model. From the different imputation models available (examples described in [van Buuren & Groothuis-Oudshoorn, 2011]), I found predictive mean matching, a semi-parametric method which preserves non-linear relations in the data [Little, 1988; van Buuren & Groothuis-Oudshoorn, 2011], a fast and sensible imputation option. In brief, predictive mean matching finds the mean and covariance of the multivariate distribution $\mathbf{X}$ with missing values (often simply based on the complete cases). Subsequently, for each incomplete sample it predicts the missing values $\mathbf{X}_{\text{miss}}$ based on $\mathbf{X}_{\text{obs}}$ and the provided predictor variables Z. In addition,

values of the complete samples for the same set of $\mathbf{X}_{\text{miss}}$ are predicted. The predicted values of the incomplete sample are than matched to the predicted values of the complete samples and the closest match is chosen. The imputed values for the incomplete sample are set to the observed values of the closest match [Little, 1988]. In this way, only realistic and theoretically observable values (assuming proper quality control of the data prior to imputation) are imputed.

Predictor variables. Collins & al. [2001] show that as many valid predictor variables as possible should be included in the imputation to obtain the least amount of bias and maximal certainty about the predictions. In addition, Schafer [1997] demonstrated that using this strategy makes MAR assumptions more plausible. However, not all predictors will be relevant and the choice of predictors can be done on a per-variable level. In order to select suitable predictors for each trait, I first computed the pairwise Spearman correlation coefficient ρ for all traits across the 303 fully-phenotyped segregants. Some of the traits like cadmium chloride or neomycin show very little correlation to any of the other traits, while many of the traits based on growth on different carbohydrate resources form a large cluster of moderate to strong correlation (figure 5.4). I tested several sets of predictor variables, either using all traits as predictors or choosing predictors based on the pairwise ρ of the traits. For each trait, I included predictors that showed a correlation higher than a predefined threshold ($\rho = \{0.1, 0.2, 0.3\}$). In addition, I restricted the predictors to traits that had been measured in at least 20% of the samples in the dataset. This excluded cadmium chloride (21% missing), hydrogen peroxide (24%), raffinose (34%), sorbitol (41%) and YPD:4C (20%) as predictor variables, but did not prevent them from being imputed.

Further parameters for MICE are the number of imputed datasets m (set to $m = 20$) and the number of iterations $maxit$ (set to $maxit = 30$). For each predictor set-up, I initiated MICE with the same seed for the random number generator to ensure comparability. After imputation, I evaluated the goodness of the imputation by computing the Spearman correlation of the imputed values (averaged across iterations m) to the experimentally observed ones (figure 5.5). Traits where the imputed values correlated to the original ones by more then 95% in at least one of the predictor set-ups were retained in the analysis. For five traits (cadmium chloride, hydrogen peroxide, raffinose, YNB:ph8, YPD:4C), no suitable predictors could be determined and these were excluded from further analyses (figure 5.5, red labels). For each trait, I chose the predictor scheme that yielded the highest correlation between the im-

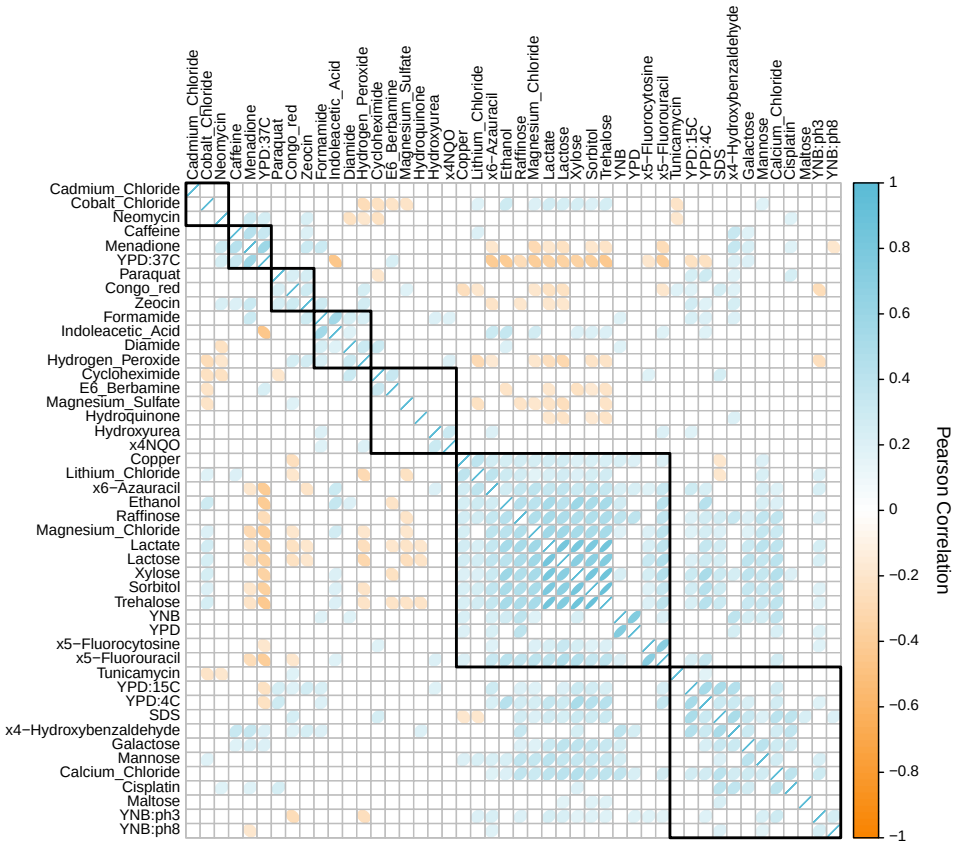

Figure 5.4: Pair-wise correlations of 46 growth traits in *Saccharomyces cerevisiae*. For each trait pair, Pearson's correlation coefficient ρ and the p-values of the correlation were computed. The p-values were adjusted for multiple testing according to Benjamini and Hochberg's method [Benjamini & Hochberg, 1995]. The strength and the direction of significant correlations ($FDR < 0.05$) are depicted above. Non-significant correlations are left blank. The traits are clustered based on complete-linkage clustering of $(1 - \rho)$ as distance measurement and the largest clusters are indicated by black squares. Generated via R function *corrplot::corrplot*.

puted and observed data for the imputation of the missing values in the full dataset. Missing values were imputed in segregants that were phenotyped for at least 80% of the traits. The final dataset contained 981 segregants with phenotypes for 41 traits each.

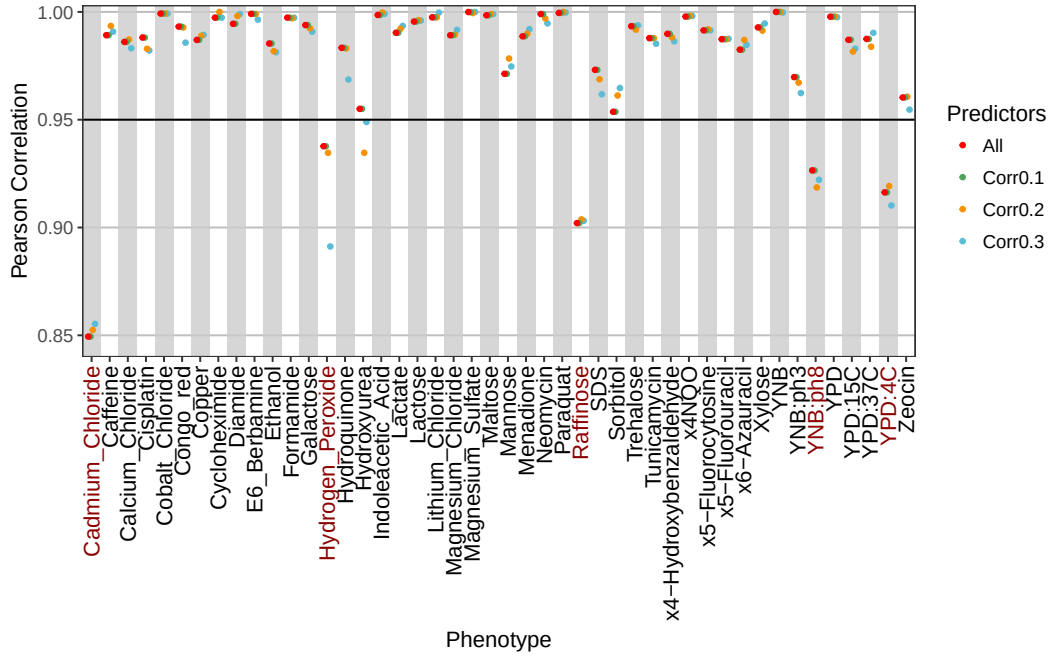

Figure 5.5: Correlation between imputed and experimentally observed trait values. In the subset of 306 fully phenotyped samples, missing values were introduced and subsequently imputed via MICE. Different predictor sets were tested based on Pearson's correlation coefficient: traits were considered predictors if their correlation with the target trait was greater than a given threshold. For each predictor setup (all traits as predictors and predictors passing the correlation threshold $\rho = \{0.1, 0.2, 0.3\}$), $m = 20$ imputed datasets and $maxit = 30$ iterations of MICE were conducted. The goodness of the imputation was evaluated by computing the correlation of the imputed values (averaged across iterations m) to the experimentally observed ones. Traits with at least one correlation greater than the 0.95 threshold (black vertical line) were retained in the dataset. For traits labelled in red, the imputation was considered to be unreliable and the traits were excluded from further analyses.

5.2. Multi-trait GWAS with LiMMBo

In order to show the utility of LiMMBo for joint high-dimensional phenotype analyses and to demonstrate the advantages over single-trait approaches, I analysed the imputed dataset both with stGWAS and mtGWAS.

5.2.1. Estimating the genetic relationship in the yeast cross

For both analyses, I used a LMM where the sample-by-sample component of the random genetic effect is based on the RRM. To obtain an estimate of the RRM I first pruned the genome-wide SNPs (11,623) for SNPs that are in LD within a window of 3kb and show a correlation $r^2 > 0.2$. As the dataset is based on an F2 cross, LD structure estimation is not straight-forward and this window size is a simple estimate derived from a study on the population genomics of domestic and wild yeasts [Liti & al., 2009]. The LD pruning reduced the SNP set for RRM estimation to 4,105 SNPs. The RRM was estimated using the method introduced by Yang & al. [2011] (section 1.7.6). *PLINK* [Chang & al., 2015] was used for both LD pruning (with parameters *–indep-pairwise 3kb 5 0.*) and RRM estimation (with parameters *–make-rel square gz*).
For the genotype to phenotype mapping the full set of 11,623 SNPs was used.

5.2.2. LiMMBo increases power in detecting genetic associations

The first step in the mtGWAS is the trait-by-trait covariance estimation via LiMMBo. 1,000 bootstraps of 10 traits each were run and their trait-by-trait covariance estimated. The combined trait-by-trait covariance estimates $\mathbf{C}_g$ and $\mathbf{C}_n$ were used as input estimates for the second step in the mtGWAS, the mvLMM (equation (1.42)) across all genome-wide SNPs. I used a mvLMM with a trait-design matrix corresponding to the any effect test, i.e. testing for an effect of each SNP on any of the traits compared versus the null hypothesis of no association (section 1.7.8).

For the stGWAS, the trait-by-trait components of the random effects are point estimates (σ_g and σ_n) derived within the LMM framework and do not require *a priori* estimation. The stGWAS was performed for each trait separately, applying univariate LMMs (equation (1.30)) to test the effect of a SNP on each individual trait. To account for the number of univariate tests, the p-values obtained from the stGWAS were adjusted for multiple testing by the effective number of conducted tests M_{eff}. M_{eff} was introduced by Galwey [2009] and adjusts for multiple testing in a manner

similar to the Bonferroni method (section 1.7.4, [Dunn, 1961]). However, it is less conservative, as it does not adjust for total number of tests, but the estimated, effective number of tests, taking correlation between the variables and tests into account:

$$M_{\text{eff}} = \frac{(\sum_{i=1}^{M} \sqrt{\lambda_i})^2}{\sum_{i=1}^{M} \lambda_i}, \tag{5.4}$$

where λ are the eigenvalues of the phenotypes' correlation matrix. To adjust for multiple testing in the stGWAS, the p-values are multiplied with M_{eff} and set to one if the multiplication leads to values greater than one. M_{eff} for the 41 growth traits was estimated to be 33.

In order to compare the single-trait and multi-trait analyses, I followed approaches of previous association studies in yeast crosses [Brem & al., 2002; Brem & Kruglyak, 2005; Ehrenreich & al., 2010], where permutations were used to estimate empirical FDR levels. With a conservative, theoretical threshold of $p_t = 10^{-5}$, at most one SNP is expected to be false positive in a total of $s = 11,623$ SNPs. To find the empirical FDR corresponding to this threshold, I generated $k = 50$ permutations of the genotypes and fitted the LMMs against these permutations. These p-values were used as the empirical p-value distribution and for $p_t = 10^{-5}$, the empirical FDRs estimated as $\text{FDR}_{\text{mtGWAS}} = 1.2 \times 10^{-5}$ and $\text{FDR}_{\text{stGWAS}} = 8.6 \times 10^{-6}$.

Figure 5.6 shows the manhattan plot of the multi-trait and single-trait GWAS. On several chromosomes (e.g. chr1, chr6 and chr15), mtGWAS peaks (blue) are observed whereas no stGWAS peaks (orange; minimum p-value per SNP across all 41 stGWAS, adjusted for multiple testing) can be detected. On the other hand, there are a few loci for the stGWAS where the multi-trait analyses either does not pass the FDR threshold (e.g. on chromosome 7) or does not detect any association (e.g. on chromosome 4). For these loci, the underlying genetics seem to be trait specific to magnesium sulfate and hydroquinone, respectively (figure B.2 in the appendix). Testing with a 41 degrees of freedom test as done in the mtGWAS hinders the detection of these strong mono-trait associations (compare distributions in figure 1.2) and confirms previous studies showing that the single-trait model for uncorrelated traits is more powerful [Korte & al., 2012]. Both, the gain in power for the multi-trait associations and the burden of a multivariate test when the underlying effect is univariate confirm the results obtained from the theoretical power analysis (section 4.7).

To quantify the increase in power, I counted the number of SNPs detected above the permutation-based thresholds for both the stGWAS and the mtGWAS. Since the number of SNPs per locus is not constant (based on LD structure in the F2 cross and

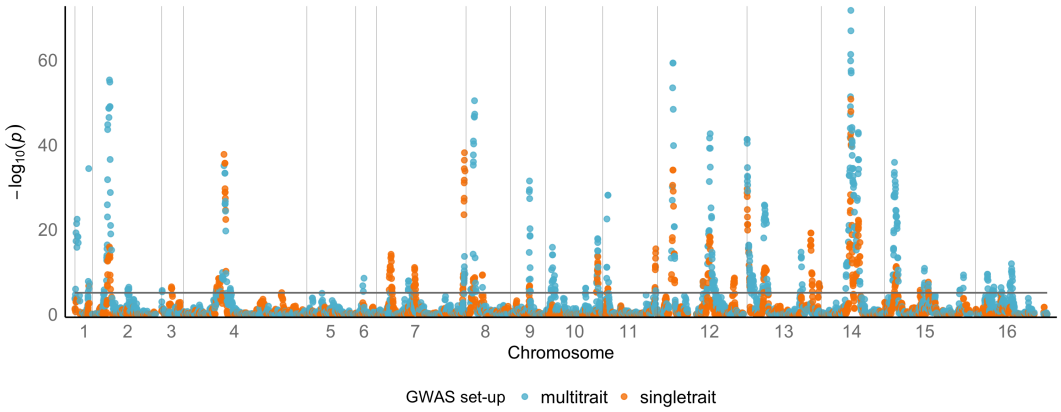

Figure 5.6: Manhattan plot of p-values from single-trait and multi-trait GWAS. The stGWAS p-values were adjusted for multiple testing by the effective number of tests ($M_{\text{eff}} = 33$) and only the minimum adjusted p-values across all 41 traits per SNP are shown. The threshold line is drawn at the empirical $\text{FDR}_{\text{stGWAS}} = 8.6 \times 10^{-6}$.

genotyping parameters), I needed a locus-based rather than a SNP-based count for a fair comparison of the two methods. In order to filter SNPs based on locus, I used *PLINK* for LD pruning of the SNPs, choosing a strict threshold of $r^2 > 0.2$ and increasing LD window sizes ranging from 3 to 100kb. The maximal LD window of 100kb covers between 6% (chromosome 4) and 43% (chromosome 1) of total chromosome length (ScerevisaeR64-1-1, ensembl release 90, [Aken & al., 2016]). Table 5.1 shows that the increase in power is present from narrow to broad LD pruning, with on average 29% more loci in mtGWAS.

Table 5.1: Comparison of loci detected in single-trait and multi-trait GWAS. In the column "All SNPs", the absolute number of SNPs beyond the FDR threshold for multi-trait and single-trait GWAS as well as their ratio (multi-trait:single-trait) are depicted. In order to limit the potential bias in the counting of the loci, introduced by different degrees of LD for different loci, the genome-wide SNPs were LD pruned and the ratio of associated SNPs determined for five different LD window sizes.

	All SNPs	LD pruned with $r^2 \geq 0.2$				
		3kb	10kb	30kb	50kb	100kb
NrSNPs	11,623	4,105	1,028	264	161	107
multitrait	1,132	384	101	24	15	9
singletrait	695	275	72	20	13	7
multitrait:singletrait	1.63	1.4	1.4	1.2	1.15	1.29

5.2.3. Multi-trait effect size estimates as indicators for common biology

As well as providing an increase in power, the mtGWAS inherently provides effect size estimates across all phenotypes for a particular locus, allowing for a richer exploration of pleiotropic effects of each of locus. To analyse the relationship between traits and SNPs based on their effect size estimates, I filtered the genome-wide SNPs for SNPs that fell within a gene body and pruned these 8,135 SNPs for SNPs in LD with $r^2 > 0.2$ and within a 3kb window (1,412 SNPs). Lastly, I filtered for SNPs passing the FDR $= 10^{-5}$ yielding 210 SNPs across 15 out of the 16 yeast chromosomes. Chromosome 5 is the only chromosome without associated SNPs in the single-trait and multi-trait GWAS (figure 5.6).

To find groups of SNPs and traits with similar effect size estimates, I clustered the effect size estimates of these SNPs both across traits and SNPs. (figure 5.7). I used the hierarchical clustering algorithm *pvclust* that provides bootstrap-based p-values as a measure for the stability of a given cluster [Suzuki & Shimodaira, 2006]. The clustering was based on their Pearson correlation coefficients, with 50,000 bootstraps for traits and 10,000 for SNPs. Clusters with $p < 0.05$ were considered stable. A heatmap of effect size estimates and the clustering results is depicted in figure 5.7. Ignoring the clustering for a first impression of the results, one can clearly see that most SNPs have non-zero effects in more than one trait (figure 5.7, strong signals across columns). Furthermore some traits have contributions from across the genome, many of which are xenobiotic growth conditions e.g. zeocin [Krol & al., 2015] and neomycin [Foiani & al., 1991]. Turning to the clustering, figure 5.7 (dendrograms) shows that the clusters are driven by specific combinations of loci and traits, and would be hard to achieve from a single-trait analysis.

There are a number of stable clusters of traits (figure 5.7, blue branches in the row dendrogram), including classically linked carbon metabolism sources (lactose, lactate and ethanol), and other clusters for which there is literature support. For example, expression of genes involved in DNA replication has been shown to change upon treatment with hydroxyurea and 4-nitroquinoline-l-oxide (x4NQO) [Elledge & Davis, 1990], two substances that are linked in this analysis by forming a stable cluster. A study demonstrating trehalose and sorbitol to have synergistic effects on viability in yeast [Hua & al., 2015] demonstrating a biological link of these sugars forming a cluster. For other clusters, such as SDS and Hydroxybenzaldehyde or magnesium sulfate and berbamine I was unable to find literature support. However, these could serve as candidate clusters for further investigation of growth phenotypes.

I discovered 31 stable SNP clusters (figure 5.7, blue branches in the column dendrogram), many of which represent linked loci. However, there are nine clusters (figure 5.7, grey boxes) spanning multiple chromosomes, and many clusters linking disjoint regions across a chromosome. Some SNP clusters have suggestive common annotation, such as *cluster a* which has two members of the nuclear pore complex (NUP1, NUP188), and *cluster b* which has a common set of vesicle associated genes (ATG5, PXA1,VPS41; figure 5.7, labelled boxes). The small size of the clusters prevented any systematic gene ontology based enrichment. Nevertheless, the ability to explore clusters of both traits and genetic loci demonstrate the utility of mtGWAS for hypothesis generation.

5.3. Summary

A particular benefit of LMMs is that complex genetic relationships can be modelled, which is useful in structured populations such as this F_2 cross in yeast. In univariate LMMs, the kinship information is used to account for background genetic effects in associations with a single trait. When used in mvLMMs, the kinship structure allows for the estimation of complex trait covariance structure. However, it is only possible through a combination of appropriate phenotype imputation and a method like LiMMBo to efficiently map all 41 growth traits together in order to investigate pleiotropic effects on a genome-wide level. I demonstrated that such a multivariate analysis through LiMMBo is more powerful in detecting genetic associations in a real dataset, than univariate tests. While the focus of this chapter was to demonstrate the applicability and power of LiMMBo, it also highlighted the potential of multivariate analysis for gaining insights into the underlying biology of pleiotropic loci. The effect sizes estimated by the multivariate LMM provide the relevant data to study shared pathways and regulation and can help to generate hypotheses for future research.

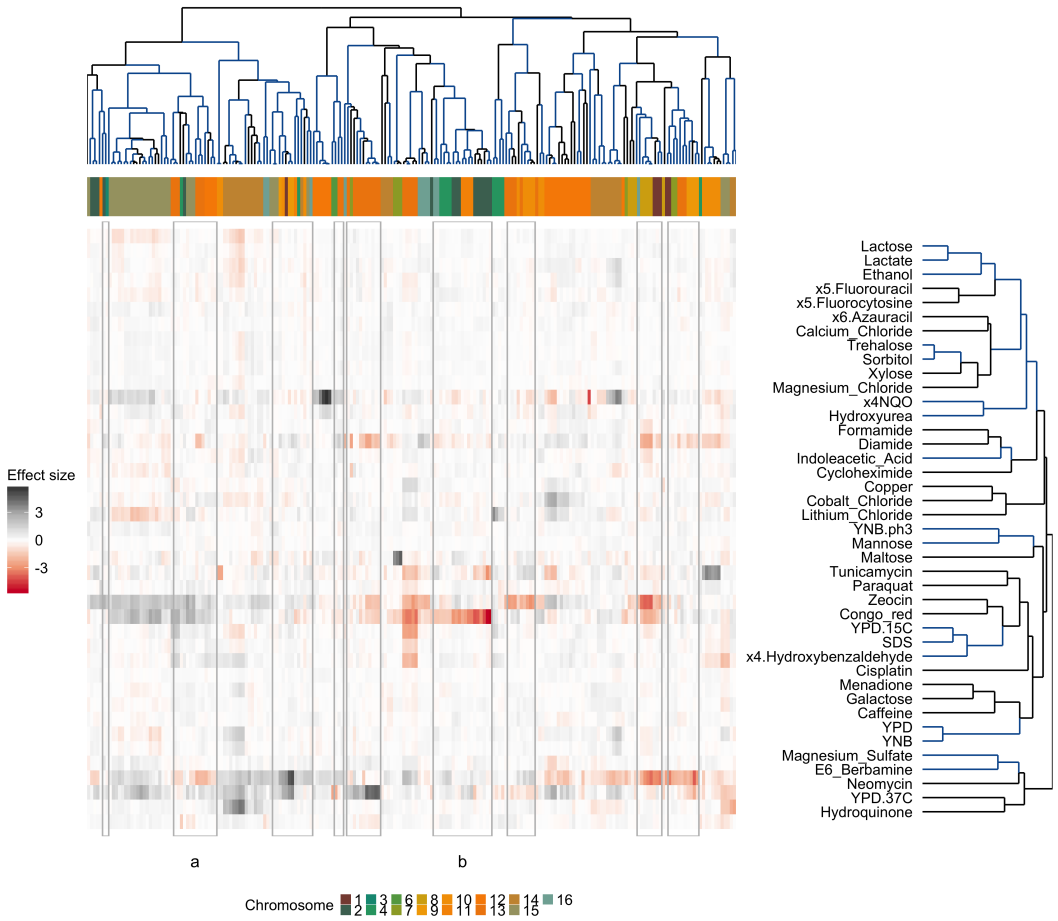

Figure 5.7: Hierarchical clustering of mtGWAS effects size estimates. Effect size estimates of LD pruned (3kb window, $r^2 > 0.2$), trait-associated SNPs located within a gene body were clustered by loci and traits (both hierarchical, average-linkage clustering of Pearson correlation coefficients). Stable clusters (pvclust $p < 0.05$) are marked in blue. Grey boxes indicate stable SNP clusters spread across at least two chromosomes. a and b label two clusters for which suggestive common annotation was found, for details see text.

6

Low-dimensional representations of very high-dimensional data

In chapter 4 and chapter 5, I developed and applied methods for the multivariate analyses of hundreds of traits. When evaluation the suitability of LiMMBo on the simulated datasets, I considered each trait as a separate, but correlated measures and used all traits for the multi-trait genotype to phenotype mapping. I used the same strategy for the application of LiMMBo to the growth traits of yeast. However, as described in chapter 3, the simulated phenotypes are generated by adding different phenotype components, and one could argue that depending on the analysis, it might proof useful to extract relevant features representing different phenotype components prior to the multivariate analyses across all traits. For instance, given very large numbers of measurements or traits, feature extracting will reduce the number of traits and therefore the degrees of freedom for the multivariate analyses. In the following chapter I will describe different methods to achieve the feature extraction by dimensionality reduction approaches. I will present two case studies that show how these approaches can be used for visualisation of high-dimensional data. Beyond that, I will demonstrate in simulations that they can not only be used for visualisation but also as valid proxy phenotypes for genetic association studies. These simulation results build the basis for the genetic association study on 3D heart

phenotypes described in chapter 7.

In biological and medical research, samples are often phenotyped for more than one trait. These traits can either be different attributes of the same underling phenotype or more independent features. In the former case, multiple phenotypes can be related measurements such as length, width and circumference of plant leafs or measurements commonly regarded as covariates such as sample height and weight. For image-based and molecular phenotyping methods, the measured traits can be a mixture of independent features and attributes of the same phenotype. For instance, in computed tomography scans, functional MRI or high-resolution microscopy, each pixel or voxel can be considered a different measurement. Groups of these describe different morphologies (features) and pixels/voxels within each group can be considered attributes of that feature. In molecular phenotyping such as gene expression or metabolite profiling, several hundred or thousand measurements are collected simultaneously. Here, the classification into features and attributes is more difficult, considering the complex structure of gene expression networks and gene regulation. In many of these cases, neither the number of attributes nor the number of independent features are know. However, when analysing these large datasets, one is often interested in extracting meaningful variables from the data or compressing the data into a more tractable number of variables. These approaches rely on the assumption that the lower number of variables are a good representation of the true complexity of the dataset. In other words, one assumes that the high-dimensional datasets occupy an intrinsically lower-dimensional space (manifold) which is embedded in the observed, high-dimensional space. Low dimensional representations of gene expression measurements might reflect common pathways or transcriptional profiles and image-derived phenotypes could reflect organ shape variation, disease status or functional MRI activity scores. For a high-dimensional dataset $\mathbf{X}$ with N samples and P dimensions (traits), dimensionality reduction techniques aim i) to provide a meaningful low-dimensional representation $\mathbf{Z}$ of K dimensions while only losing minor amounts of information:

$$\mathbf{X} \in \mathcal{R}^{N,\,P} \xrightarrow{\text{DimReduction}} \mathbf{Z} \in \mathcal{R}^{N,\,K}, \tag{6.1}$$

ii) to use only a small number of free parameters and iii) to preserve the quantities of interest in the data. Depending on the algorithm employed, these might be local proximity or global structure.

There are a variety of approaches for dimensionality reduction with different un-

derlying mathematical concepts and parameters and choosing the most appropriate method for a given dataset is not trivial. Fundamentally, the problem is finding an objective criterion of what a good dimensionality reduction method is.

In the following, I will first present a small review of current dimensionality reduction methods. I will use these methods to demonstrate the application of dimensionality reduction for visualisation on small datasets with known structure. I will compare the visual results to two published criteria for measuring the quality of dimensionality reduction in terms of neighbourhood-similarities in the low- and high-dimensional space. Then, I will describe the results of the different dimensionality reduction techniques on simulated high-dimensional datasets and propose an additional stability criterion which aids in choosing the dimensionality of the lower-dimensional phenotype space. Finally, I show that low-dimensional representations of the phenotypes can capture underlying genetic structure. The methods and criteria used in this chapter will be applied on clinically relevant high-dimensional heart morphology data in chapter 7.

6.1. Review of dimensionality reduction methods

The earliest dimensionality reduction techniques were two linear methods based on spectral decomposition: PCA and classical multi-dimensional scaling (MDS).

The general concept of PCA was described by Pearson in 1901 [Pearson, 1901]. In [1933], Hotelling was the first to describe it as a method for dimensionality reduction. In PCA the components of the new phenotype representation are the PCs and are the eigenvectors $\mathbf{W}$ of the empirical covariance matrix $\mathbf{C}$: $\mathbf{C} = \mathbf{X}\mathbf{X}^T = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$. The eigenvalues in the diagonal matrix $\mathbf{\Lambda}$ corresponding to the PCs are equivalent to the variance explained by their components. The transformation of the phenotype data into PCs leads to a projection where the highest amount of phenotypic variance explained lies in the first component, the second highest variance in the second component and so on. The dimensionality reduction is achieved by using the first K PCs until the cumulative sum of the eigenvalues reaches a predefined threshold of total phenotypic variance that should be retained: $\mathbf{Z} = \mathbf{W}_1, \dots, \mathbf{W}_K$.

MDS was introduced by Gower [1966], motivated by his dissatisfaction about the overuse of PCA in biology. MDS is based on the spectral decomposition of a dissimilarity matrix $\mathbf{D}$ between the samples in $\mathbf{X}$. Classical MDS finds the low-dimensional representation $\mathbf{Z}$ whose pairwise distance matches the dissimilarity d_{ij} of the original data: $d_{ij} \approx \hat{d_{ij}} = \|\mathbf{z}_i - \mathbf{z}_j\|$. $\mathbf{Z}$ can be found by the eigendecomposition of the

squared dissimilarity matrix $\mathbf{D^2} = \mathbf{V\Lambda V}^{\mathrm{T}}$, where $\mathbf{Z} = \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^T$. As in PCA, $\mathbf{Z}$ will be ordered with the components explaining most variance ranked first and dimensionality reduction can be achieved by selecting the first K vectors [Gower, 1966]. MDS finds an embedding that preserves the inter-point distances and is equivalent to PCA when those distances are Euclidean.

Several decades after the introduction of PCA as a means of linear dimensionality reduction, Schoelkopf & al. [1998] and colleagues proposed its non-linear extension based on the transformation of $\mathbf{X}$ into a feature space $\mathbf{F}$ via the mapping function Φ. Instead of finding the eigendecomposition of the covariance matrix of $\mathbf{X}$, the aim is the diagonalisation of the covariance $\mathbf{K}$ of the features of the data $\Phi(\mathbf{X})$: $\mathbf{K} = \Phi(\mathbf{X})\Phi(\mathbf{X})^T$. Using the kernel representation $k(\mathbf{x}_i, \mathbf{x}_j) = (\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j))$ to compute the dot products of $\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)$ allows computation of the dot product in $\mathbf{F}$ without having to carry out the map Φ. This technique is commonly referred to as the kernel trick and yields the feature covariance matrix $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)_{ij}$. The normalised eigenvectors of $\mathbf{K}$ are used to extract the PCs. This kernel principal component analysis (kPCA) approach allows for non-linear feature extraction, whilst the possibility to select different kernels (e.g. gaussian or sigmoid) makes it applicable for a wide range of cases when non-linearity is assumed [Schoelkopf & al., 1998].

PCA, kPCA and MDS build the basis for many other dimensionality reduction techniques. Notably, Ham and colleagues show that the class of kernel-eigenmap-based dimensionality reduction methods, such as Isomap, Locally linear embedding (LLE) and Laplacian Eigenmaps can be understood as a variant of kPCA with different kernel matrices. Methods of this class are described in detail later, but common to all these methods is the aim to obtain a global representation of the data $\mathbf{X}$ by using information about local interactions between the data points in $\mathbf{X}$. The data points are represented as the nodes of a symmetric graph, whose kernel function k describes a local geometry of $\mathbf{X}$. The graph specified by $(\mathbf{X}, k)$ is used to construct a square matrix $\mathbf{M}$, which describes the transitions on the graph as a Markow chain. Using this Markov matrix $\mathbf{M}$, one can map the data into a lower dimensional Euclidean space. The difference of the algorithms lies in the definition of the neighbourhood structure and the means to find a global embedding. Table 6.1 summarises these and other commonly used linear and non-linear techniques and the list below gives a short summary of the mathematical principles.

1. **Probabilistic estimation of expression residuals**: PEER implements factor analysis methods to estimate variance components in $\mathbf{X}$. The model assumes additive effects from independent sources that influence $\mathbf{X}$ and aims at estim-

Table 6.1: Dimensionality reduction methods. The different dimensionality reduction techniques can distinctly be classified into linear and non-linear types. The methods column broadly groups techniques based on their main mathematical concept and parameters gives the number of parameters that need to be specified for the mathematical model.

Type	Method	Name	Parameters
linear	spectral	PCA	0
		MDS	0
	Factor analysis	PEER	1
	Generative model	ICA	2
non-linear	rank-based	nMDS	2
	PCA-based	DRR	>1
	spectral	kPCA	0
	Kernel eigenmap	Isomap	1
		LLE	1
		Laplacian Eigenmaps	2
		DiffusionMaps	>2
	Probability distributions	tSNE	4

ating these effects in a joint Bayesian inference model. By specifying the only source of variation to be due to unknown effects, PEER can be used to extract latent variables from high-dimensional datasets, where the latent variables are modelled based on a standard normal distribution and are initiated based on PCA of $\mathbf{X}$. The model specifications are complex and the interested reader is referred to the paper describing the details of the methodology [Stegle & al., 2010]. In the most simple scenario, only the parameter of the maximum number of unobserved latent factors, i.e. the column dimensionality K of $\mathbf{Z}$ have to be specified [Stegle & al., 2012].

2. **Independent Component Analysis**: ICA belongs to the class of generative models, which describes how the data $\mathbf{X}$ could have been generated by a process of mixing independent components $\mathbf{S}$ according to a mixing scheme $\mathbf{A}$: $\mathbf{X} = \mathbf{SA}$. Both the independent components and the mixing matrix are unknown. The key to "un-mixing" the signals are the underlying assumptions that the latent components are independent and have a non-Gaussian distribution. In order to find $\mathbf{A}$ and $\mathbf{S}$, ICA finds the un-mixing matrix $\mathbf{U}$ which maximises the non-gaussianity of $\mathbf{S}$: $\mathbf{XU} = \mathbf{S}$. Non-gaussianity can be quantified

by approximating the negentropy of $\mathbf{S}$, i.e. the difference in entropy between a Gaussian random variable of the same covariance matrix as $\mathbf{S}$ and the entropy of $\mathbf{S}$ itself. The parameters to be specified are the threshold for the tolerance at which the un-mixing matrix is considered to have converged and the number of components to be modelled. ICA was first described by Herault & Jutten [1986] and has seen many implementations for finding the maximum of the non-gaussianity (reviewed in [Comon, 1994]), including FastICA [Hyvärinen & Oja, 2000]. ICA often includes a pre-processing step to make the columns of $\mathbf{X}$ uncorrelated and scale their variances to unity. This process is termed "whitening" and is achieved through PCA of $\mathbf{X}$.

3. **non-metric MDS**: Extensions of the classical MDS described above relax the matching criterion of dissimilarities and distances to finding the closest match of a monotonic function of the distances to the dissimilarities: $f(d_{ij}) \approx \widehat{d_{ij}} = \|\mathbf{z}_i - \mathbf{z}_j\|$. The closest match is determined by minimising a stress function [Kruskal, 1964a; Kruskal, 1964b]. In the non-metric version of these extensions, $f(d_{ij})$ simply considers the rank order of the input dissimilarities such that the rank order agreement between the distances and the dissimilarities is maximized [Minchin, 1987]. The parameters to be specified are the threshold for minimum stress at which the distances and dissimilarities are considered to have converged and the number of components to model.

4. **Dimensionality reduction via regression**: DRR is a PCA-based regression technique which aims to remove redundant information present in the PCs $\mathbf{W}$ of $\mathbf{X}$. While standard PCA yields decorrelated dimensions, complete independence of its components is only certain if the high-dimensional data had a Gaussian probability density function [Laparra & al., 2015]. The main idea in DRR is to remove the redundant information contained in partially dependent components and only keep the remaining, non-predictable information in the low-dimensional representation. The removal of the redundant information is achieved in a step-wise manner by starting at the lowest variance component (i.e. smallest eigenvalue) and using it as the response variable for a multivariate non-linear regression function f with all higher variance components as predictors. This process is repeated for each PC until the component with the second highest eigenvalue is reached and all redundant information has been regressed out. Formally, this iterative prediction scheme can be described as $z_i = \mathbf{w}_i - f_i(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{i-1})$, where z_i is the non-predictable information. As

in PCA, the first components account for the highest variance. The number of parameters depends on the function f specified for the non-linear regression. The standard method described in the original paper uses Kernel Ridge regression with a Gaussian kernel function, i.e. one free parameter for the band width of the kernel [Laparra & al., 2015].

5. **Isomap**: Isomap builds on classical MDS for the dimensionality reduction and kernel eigenmaps to find the required dissimilarity matrix of $\mathbf{X}$. The dissimilarities are defined as the geodesic manifold distances between all pairs of data points. Isomap constructs a graph of all data points and sets the edge length between neighbouring points to the geodesic distance. For data points in proximity (based on n nearest neighbours or threshold on the distance measure), the euclidean distance in input space serves a good approximation. The geodesic distance for points outside the proximity criterion is approximated by adding up a sequence of 'short hops' jumps between neighbouring points. The shortest distances between points of the graph are a measure for the dissimilarity between data points and serve as the input data for classical MDS [Tenenbaum & al., 2000]. The proximity threshold is the parameter to specify.

6. **Local linear embedding**: LLE uses kernel eigenmaps based on the local structure in the data to recover the non-linear global data structure. It assumes that any data point in $\mathbf{X}$ lies on a close to linear patch with its neighbours and can be reconstructed through linear recombination of these neighbours. The linear recombination is described in the weight matrix $\mathbf{H}$. The objective of the algorithm is to find $\mathbf{H}$ which minimises the reconstruction error between all data points and their reconstructions. Based on the optimised $\mathbf{H}$, the data points $\mathbf{X}$ can be transformed into lower dimensional space $\mathbf{Z}$ by solving the eigendecomposition of $(\mathbf{I}_N - \mathbf{W})^T(\mathbf{I}_N - \mathbf{W})$ [Roweis & Saul, 2000]. LLE requires the specification of the local neighbourhood size n.

7. **Laplacian Eigenmaps**: Laplacian Eigenmaps are based on an adjacency graph representing $\mathbf{X}$. For adjacent data points (proximity measures as in Isomap), the edges of the graph are weighted based on a heat kernel of the euclidean distance: $\mathbf{H}_{i,j} = exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{n})$. Edges for points that do not fall within the proximity threshold n are set to zero. Based on the weight matrix $\mathbf{H}$, a diagonal matrix $\mathbf{D}$ is constructed by $D = \sum_j \mathbf{H}_{i,j}$ and the positive, semi-definite Laplacian matrix $\mathbf{L}$ computed as: $\mathbf{L} = \mathbf{D} - \mathbf{H}$. The eigendecomposition of $\mathbf{LV} = \boldsymbol{\lambda}\mathbf{DV}$ and selection of the first K eigenvectors $\mathbf{V}$ yields the K-dimensional em-

bedding of $\mathbf{X}$ in $\mathbf{Z}$ [Belkin & Niyogi, 2003]. For dimensionality reduction via Laplacian Eigenmaps, the threshold for the proximity criterion and n, the free parameter in the heat-kernel have to be specified. Large values of n yield less weight to differences in distance, with $n = \infty$ setting all non-zero distances to one.

8. **DiffusionMaps**: As for all kernel eigenmap methods, DiffusionMaps first constructs a graph representation of $\mathbf{X}$ which is turned into the Markow matrix $\mathbf{M}$, used for the low-dimensional embedding. The length of the edges between points on the graph are computed by a kernel $k(\mathbf{x}_i, \mathbf{x}_i)$ normalised to the local connectivity of the graph, and in such capture the local geometry in the data. This normalised kernel can be interpreted as the transition kernel of $\mathbf{M}$, representing the transition probability from point $\mathbf{x}_i$ to $\mathbf{x}_j$ in one time step. Based on the eigenvalues and eigenvectors of $\mathbf{M}$, diffusion distances and maps between the data points can be computed. These are in turn used to map the data into a Euclidean space, where the distance describes the relationship between data points in terms of their connectivity. The dimensionality of the re-mapped data depends on the number of eigenvectors used for the embedding into Euclidean space. These are chosen based on the number of transitions t on $\mathbf{M}$ and an accuracy term ϵ, which specify the maximum eigenvalue considered informative in the mapping [Coifman & al., 2005; Coifman & Lafon, 2006]. Depending on the kernel function, additional parameters might have to be specified. Typical kernel functions are the Gaussian kernel and heat kernels.

9. **t-Distributed stochastic neighbourhood embedding**: In tSNE, the Euclidean distance of the data points in $\mathbf{X}$ are converted into joint probabilities $p_{i,j}$. Similarly, for a low-dimensional representation $\mathbf{Z}$ of $\mathbf{X}$, the distance $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ is converted into the joined probabilities $q_{i,j}$. The objective of tSNE is to find the configuration of $\mathbf{Z}$ which minimises the Kullback-Leibler divergence KL between the probability distributions P and Q: $KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$. KL in general is a measure for how much one probability distribution diverges from another [Kullback & Leibler, 1951] and serves in tSNE as the criterion for finding a good low-dimensional representation. The mapping of similarities to probabilities in the low-dimensional space are based on a Student t-distribution with one degree of freedom, whereas the mappings in high-dimensional space are converted using a gaussian distribution. Depending on the data density around each $\mathbf{x}_i$, the standard deviation is adjusted for each

gaussian p_i based on the specified perplexity, a smooth measure of the effective number of neighbours. In addition, parameters for the gradient descent function used to find the minimum KL have to be specified: the number of iterations, the learning rate and the momentum. For details of these parameters refer to [Maaten & Hinton, 2008].

Despite the diversity of the dimensionality reduction techniques, there are a number of underlying features which define common properties and can give an indication for their applicability. Methods directly based on PCA (PCA, MDS and DRR), are easy to apply and the extracted features are interpretable (directions of variance). While PCA and MDS mainly work well for linear manifolds, DRR extends the applicability to non-linear manifolds. The ability to learn non-linear manifold structures in the data is also shared by the kernel eigenmap methods, nMDS and DRR [Coifman & Lafon, 2006]. However, non-linear models introduce a number of free parameters, whose choice requires prior assumptions about the manifold characteristics. Dimensionality reduction via kernel eigenmaps and tSNE depend on the assumption that distances of points for apart in the global space do not contain information and need not be preserved. Hence, these techniques are simply based on local neighbourhoods and preserve these in the low-dimensional space. This in turn requires dense data points in the low-dimensional space for these strategies to be a good estimation.

There are two main purposes for dimensionality reduction, visualisation and feature selection. For visualisation, K is commonly chosen in a range from one to three such that the data can be presented in a one, two or three dimensional graphic. The choice of dimensionality for feature selection is less trivial, as the dimension of the low-dimensional manifold is unknown. In general, choosing the dimensionality is easiest for PCA and PCA-based methods, where the principal components that cumulatively explain a certain fraction of the variance in the data define the dimensionality. For other methods, the task is less straight forward and different strategies have to be developed. In the next section, I will show the results of applying the techniques described above for the visualisation of two small datasets with known structure.

6.2. Visualisation of data structures by dimensionality reduction

In high-dimensional data analysis, one is often interested in finding a clear visualisation of the data, which leads to a minimal loss of information and is capable of

summarising underlying data structures. Data can be of biological origin, representing features of interest like cell populations or tissue types, or of technical origin such as batch effects. In high-dimensional datasets, visualisation requires either *a prior* selection of the dimensions of the original data to be displayed or the reduction to a dimension that can be represented. Common choices of dimensionality reduction for this task are PCA or tSNE [Deng & al., 2014; Crowley & al., 2015; Corces & al., 2016; Martinez-Jimenez & al., 2017; Huisman & al., 2017].

To understand how the visualisation via dimensionality reduction depends on the underlying dataset, and to see how the true dimensionality of the data is reflected in the visualisation, I needed datasets with known properties. As outlined above, one high-level classification of the dimensionality reduction methods is their grouping into linear and non-linear methods. To understand the relationship between input data and linearity of the dimensionality reduction methods, I selected one dataset with approximately linear structure and created a second dataset with non-linear properties. The datasets are described below and their properties depicted in figure 6.1 and figure 6.3. To allow for an easier comparison of the input data and their visualisation, these figures are located with the figures for the low-dimensional visualisation further down in the document.

The first, linear dataset is a commonly used sample dataset for statistical functions in R (and is distributed with the R software) and consists of 150 samples of three *Iris* species (*I. setosa*, *I. versicolor*, *I. virginica*) for which four phenotypes were measured: sepal width, sepal length, petal width and petal length. In order to get an understanding of the phenotype structure, I computed the pair-wise Pearson correlation coefficient across the three species and across the four phenotypes (one sample appears twice in the dataset and was removed for subsequent analyses). The strongest correlation on species level is observed for *I. virginica* and *I. versicolor* ($r^2 = 0.9$). On phenotype level, petal length and width correlate strongly across species ($r^2 = 0.96$, figure 6.1).

For the second, non-linear dataset, I simulated 2,000 data points uniformly distributed on a (x,y)-plane and transformed the plane into (x,y,z) coordinates by $z = x\sin(x)$ and $x = x\cos(x)$. The resulting "roll" structure is depicted in figure 6.3.

These datasets represent two distinct types of data: the *Iris* data is a four-dimensional dataset comprised of three subgroups, whereas the roll data is two-dimensional manifold non-linearly embedded in a three-dimensional space. In the following, I applied the twelve dimensionality reduction techniques described above to both datasets and compared their low-dimensional representations.

For each technique, I used corresponding functions already implemented in publicly available R-packages. Table 6.2 summarises the R packages, functions and their parameters used for the dimensionality reduction. Most functions require specification of the expected number of dimensions $ndim$. For the purpose of visualisation in a Cartesian coordinate system, this parameter choice is straightforward (one, two or three) and was set to $ndim = 2$. In the case of kernel eigenmap methods and tSNE, the number of n nearest neighbours used in the graph construction and probability function have to be provided. This task is less intuitive and different algorithms have been implemented to estimate the optimal number of neighbours for the reconstruction. Choosing a suitable n is important, as neighbourhoods chosen too large might eliminate fine structures in the data, while neighbourhoods too small can lead to the division of the continuous input space into smaller, unconnected sub-manifolds [Kayo, 2006].

For any method that required specification of n, I provided n estimated according to the method proposed by Kayo [2006], implemented as the function *calc_k* in the *lle* package. Some methods require additional, specific parameters. These are either specified in table 6.2 or the default setting was chosen. For functions that required a distance matrix or metric for the local neighbourhood estimation (MDS, DiffusionMap, Isomap, nMDS), the default is the Euclidean distance. Methods that require a kernel function (DRR, kPCA) use a gaussian radial basis kernel by default. For ICA and DRR, I choose the default setting of the PCA pre-processing step. For PEER, the functions are implemented in an object-oriented manner and I followed the protocol described in Stegle & al. [2012]. I choose to include the optional parameter of adjusting for the mean.

Before applying dimensionality reduction functions to both datasets, I estimated the optimal number of neighbours for the dimensionality reduction techniques based on local neighbourhoods. For the *Iris* data with 596 data points, the optimal number of neighbours is estimated to be $n = 26$. For the roll data with 2,000 data points it was estimated to be $n = 36$. Figure 6.2 shows the two-dimensional representation of the *Iris* data after dimensionality reduction by the four linear and eight non-linear dimensionality reduction techniques. Assuming that the allocation to species is the correct intrinsic low-dimensional representation of the *Iris* dataset, I coloured the data points according to species to enable the visual comparison of the goodness of the dimensionality reduction.

PCA, i.e. the representation of the data based on the direction of highest variation in the data is able to clearly separate the *I. setosa* from *I. versicolor* and *I. virginica* across

Table 6.2: R functions for dimensionality reduction methods and their parameters. Most functions require *a priori* specification of the number of n nearest neighbours and the expected intrinsic dimensionality $ndim$. Any function-specific parameters different to the default settings are listed. The reference column specifies the publications the R packages are based on. LE: Laplacian Eigenmaps, DM: Diffusion Maps.

Name	R function	Parameters	Reference
PCA	stats::prcomp	-	[Hotelling, 1933]
PEER	peer	ndim,	[Stegle & al., 2010]
ICA	fastICA::fastICA	ndim, fun=logcosh, method="C"	[Hyvärinen & Oja, 2000]
MDS	stats::cmdscale	ndim	[Gower, 1966]
nMDS	vegan::metaMDS	ndim	[Ripley, 1996]
DRR	DRR::drr	-	[Laparra & al., 2015]
kPCA	kernlab::kpca	-	[Schoelkopf & al., 1998]
Isomap	vegan::isomap	ndim, k, fragmentedOK=TRUE	[Tenenbaum & al., 2000]
LLE	lle::lle	ndim, k	[de Ridder & Duin, 2002]
LE	loe::LOE	ndim, k	[Belkin & Niyogi, 2003]
DM	diffusionMap::diffuse	k	[Lafon & Lee, 2006]
tSNE	Rtsne::Rtsne	ndim, k	[Maaten & Hinton, 2008]

the first principal component. However, the separation of the strongly correlated *I. versicolor* and *I. virginica* species based on the first two principal components alone is not possible. MDS with Euclidean distance is equivalent to PCA and the resulting MDS plot is a mirror image of the PCA result on the x-axis. ICA for this dataset shows the strong influence of the pre-processing via PCA, as it is the mirror image of the PCA result on the x- and y-axis. PEER is capable of separating *I. setosa* from the other species, but similarily fails at completely separating *I. versicolor* and *I. virginica*. Visually the best results of the non-linear methods are obtained from DRR, Isomap and nMDS and perform similarly in their ability to separate the species as the linear methods. The other non-linear methods are able to separate *I. setosa*, but do worse in separating the other two species.

The results of the dimensionality reduction for the non-linear projection of the 2D manifold into 3D space demonstrate the difficulty of the linear methods to deal with non-linear structures (figure 6.4). The color scheme of the original embedding simply represents the location of points in the 2D plane ordered in x-direction. In a good low-dimensional representation, one should be able to observe the gradient of the original (x,y)-plane linearly across either one of the dimensions. While the general order of the points is conserved in the low-dimensional representation for the linear methods, none are able to separate them linearly in either dimension (fig-

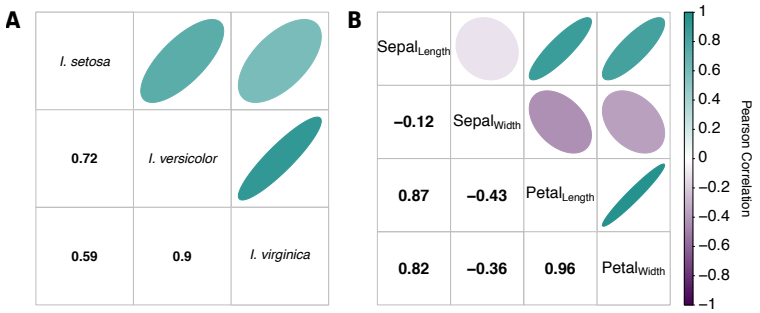

Figure 6.1: Correlation of flowering phenotypes. For the 149 unique samples in the *Iris* dataset, the pair-wise Pearson correlation for the three different *Iris* species across all measurements (A) and the four flowering phenotypes sepal width, sepal length, petal width and petal length across the three species (B) are depicted. The color scheme and shapes in the upper triangle of the matrix represent the strength and direction of the correlation, the lower triangle depicts the value of the correlation. Generated via R function *corrplot::corrplot*.

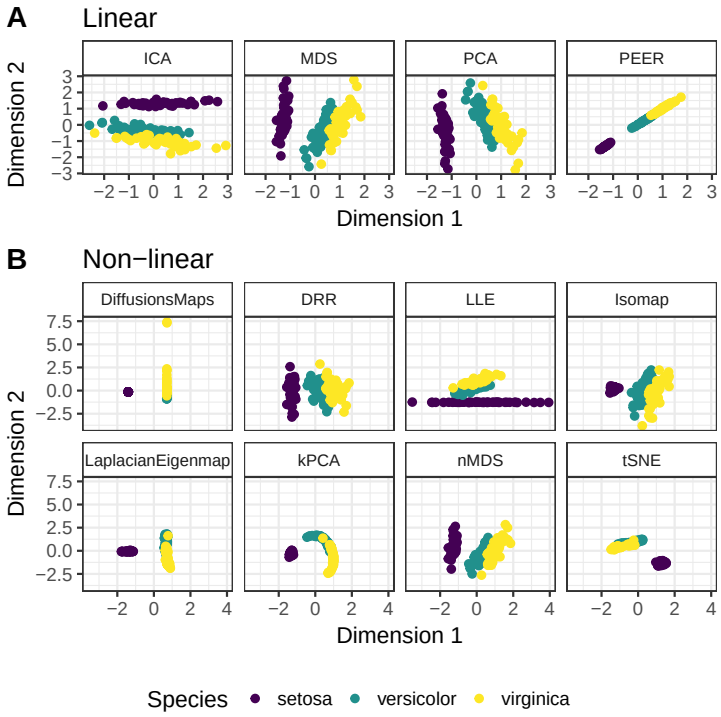

Figure 6.2: Visualisation of the *Iris* dataset in two dimensions. The number of dimensions in the *Iris* dataset was reduced form four to two by the dimensionality reduction techniques described in table 6.1 and computed with the functions and parameters listed in table 6.2. The number of nearest neighbours provided to the local-proximity-based methods was estimated to be $n = 26$.

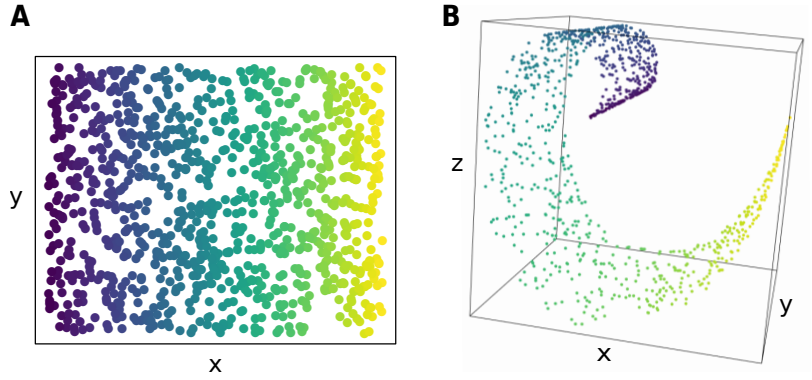

Figure 6.3: Three-dimensional embedding of data points lying on a two-dimensional plane. Data points uniformly distributed on a (x,y)-plane (A) are transformed into (x,y,z) coordinates by $z = x\sin(x)$ and $x = x\cos(x)$. The color scheme simply represents the location in x-direction of the (x,y)-plane. Generated via R function *plot3D::scatter3D*

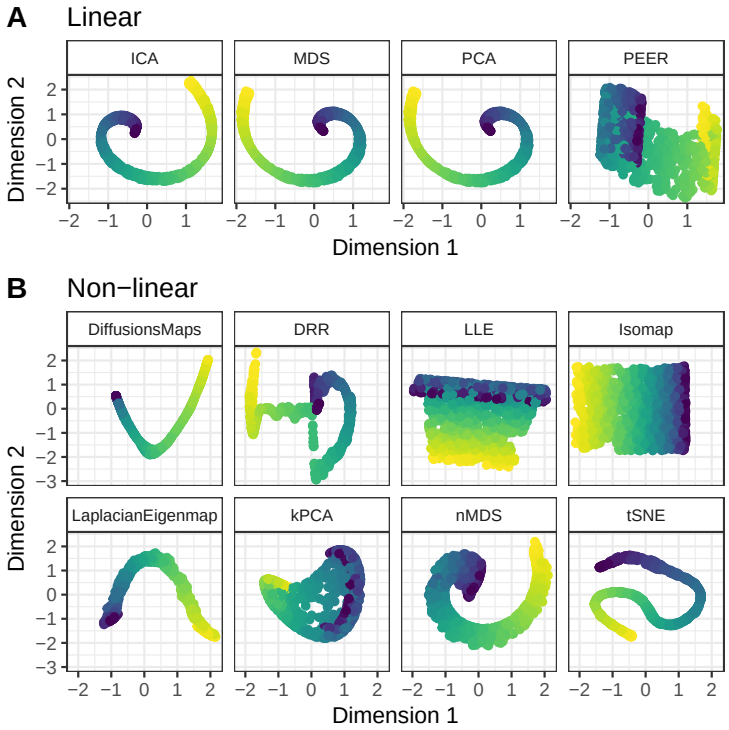

Figure 6.4: Visualisation of the roll dataset in two dimensions. The dimensionality reduction methods described in table 6.1 were analysed for their ability to recover the original 2D plane embedded into 3D space (figure 6.3). The 2D-representation was computed with the functions and parameters listed in table 6.2, with the number nearest neighbours provided to the local-proximity-based methods estimated to be $n = 36$.

ure 6.4A). PEER performs best in capturing the spread in y-direction compared to the other linear methods, but equally fails in separating the tight curvature x-direction. In contrast, the non-linear method Isomap completely recovers the original 2D plane. DiffusionMap and Laplacian Eigenmaps are able to separate the structure linearly, but underestimate the spread of the original data in y-direction. LLE recovers the spread in y-direction, but fails to find the order in x-direction for the tight curvature (dark colors) in the 3D space. DRR, nMDS and tSNE suffer from the same issues as the linear methods, with DRR additionally introducing non-smoothness. kPCA recovers the plane structure for the mid-section of the roll, but scrambles the order at both ends.

The visualisations clearly demonstrate the difference in ability of the dimensionality reduction methods to find a good low-dimensional representation of the original, known data structures. As a generalisation and unsurprisingly, linear methods perform well in separating linear data structures (*Iris* data) but fail in in recovering non-linear structures (roll data). Non-linear methods perform better in recovering the non-linear structure, but underperform on linear datasets compared to the linear methods.

6.3. Quantification of dimensionality reduction performance

In addition to the visualisation, it would be desirable to have a quantitative assessment of the performance of the dimensionality reduction techniques. Lee & Verleysen [2009] reviewed different methods for evaluating the quality of dimensionality reduction methods. Two criteria for the goodness of the low-dimensional representation contained in three out of the five methods reviewed are the closeness of neighbouring samples in the low-dimensional space compared to the original space (trustworthiness of the projection) and the conservation of original neighbourhoods in the low-dimensional space (continuity of the projection). Kaski and colleagues [Kaski & al., 2003] proposed two metrics quantifying the extend of trustworthiness and continuity based on the ranking of k neighbours in the original and low-dimensional space. For trustworthiness, they define $r(x_i, x_j)$ as the rank of the distance of x_j to x_i in the original data space and $U_k(x_i)$ as the set of $x_{j \neq i}$ that are in the neighbourhood of x_i in the low-dimensional space but not in the original space. Similarly, continuity is based on $\hat{r}(x_i, x_j)$, the rank of the distance of x_j to x_i in the low-dimensional space and $V_k(x_i)$ as the set of $x_{j \neq i}$ that are in the neighbourhood of x_i in the original space but not in the low-dimensional space. The trustworthiness T and the

continuity C are defined as:

$$T = 1 - A(k) \sum_{i=1}^{N} \sum_{x_j \in U_k(x_i)} (r(x_i, x_j) - k) \tag{6.2}$$

and

$$C = 1 - A(k) \sum_{i=1}^{N} \sum_{x_j \in V_k(x_i)} (\hat{r}(x_i, x_j) - k), \tag{6.3}$$

where $A(k) = \frac{2}{Nk(2N-3k-1)}$ is introduced as a normalising parameter scaling the values between zero and one. The projection into low-dimensional space is considered trustworthy if the set of k closest neighbours of a sample in the low-dimensional space are also close in the original space. Continuity quantifies how well the original neighbourhoods are preserved, i.e. it measures if there are neighbourhoods of k points in the original space which are not preserved because of discontinuities in the low-dimensional space.

I applied these metrics to the results of the low-dimensional projections obtained in section 6.2. Both metrics are dependent on the number of k neighbours that they are evaluated on, so I chose different neighbourhood sizes ranging from 1 to 3% of samples (rows) in the dataset. The results are depicted in figure 6.5 and figure 6.6. tSNE, LLE, the PCA-derived linear methods (PCA, ICA, MDS) and nMDS have a trustworthiness measure of more than 0.95 across all neighbourhood sizes in the *Iris* data (figure 6.5A). The PCA-derived non-linear method DRR performs slightly worse, as do kPCA and Isomap. Laplacian Eigenmaps perform worst only reaching 0.9 for high neighbourhood sizes. In general, the dependency of the local methods on neighbourhood size becomes apparent, as the kernel-eigenmap methods' trustworthiness varies strongest across the different neighbourhood sizes. The six methods performing well in terms of trustworthiness for the *Iris* data (tSNE, LLE, PCA, ICA, MDS and nMDS) also keep the level of discontinuities introduced in the low-dimensional space low as seen by high measures of continuity (figure 6.5B). To get an estimate for T and C for a poor representation of the original data, I randomly chose neighbourhoods in the original space and computed trustworthiness and continuity measures for these and the original *Iris* data, leading to median measurements of 0.51 for both T and C (results not shown in graphic to allow for a clearer visualisation of the trustworthiness range 0.85 to 1). For the roll data, Isomap has by far the best performance in terms of trustworthiness (figure 6.6A) and confirms the visual results (figure 6.4B). LLE and nMDS also score above 0.9. These three methods together with kPCA and DiffusionMaps are best in preserving continuities (figure 6.6B). The

trustworthiness for all linear methods is similar and consistently lower than the best scoring non-linear methods. The worst results in terms of continuity are observed for tSNE and DRR and both methods show discontinuities in the visualisation (figure 6.4B). For the reference point of trustworthiness and continuity based on random neighbourhoods, results similar to those found for the *Iris* dataset were observed, with median $T = 0.52$ and $C = 0.52$.

Overall, the trustworthiness and continuity measures reflect the results obtained from the visualisation of the data by their low-dimensional representation: linear methods are most suitable for linear data and non-linear methods for non-linear data.

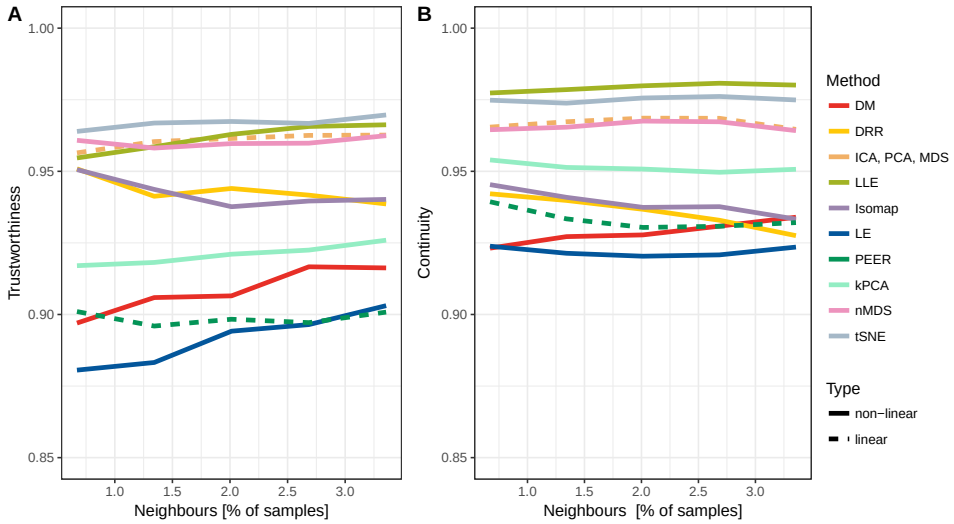

Figure 6.5: Quality of the dimensionality reduction in the *Iris* dataset. The trustworthiness (A) and Continuity (B) of the projections into the low-dimensional space for the *Iris* dataset were computed according to equation (6.2) and equation (6.3). The neighbourhood sizes ranged from one to five neighbours, corresponding to 0.6 to 3.4% of samples.

6.4. Dimensionality reduction for feature extraction

Apart from serving as a tool for visualisation, dimensionality reduction is often used for feature extraction. While visualisation is limited to one, two or three dimensions, for feature extraction one is interested in the intrinsic dimensionality of the data which can be of much higher dimension. Metrics, such as the one introduced in the previous section, can indicate which methods provide a trustworthy dimensionality reduction. However, they do not help with choosing the number of dimensions in

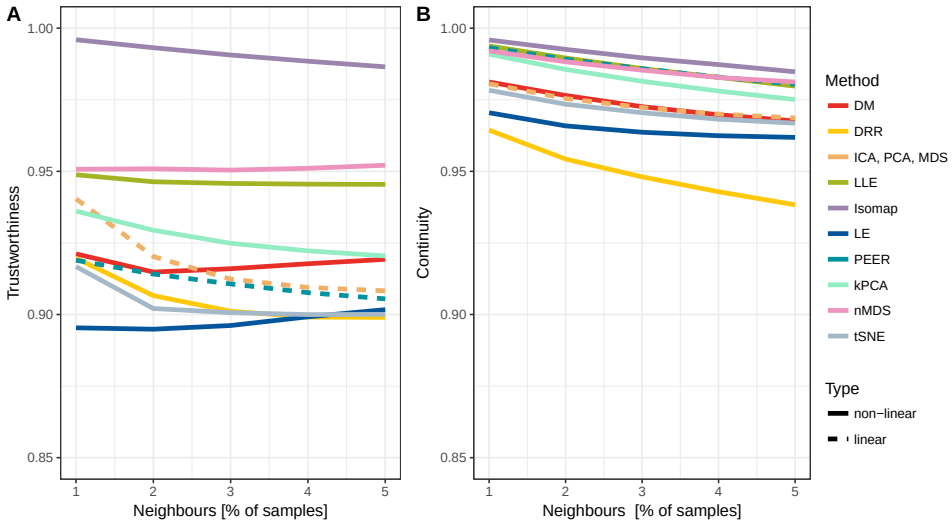

Figure 6.6: Quality of the dimensionality reduction on the 2D manifold embedded in 3D. The trustworthiness (A) and Continuity (B) of the projections into the low-dimensional space for the 2D manifold were computed according to equation (6.2) and equation (6.3). The neighbourhood sizes ranged from 10 to 50 neighbours, corresponding to 1 to 5% of samples.

the low-dimensional space. Here I propose a novel, simple stability criterion for the choice of dimension (section 6.4.1) and show that features selected based on the stability criterion are able to capture underlying genetic structure (section 6.4.2).

6.4.1. Stability of dimensionality reduction

An assumption in dimensionality reduction for feature selection is that these techniques capture the variation or structure of the high-dimensional space in the low-dimensional components. In an ideal scenario, any technical or unwanted covariates have been accounted for *a priori* (e.g. through regression) and the low-dimensional components will only capture the true biological structure in the data. While the dimensionality reduction techniques intrinsically learn structures based on the observed data, they should be robust against small changes in the data such as removing or adding a moderate number of samples. In the following, I will describe a method of finding robust low-dimensional representations and will call these representations stable. As such, stability is a simple but effective way of ensuring reproducibility, but cannot be used to distinguish an appropriate from a less appropriate low-dimensional representation. In contrast, a dimensionality reduction that is *not* stable is certainly not capable of producing reliable results.

In order to estimate the stability of the dimensionality reduction techniques and

to investigate different parameters potentially influencing the stability, I used *PhenotypeSimulator* (chapter 3) to simulate datasets of 1,000 phenotypes with different numbers of samples and phenotype components as described in section 3.2. The sample sizes ranged from 500 samples as observed in small cohort studies with dimensionality-reduced phenotypes [Pausova & al., 2007] to 10,000 [Liu & al., 2012]. All phenotypes were simulated with genetic variant and infinitesimal effects and noise effects. A total of 50,000 SNPs was simulated with allele frequencies of 0.1, 0.2 and 0.4 chosen at equal probability. 20 SNPs were selected for the simulation of genetic variant effects with effect sizes drawn from $\mathcal{N}(0, 1)$. The genetic kinship matrix was estimated based on all simulated SNPs. For each sample size, an additional phenotype set was simulated that also contained non-genetic covariate and correlated noise effects. The parameters for the simulation are summarised in table 6.3. For each simulation set-up, ten independent datasets were simulated and subsequent analyses applied to each dataset individually.

Table 6.3: Simulation parameters of phenotypes used for stability estimation. N: number of samples, P: number of traits; h_2: total genetic variance, h_2^s: variance of genetic variant effects, h_2^g: variance of genetic random effects, $1-h_2$: total noise variance, δ: variance of non-genetic covariate effects, rho: variance of correlated noise effects; pcorr: correlation of correlated noise effects, θ: proportion of shared genetic variant effects, η: proportion of shared genetic random effects,γ: proportion of shared non-genetic covariate effects, α: proportion of shared noise random effects.

Parameter	Parameter values
N	500, 1,000, 10,000
P	1,000
h_2	0.4
h_2^s	0.01
h_2^g	0.99
(1-h_2)	0.6
(1-h_2)δ	0.4, 0.4
(1-h_2)(1-δ)ρ	0.2, 0
(1-h_2)(1-δ)(1-ρ)	0.4, 0.6
pcorr	0.4
θ	0.8
η	0.8
γ	0.8
α	0.8

To test the stability of dimensionality reduction techniques, I chose a cross-vali-

dation approach, where I randomly selected 80% of the simulated samples, applied a dimensional reduction technique and recorded the results. For each dataset, I repeated this step ten times. Subsequently, I did a pairwise comparison of the ten low-dimensional representations of the dataset, hence 45 comparisons. For each pairwise comparison, I selected the samples common to both datasets and computed the Spearman correlation of the components across these samples. I matched each of the components in the first dataset to the component in the second dataset with which it had maximum correlation. The matching algorithm started at the highest correlation and allowed for each component to be exactly matched once. In case of a tie, it was matched to the closest component in rank that had not been matched yet. After finding the pairs of highest correlation, I counted the number of components that passed a given threshold. Components that showed more than 90% correlation were considered stable.

I applied the twelve dimensionality reduction methods described in section 6.1 with the parameters summarised in table 6.2 to the different simulated datasets and determined the trustworthiness, continuity and stability of each method. Instead of directly using the raw simulated data as input for the dimensionality reduction, I followed standard methods used the residuals from a linear regression of the simulated data with the known confounders (introduced as non-genetic covariate effects in the simulation). For methods that required the specification of the dimensionality, I provided an initial estimate of $ndim = 100$. These 100 dimensions will be the 100 components explaining most variance in the data for methods based on or including a pre-processing step that uses variance selection (PCA, DRR, tSNE, ICA, MDS, nMDS and Laplacian Eigenmaps). For PEER, which uses iterative model updates, selecting a dimensionality that is too high, will be compensated for by the weights associated with the components, which will effectively set the contribution of the non-informative components to zero. In this way, an initial poor choice of too many dimensions will affect the final estimated components only minimally. In LLE, the provided dimension is only used as a maximum value and the estimation of any component is not affected by the estimation of subsequent components [Roweis & Saul, 2000; Kayo, 2006].

Figure 6.7 summarises the effects of sample size and background structure on the different dimensionality reduction methods. The effect is measured as the trustworthiness and continuity of the projection across the ten subsets of each dataset. For most methods, the sample size has only minor effects on the trustworthiness of the low-dimensional projection. Laplacian Eigenmaps are the exception to this ob-

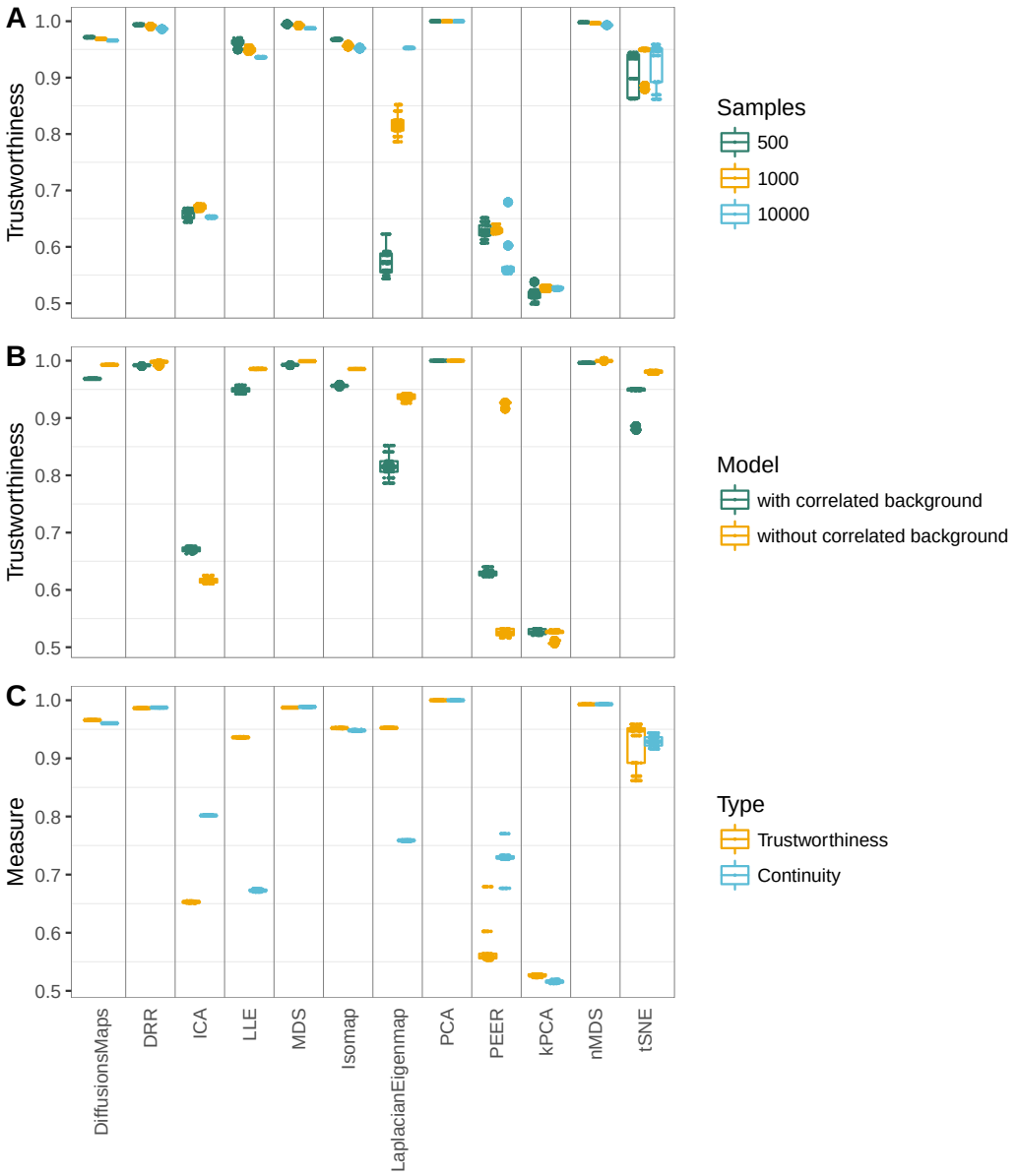

Figure 6.7: Performance of dimensionality reduction techniques on simulated datasets. The trustworthiness and continuity (equation (6.2) and equation (6.3)) of twelve dimensionality reduction methods on ten independent simulated datasets for each phenotype setup were computed. 1,000 phenotypes with non-genetic covariates and observational noise effects or non-genetic covariates, observational noise effects and correlated noise effects were simulated for datasets of 500, 1,000 and 10,000 samples. For each dataset, a ten-fold cross-validation of the dimensionality reduction and subsequent computation of trustworthiness and continuity was conducted. The results of ten evaluations on the ten independent datasets are summarised in the boxplots. A. Trustworthiness of the dimensionality reduction depending on the number of samples in the simulated dataset (noise background model: no correlated background). B. Trustworthiness depending on the background noise structure of the phenotypes (sample size: 10,000). C. Performance of the dimensionality reduction techniques in terms of trustworthiness and continuity (sample size:1,000, noise background model: correlated background).

servation, as the trustworthiness of the dimensionality-reduced datasets sharply increases with sample size (figure 6.7A). The effect of the background structure of the phenotype is shown in figure 6.7B. Most models perform marginally better on data without correlated background structure, while the trustworthiness of the representation found by Isomap and PEER is distinctly better on this data type. In contrast, ICA performs slightly better on datasets with correlated background structure. Two thirds of the models that yield trustworthy projections, also perform well in terms of continuity (figure 6.7C). PEER and ICA seem to be better at protecting original neighbourhoods (continuity), than they are at ensuring that the samples in low-dimensional space were in proximity in the original space (trustworthiness). The opposite trend can be observed for LLE and Laplacian Eigenmaps. kPCA performs worst overall and is only marginally better than randomly simulated neighbourhoods as a low-dimensional representation (section 6.3).

The stability of the dimensionality reduction techniques dependent on the background model is displayed in figure 6.8. For the majority of methods (DRR, MDS, Isomap, PCA, PEER and nMDS), the background structure of the dataset does not influence the stability of the components, with three components reliably recovered in the ten-fold cross-validation. DiffusionMaps and LLE detect more stable components for both data types with five and seven stable components in the data with correlated background and seven and five without correlated background, respectively. kPCA performs worse for both data types, while ICA has no components that pass the 0.9 correlation threshold for either of the data types. tSNE only finds stable components in the model with correlated background structure. Results for the datasets with 500 and 10,000 samples were consistent with these observations.

6.4.2. Stable features enable discovery of genetic associations

In genetic association studies of high-dimensional phenotypes, features selected by dimensionality reduction methods serve as the response variable and one aims to find genetic components that are associated with this low-dimensional phenotype representation. Studies employing these techniques range from genotype association studies on features extracted from facial images [Liu & al., 2012] and metabolic profiles [Avery & al., 2011] to genome-wide pathway association studies of multiple correlated phenotypes [Zhang & al., 2012]. These studies commonly test the association between SNPs and the top few components that explain most phenotypic variance. For instance, the first eleven PCs capturing more than 90% of variance of facial features were used as the phenotypes in the study by Liu and colleagues.

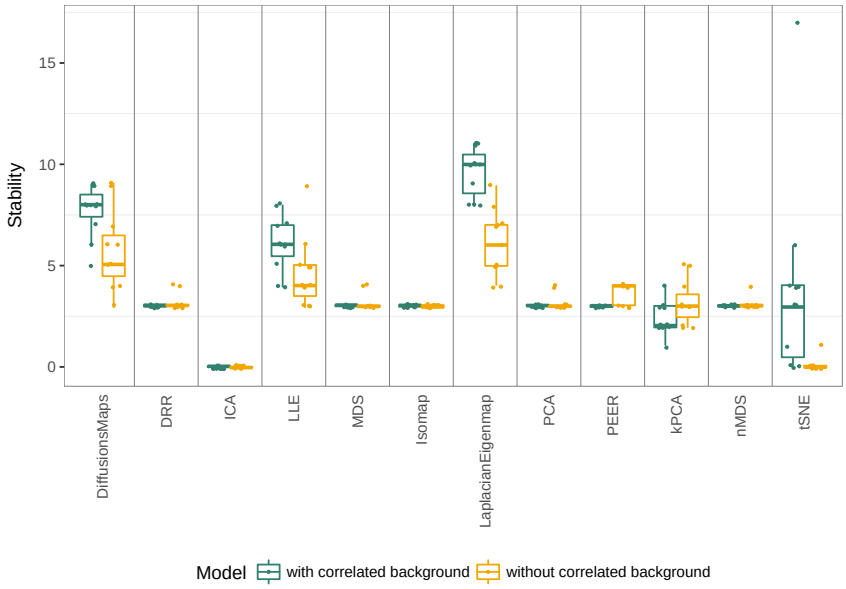

Figure 6.8: Stability of dimensionality reduction techniques for different background noise models. The stability of twelve dimensionality reduction methods on the ten independently simulated datasets per setup were computed. 1,000 phenotypes with non-genetic covariates and observational noise effects or non-genetic covariates, observational noise effects and correlated noise effects were simulated for the datasets with 1,000 samples. For each dataset, a ten-fold cross-validation of the dimensionality reduction and subsequent evaluation of the stability was conducted. Components that passed the correlation threshold of $cor = 0.9$ were considered stable and the number of stable components per method is displayed. For ICA, no stable components were detected for either dataset, for tSNE the same was true for the dataset without correlated background structure.

Similarly, Avery and colleagues used the first eight PCs extracted from the metabolic profiles based on 19 traits for the genotype to phenotype mapping analysis. Contrary to this common practice, Aschard & al. [2014] showed in simulations and in application to a datasets of coagulation traits that only testing the top PCs can lead to a loss in power for detecting genetic associations. They demonstrated that combining signal across PCs can increase power and that components explaining little phenotypic variance can be equally important as components explaining large variation. However, as seen in the previous section, phenotype components that reflect lower variance structures might reflect technical or biological noise and may not be recovered when subsampling the dataset. As such, the choice of dimensionality when using the extracted features for genotype to phenotype mapping comes down to a trade-off between gain in power and stability.

In order to test if the dimensionality reduction techniques employed so far can stably capture phenotypic components that yield enough power to serve as proxy phenotypes in association studies, I simulated a new set of phenotypes with genetic variant effects that affect different proportions of traits. I used the same strategy and parameter settings for the simulation of the noise effects as described for the phenotype simulation of the stability analysis, i.e. datasets with non-genetic covariates and observational noise effects or non-genetic covariates, observational noise effects and correlated noise effects (table 6.3). For each of these datasets, I simulated different structures of genetic variant effects, by adding 20 SNP effects to a subset of traits. The percentage of affected traits ranged from 1 to 100, corresponding to ten and all 1,000 simulated traits. Independent of the subset size, the proportion of variance of the genetic variant effects in relation to the total phenotypic variance was set to 0.05, corresponding to $h_2^s = 0.02$ for $h_2 = 0.4$. The basis for the simulation of the genetic effects were the genotypes and kinship estimate of the simulated cohort with related individuals described in section 3.1. For each setup, i.e. each background noise model (with/without correlated background structure) and percentage of traits affected, I generated ten datasets and applied the twelve dimensionality reduction methods to each dataset. To determine the stability of the dimensionality reduction and decide which components to use for the genetic association study, I employed the cross-validation approach described in section 6.4.1.

For the majority of dimensionality reduction methods, the percentage of traits affected does not affect their stability in the dataset with correlated background structure (figure 6.9A). LLE and LaplacianEigenmaps do not follow this general observation and show some fluctuations in the stability, without showing an obvious trend.

ICA and tSNE on average do not find stable components for any number of traits affected. In the model without background structure (figure 6.9B), there is a general trend towards more stable components in the dataset when a larger subset of traits was affected by the genetics. DiffusionMaps and Laplacian Eigenmaps show the opposite behaviour, while there is no clear trend for tSNE. ICA can again not stably recover any components. For all methods, the median number of stable components across different proportions of traits influenced by genetics in the model without background structure is approximately the same as the number of components in the model with background structure.

For every setup, stable components were selected and used as the response variables in a multivariate LMM with an any effect trait design matrix (section 1.7.8) for the 20 causal SNPs and the kinship matrix as the random genetic effect. The significance of the association was assessed by the permutation approach described in section 4.7, where the original p-values are compared to p-values from the same association model on permuted genotypes to obtain an empirical p-value. Figure 6.10 shows the percentage of causal SNPs that could be detected with this approach ($p_{\text{empirical}} < 0.01$). ICA is not depicted as it was not possible to find stable components for any of the phenotype sets. In general, the percentage of detected true SNPs is lower for components derived from phenotypes with correlated background structure (figure 6.10A) as compared to those from phenotypes without correlated structure (figure 6.10B). Similar to the observation for the stable number of components (figure 6.9), the percentage of detected SNPs does not vary much depending on the number of traits affected in the datasets with correlated background structure. For the phenotypes without correlated background structure, there is a trend towards detecting more SNPs for larger subsets of traits affected by the genetics. For both phenotype models, the PCA-based methods (DRR, MDS, PCA and PEER) and nMDS perform better in recovering the underlying genetics. These methods all perform best in finding components that allow for detecting causal SNPs in phenotypes where 40% of all traits where affected by the genetics, with up to 80% of SNPs detected on average.

The power to detect SNPs in the standard genotype-phenotype mapping approach depends, among other factors such as sample size and allele frequency, on the effect sizes of the SNP [Cohen, 1992; Halsey & al., 2015; Astle & al., 2016]. For phenotypes derived via dimensionality reduction, the effect size of the SNP has an additional influence on the outcome of the association. While the effect size of the SNP is linked to power as in any genotype-phenotype mapping, its influence is likely to also occur

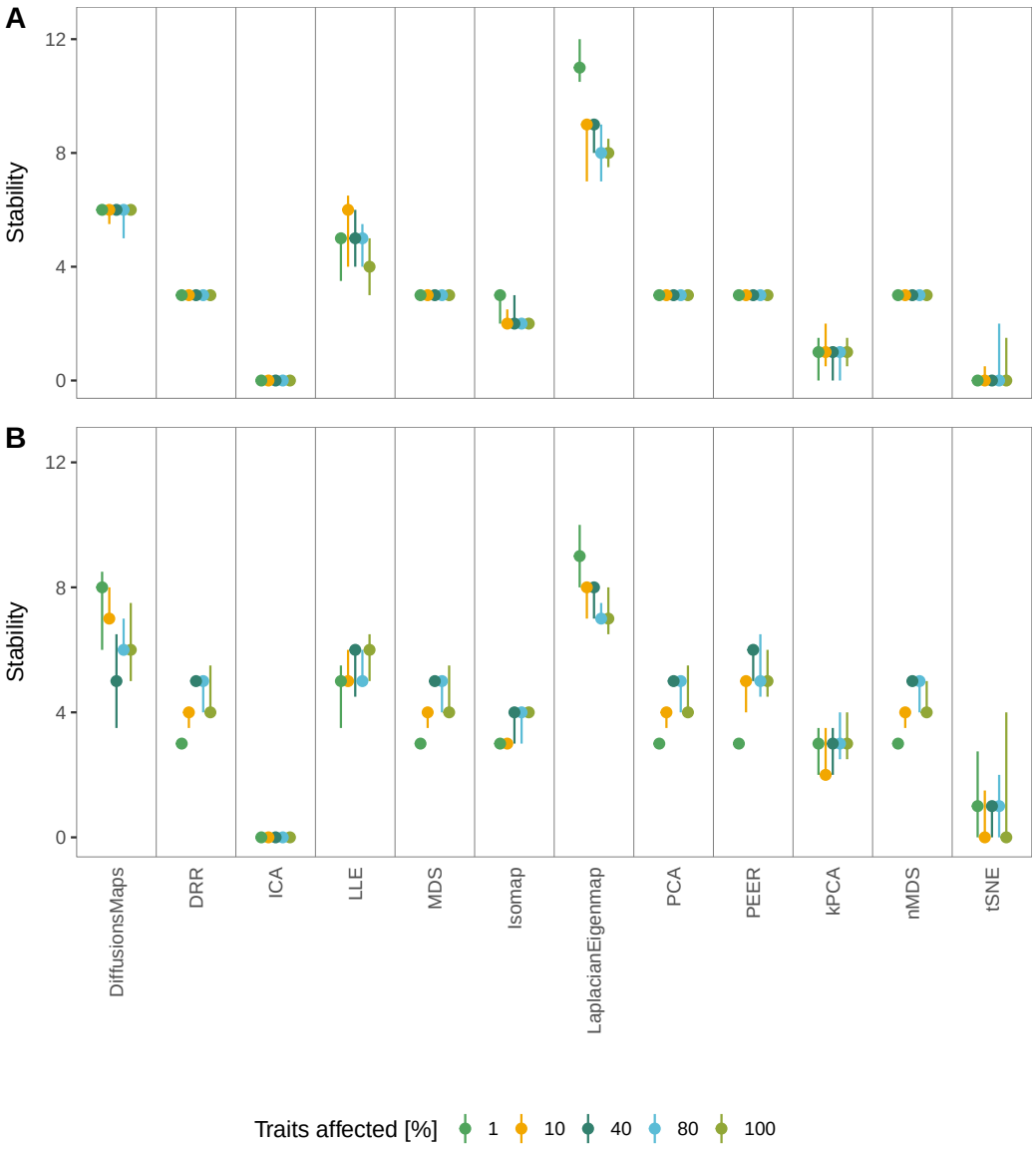

Figure 6.9: Stability of dimensionality reduction techniques for different genetic variant and observational noise models. A. Components from datasets with correlated background structure. B. Components from datasets without correlated background structure. The stability of twelve dimensionality reduction methods on ten independent simulations of ten datasets (two different noise background models, five subset sizes of traits affected by the genetic variant effect, 1,000 phenotypes) were computed. For each dataset, a ten-fold cross-validation of the dimensionality reduction with 80% of the 1,000 samples and subsequent evaluation of the stability was conducted. Components with $cor \geq 0.9$ were considered stable and the median number of stable components per method and dataset is displayed (points). The vertical lines indicate the 25% and 75% quantile for the ten independent simulations.

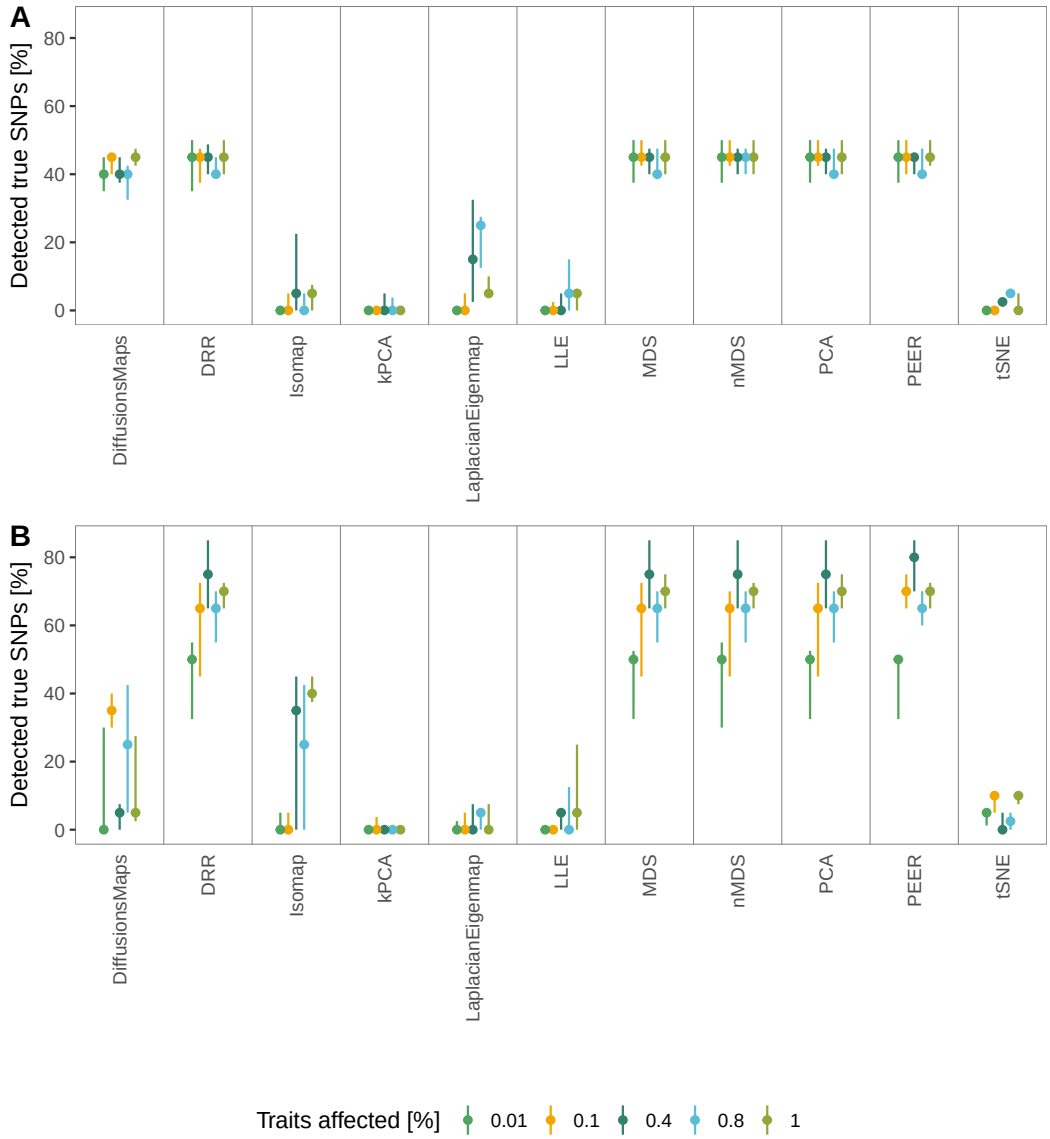

Figure 6.10: Genetic association of stable components from dimensionality reduction. A. Detected SNPs from datasets with correlated background structure. B. Detected SNPs from datasets without correlated background structure. The stable components for each dataset were used as the response variables in a multivariate LMM with an any effect trait design matrix for the 20 causal SNPs and the kinship matrix as the random genetic effect. Vertical lines indicate the 25% and 75% quantile, points represent the median for the ten independent simulations.

before the mapping, namely in finding stable components that reflect this genetic structure.

To test if finding low-dimensional components that capture the underlying genetics depends on the effect size of the causal SNPs, I computed the mean absolute value of effect sizes from the causal SNPs for all simulated datasets. I then classified these SNPs into two categories, based on passing the FDR threshold of $p_{\text{empirical}} < 0.01$. SNPs with empirical p-values below that threshold are considered "detected", the remainder are "not-detected". Figure 6.11 depicts the effect sizes of these SNP categories dependent on the dimensionality reduction technique that was used for deriving the phenotypes, summarised across all proportions of traits affected by the genetic variant effects. ICA and kPCA are not depicted as they either did not detect stable components or their stable components did not detect associations. On average, the effect size of the detected SNPs are larger than the ones for SNPs that are not detected. The results for the linear methods (MDS, PEER, PCA) and nMDS are mostly identical, with median effect sizes for detected SNPs slightly higher in the model with correlated background (figure 6.11A) than without (figure 6.11B). DRR follows the same trend as does LLE, albeit on marginally higher effect size levels. DiffusionMaps, Laplacian Eigenmaps, Isomap and tSNE require higher effect sizes to detect SNPs in the model without correlated background structure. The spread and number of outliers of effect sizes for undetected causal SNPs is smallest for DRR, MDS, PEER, PCA and nMDS for both noise background models. For the background model with correlated structure, the spread of effect sizes for DiffusionMaps is equally low with a median of 0.2. For the other methods, large numbers of outliers for SNPs with high effect sizes that could not be detected are observed, i.e. SNPs with high effect sizes that were not detected.

6.5. Dimensionality reduction is a powerful tool for genetic association studies

In this chapter, I reviewed dimensionality reduction methods with different properties and underlying mathematical concepts. I analysed their performance in terms of trustworthiness and continuity and introduced a new measure, stability, to asses the low dimensional phenotype representations they generate. Finally, I investigated if using low-dimensional representations of the original phenotypes are capable of recovering the underlying genetic structure in simulations.

I was able show on datasets with known structure (*Iris* and roll dataset) that the

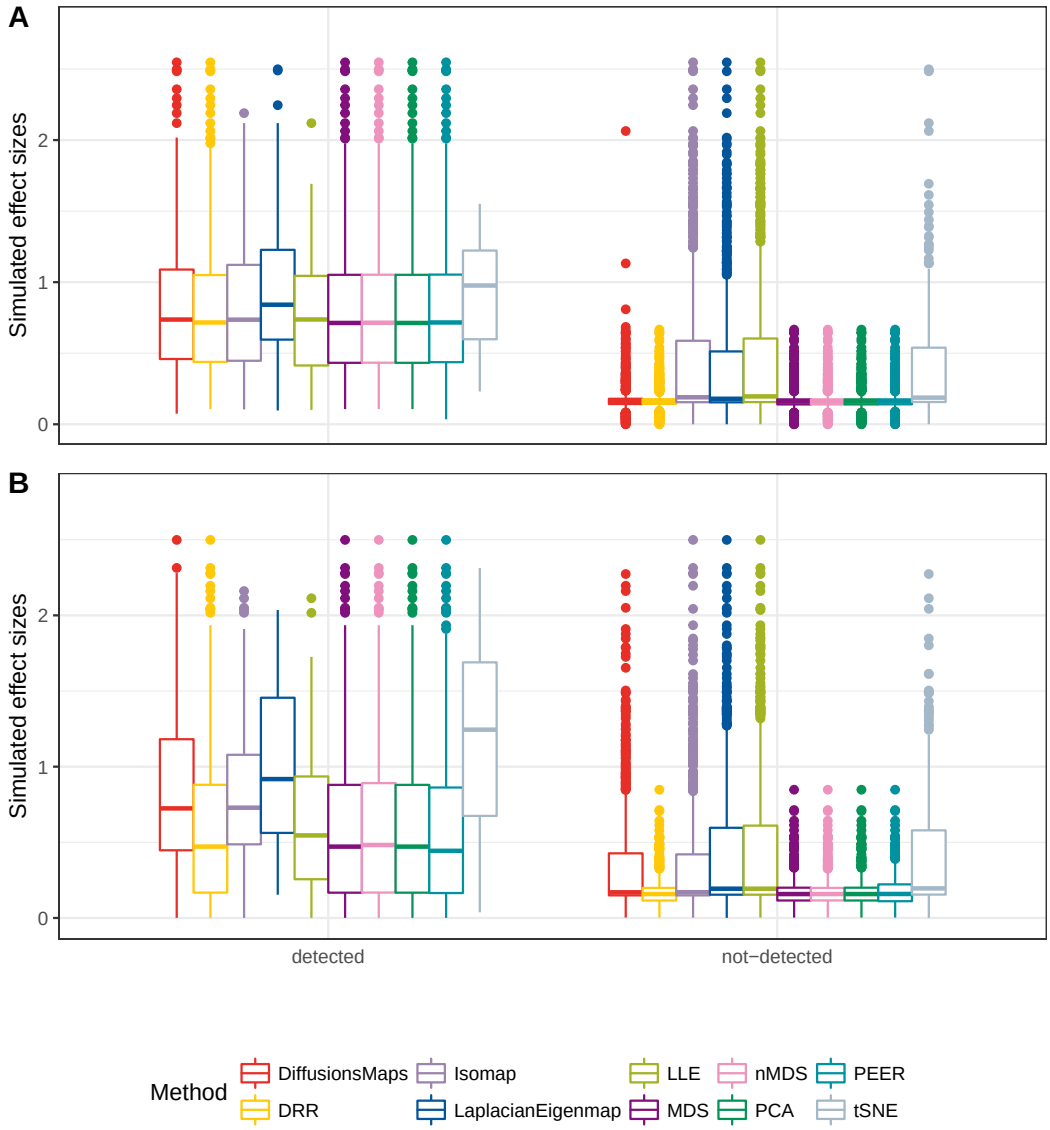

Figure 6.11: Effect size distribution of discovered SNPs. A. Power from datasets with correlated background structure. B. Power from datasets without correlated background structure. The mean of the simulated effect sizes per SNP across all traits was computed. SNPs were classified into "detected" and "not-detected" based on the threshold $p_{\text{empirical}} < 0.01$. The plot shows the dependence of detecting causal SNPs on their effect size for different background models and dimensionality reduction techniques across all proportions of traits affected by the SNPs.

trustworthiness and continuity criteria agree with the visual assessment of the methods' performance. Based on these results, I used the trustworthiness and continuity criteria to evaluate the effect of sample size and phenotype structure on the performance of the different methods. For the majority of methods analysed in this thesis, the sample size has only minor effects on the performance. In general, most models perform marginally better on data without correlated background structure. Trustworthiness and continuity are helpful in determining the correspondence of the high- and low-dimensional space. The stability criterion that I defined in this chapter evaluates a different aspect of the dimensionality reduction. It measures the number of components that can be reliably recovered in cross-validation and thus helps to determine the stable dimensions of the low-dimensional space. Applied to the two different data types, with and without background structure, it shows that background structure alone does not influence the number of stable components much. A stronger effect on the number of stably recovered components is observed when varying the proportions of traits influenced by the genetic variant effects. This seems intuitive since SNP effects are mathematically equivalent to any other type of fixed effect confounders that are present in the data. An increase in the proportion of traits affected generally leads to an increase in components recovered. This increase is maximal for about 40 to 80% of traits affected. This trend is reflected in the number of causal SNPs that can be detected when using the stable components as phenotypes in a genetic association model. The higher number of stable components at 40 to 80% of traits affected captures more of the underlying genetics.

In the analyses of stability and power to detect genetic associations, the linear and PCA-derived methods seemed to outperform the other methods. In particular, kPCA, ICA and tSNE yielded the least promising results: ICA did not recover any stable components, while the number was very low for kPCA and tSNE did only find stable components in the model of correlated background structure. In the association analyses, these components were either not associated at all (kPCA) or only for SNPs with large effect sizes (tSNE). However, there is a point of caution in these conclusions. Foremost, the performance of all these methods is intrinsically linked to the underlying data structure. Thus, in general, the different mathematical models of the dimensionality reduction methods will make some models more suitable for the analysis of a given dataset than others. In this simulation study, the high-dimensional datasets for stability and genetic association analyses are more similar to the *Iris* data than to the roll dataset. As such, it is encouraging that the chosen evaluation criteria trustworthiness and continuity show similar results for suitability of

the methods, i.e. linear methods seem to perform better on linear data than the non-linear methods. In addition, the non-linear methods all require the specification of model parameters for which I chose the default settings. Improved results might be observed when different parameter settings are evaluated for the different methods. To extend and improve this study, high-dimensional datasets more reflective of the non-linear structure of the roll dataset could be simulated and the non-linear dimensionality reduction methods evaluated on a range of parameter settings.

This simulation study has shown that dimensionality reduction methods are a valid intermediate step in genotype to phenotype mapping of high-dimensional datasets. Although methods like LiMMBo (chapter 4) enable association studies with large numbers of phenotypes, there is always a trade-off between exploiting correlated structure in the phenotypes and the joint mapping cost in form of degrees of freedom when evaluating the test statistic. Employing dimensionality reduction techniques to find the correlated background structures in the phenotypes while simultaneously reducing the degrees of freedom offers huge potential for the multivariate analysis of these phenotypic traits. For applications on real data, one should carefully evaluate different dimensionality reduction methods as the choice strongly depends on the data and investigate parameter settings to find components that best reflect the original data. The introduced stability criteria is particularly useful in genetic association studies as dimensionality reductions that are not stable are guaranteed not to produce reliable results.

7

GWAS of left ventricular wall thickness

The structure of the human heart is determined by an interplay of genetic factors and and complex environmental influences [Payne & al., 1995; Sanoudou & al., 2005; O'Toole & al., 2008]. One common, heritable trait used to predict clinically relevant heart conditions is left ventricular mass (LVM). In particular, the increase in LVM is associated with an increased risk of heart failure and sudden death [Haider & al., 1998; Post & al., 1997; Lorell & Carabello, 2000]. The increase in LVM through the thickening of the left ventricular wall is a direct response to a rise in hemodynamic burden which causes the hypertrophy of existing myocytes [Lorell & Carabello, 2000]. The thickening of the wall can occur in a symmetric fashion through concentric thickening of the ventricle with a small cavity dimension. However, about 58% of all cases of left ventricular hypertrophy are asymmetric [Davies & McKenna, 1995] and the observed asymmetry patterns are diverse in distribution and occurrence [Hughes, 2004; Florian & al., 2012]. A number of genetic factors have been shown to be involved in these asymmetric changes in the structure of the left ventricle [Davies & McKenna, 1995; Chen & Chien, 1999; van der Merwe & al., 2008]. To date, GWAS in African American [Fox & al., 2013], Caucasian [Vasan & al., 2007; Vasan & al., 2009; Arnett & al., 2009] and more recently Japanese cohorts [Sano & al., 2016] have attempted to identify genomic loci that are associated with LVM, where LVM was assessed using echocardiographic measures or 2D cardiac magnetic res-

onance imaging. However, none of the studies find associations that pass the commonly applied genome-wide significance threshold. Many factors might have influenced the success of the studies and the lack of finding genetic associations such as lack in power through small sample or effect size. Given the genetic effects of the clinical LVM phenotypes observed [Davies & McKenna, 1995; Chen & Chien, 1999; van der Merwe & al., 2008], the assumptions for a genetic contribution to the natural variation in heart morphology holds, despite the negative results obtained in these studies. However, the asymmetric nature of changes in heart morphology might make LVM an inaccurate phenotype for detecting these genetic effects. To investigate genetic influences on overall heart structure instead of on a reduced representation such as LVM, spatially resolved, quantitative heart phenotypes are needed.

A recent advance in cardiac MRI is the use of 3D imaging of the heart as a whole as opposed to multiple transverse sections of the heart by 2D imaging. The latter technique has been the clinical gold standard but recent studies have shown that 3D imaging improves spatial resolution especially at the base and apex of the heart (figure 2.1) and can avoid technical issues arising from 2D imaging [de Marvao & al., 2014]. Detailed images derived from the 3D imaging technique combined with genotype data would allow for an investigation into spatially-confined changes in heart morphology. Genetic association studies based on imaging phenotypes are widely applied in the field of neuroscience [Filippini & al., 2009; Ho & al., 2010; Jahanshad & al., 2013; Hibar & al., 2015]. The first unbiased study using genome-wide genetic markers to find genetic associations with brain activity patterns was conducted by Stein and colleagues. They associated every voxel of 3D brain scans with all genetic markers. Following this approach, associating heart morphology as represented in the 3D scans would require testing approximately 140,000 voxels. However, voxel-wise GWAS is limited in power and does not take into account any spatial correlation between the voxels [Ge & al., 2014].

To overcome these limitations and offer more practical measurements for clinical use, De Marvao and colleagues have developed a technique to extract 3D features of the cardiac morphology from the 3D scans [de Marvao & al., 2014]. As part of the *digital heart project* [Cook & O'Regan, 2010], they created the first at scale cohort of about 1,500 detailed 3D statistical models of the variation in cardiac morphology from healthy volunteers. Based on these models, standard clinically relevant measurements such as LVM can be computed. Far beyond these simple 1D measurements, the 3D models allow spatially derived phenotypes such as left-ventricular wall thickness or curvature to be resolved for over 27,000 coordinates. However, the substan-

tial challenge in handling this still large number of correlated dimensions present in these models remains.

In the following chapter, I describe the GWAS of phenotypes derived from the 3D statistical models of the *digital heart project*. Within this project, I was responsible for the quality control and imputation of the genotypes, and conducted the GWAS from the 3D phenotypes. My colleagues collected the DNA samples, performed MRI scans and provided the 3D phenotyping. I will first describe the genotyping and phenotyping strategy and then show the results from applying different dimensionality reduction techniques to the 3D heart phenotypes. Based on the criteria described in chapter 6, I chose the most suitable methods and conducted a GWAS with components derived thereof as proxy phenotypes. Finally, I investigated the associated loci for any spatial association with the 3D heart phenotypes.

Using the genotype information which I processed and imputed, a preliminary publication on genetic associations was accepted for publication [Biffi & al., 2017] and we are currently planning the publication of the analyses and results described in this chapter.

7.1. Data

7.1.1. Genotypes

Quality Control. Genotyping and genotype calling were carried out at the Genotyping and Microarray facility at the Wellcome Trust Sanger Institute, UK and Duke-NUS Medical School, Singapore. Genotypes were assessed in five batches using Illumina HumanOmniExpress- 12v1-1 (Sanger, two batches), Illumina HumanOmniExpress-24v1-0 (Duke-NUS, two batches) and Illumina HumanOmniExpress- 24v1-1 chips (Duke-NUS). SNPs were called via the GenCall software for clustering, calling and scoring of genotypes [Teo & al., 2007]. For batches run on the same platform, genotype signals were combined and called in a single analysis, leading to three independent genotype batches: Sanger12 (1,344 samples), Duke-NUS12 (284 samples), Duke-NUS3 (96 samples). I carried out the quality control (QC) on the raw genotype calls, the phasing and the imputation at a per-batch level. The final QC of the imputed data was conducted across all batches and only SNPs passing the control in every batch were used in subsequent analyses.

Prior to QC, I matched the rsID descriptions (chromosome, chromosomal positions and allele order) of the three batches to the reference set I would use for imputation, a combined UK10K [UK10K Consortium, 2015] and 1,000 Genomes [1000

Genomes Project Consortium, 2015] reference panel. For rsIDs not included in the reference panel I retrieved location and allele order from the ensembl human variation annotation (GRCh37p13, 15.04.2016). rsIDs that matched to neither reference were removed from further analyses (4,681 across all chips). In order to avoid batch effects in SNP calling simply based on the probe sequences, I confirmed that probes targeting the same SNP on different chip versions had the same sequence. As this was the case, no SNPs were removed at this stage. I followed an adapted quality control protocol from Anderson & al. [2010] to asses the quality of the genotyping on a per-individual and per-marker level. Unless stated otherwise, the PLINK software (version 1.9) [Purcell & al., 2007; Chang & al., 2015] was used for all QC analyses. In summary, the per-individual QC included the identification of individuals with discordant sex information, missing SNP rates (more than 3% of SNPs not called) and heterozygosity rate outliers (three standard deviations outside of the mean heterozygosity rate). Population substructures arising due to different ethnical origins of samples were examined by comparing the sample genotypes to genotypes from the HapMap Phase III study [The International HapMap Consortium, 2005] for four ethnic populations (with subpopulations, figure B.7 in the appendix). Samples that clustered with HapMap III individuals of European ancestry were kept for further analyses. The per-marker QC included filtering of SNPs with missing call rate in more than 1% of the samples and SNPs which significantly deviate from Hardy-Weinberg equilibrium (HWE, $p < 0.001$). After removing samples and SNPs that failed QC, I confirmed that any pattern of missing genotype information was not batch-specific. To analyse these patterns, I treated each pair-wise combination of batches as a case-control set-up and computed the differential missingness of SNPs common to all batches. None of the 631,877 common SNPs had to be removed due to significant differential missingness ($p < 10^{-5}$). Table 7.1 shows an overview of sample and SNP numbers before and after the QC described above. The QC plots for each step can be found in figures B.5 to B.7 in the appendix.

Phasing and imputation. Phasing and imputation were conducted in two separate steps. For phasing, I used SHAPEIT (version 2.r727) [Delaneau & al., 2012; Delaneau & al., 2013] to generate estimated haplotypes for each sample that passed the quality control. The window size for phasing was set to 2Mb, and the number of conditioning states per SNP to 200. All other parameters were set to default values. The phased genotypes were then imputed with IMPUTE2 (version 2.3.0) [Marchini & al.,

Table 7.1: Sample and SNP numbers before and after the QC. For each batch (first column), the number of male (m)/female (f) samples and SNPs before and after QC are listed. Rate specifies the genotyping rate of samples within one batch after QC.

	pre-QC		post-QC		
	samples (m/f)	SNPs	samples (m/f)	SNPs	Rate
Sanger12	1,344 (614/730)	719,665	998 (463/535)	677,036	0.998
Duke-NUS12	284 (118/166)	716,503	179 (68/111)	682,016	0.998
Duke-NUS3	96 (48/48)	7,713,014	62 (34/28)	657,497	0.998

2007; Howie & al., 2009] based on the combined 1,000 Genomes [1000 Genomes Project Consortium, 2015] and UK10K [UK10K Consortium, 2015] reference panel. I set the imputation interval to 3Mb, with a buffer region of 250kb on either side of the analysis interval. As suggested in the user manual, I used an effective population size of 20,000 and set the number of reference haplotypes to 1,000. Again, for the additional, non-specified parameters the default was used.

Combining datasets. I combined the three genotype batches after imputation and filtered them again on a per-sample and per-marker level. On the per-sample level, I excluded related individuals because of the difficulties that might arise in adjusting for relatedness in the processing of the phenotypes via dimensionality reduction. A more detailed explanation will follow in section 7.2. Relatedness was estimated by the proportion of SNPs shared between two individuals and subsequent calculation of IBD estimated as PI_HAT on the genotyped SNPs via PLINK as described by [Anderson & al., 2010]. For any pair of individuals with a PI_HAT of greater than 0.125, the individual with the higher SNP calling rate was retained in the analysis. For the quality control on the per-marker level, I used the statistical information about the imputation certainty, the *info* metric, given as additional output by IMPUTE2. The metric typically takes values between zero and one, with values closer to one indicating high imputation certainty. I excluded any SNP with an info score of less than 0.4 in at least one of the batches. Approximately 60% of all imputed SNPs were excluded based on this criterion. After combining the datasets, I used SNPTEST (v2.5) [Marchini & Howie, 2010] to compute the minor allele frequency (MAF) and p-value for deviation from Hardy-Weinberg equilibrium per SNP. SNPs with a significant deviation from Hardy-Weinberg equilibrium ($p < 0.001$) and a minor allele count of less than 20 alleles (corresponding to a minor allele frequency of 0.008) were removed, leading to a decrease in SNPs of another approximately 41%, a total

reduction from imputed SNPs to SNPs that passed every filtering criteria of 23%. A summary showing the magnitude of the number of imputed SNPs per batch, the number of SNPs after imputation quality filtering and filtering for MAF and Hardy-Weinberg-equilibrium deviation is depicted figure 7.1. Exact numbers can be found in table A.2.

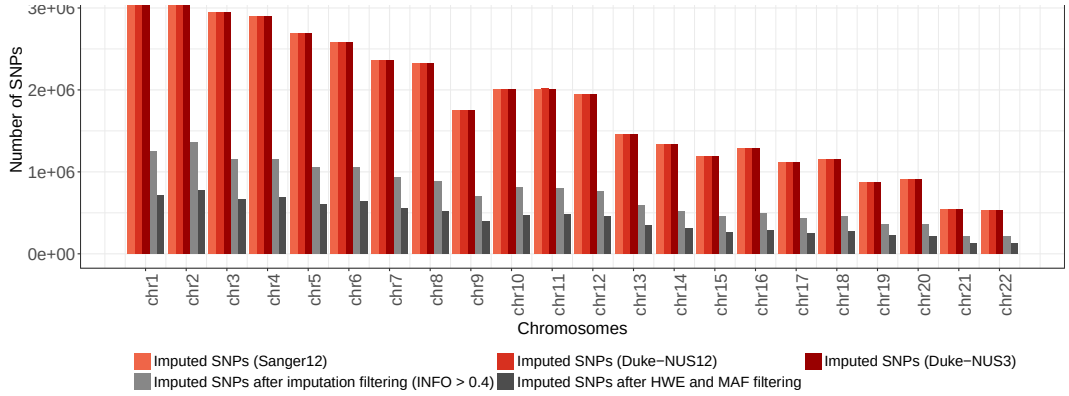

Figure 7.1: Overview of SNP numbers after imputation and imputation quality control. The imputation of the SNPs based on the genotypes from SNP arrays was done on a per-batch level. The number of SNPs for each batch after imputation is shown as red bars and is very similar for each of the three batches (exact numbers in table A.2). About 40% of SNPs are retained after filtering for the 'info' metric (light grey bars). The bars in dark grey show the final number of SNPs per chromosome.

After imputation and imputation quality control, the dataset contains 9,233,118 SNPs from 1,207 samples. IMPUTE2 yields imputed genotypes encoded in triplets of posterior probabilities for the possible allele combinations (AA, AB, BB). These probabilities were converted into expected genotypes G by the dosage model [Howie & al., 2011]:

$$G = 0 \times p(AA) + 1 \times p(AB) + 2 \times p(BB) = p(AB) + 2 \times p(BB) \tag{7.1}$$

7.1.2. Phenotypes

The phenotyping was done by my collaborators, in particular Antonio de Marvao. CMR imaging and generation of 3D models of the left ventricle derived from these images were conducted at Hammersmith Hospital, London. In the following, I will briefly describe the methodology of their automatic phenotyping approach. The technical details of the image acquisition, the analysis and their improved performance over standard methods are described in detail in [de Marvao & al., 2014].

In the automated phenotyping approach developed by Antonio de Marvao and colleagues, cardiac structures are accurately extracted from raw 3D cadiac magnetic resonance images via a multi-atlas PatchMatch (figure 7.2, 1) to generate 3D models of the individuals' hearts (figure 7.2, 3). The cardiac structures of interest in this study were left ventricular cavity, myocardium and right ventricular bloodpool at end-diastole and end-systole. The multi-atlas PatchMatch algorithm uses a local database of segmented and quality controlled cardiac MRI atlases, to which each newly acquired image is compared. The database of atlases was created by Antonio, who initially selected 20 subjects and manually labelled the approximately 140,000 voxels per image into the three categories named above (left ventricular cavity, myocardium and right ventricular bloodpool). These manually classified images were then divided into smaller patches – atlases– which served as the initial training dataset for the segmentation algorithm. In the database generation phase, subsequent successful segmentation of new images described by the method below were added, yielding a total of 1,072 images in the final database. In addition to serving as a database for the segmentation algorithm, the database images were used to generate a template image of average heart size, position and orientation.

For each new image, six landmarks are manually placed on the image, which enables the subsequent image registration between the target and the atlas images. After registration, a multi-atlas PatchMatch algorithm finds corresponding patches of adjacent voxels within the atlas and target images (figure 7.2,2). Each patch in the target image is given the label of the closest matching atlas patches and combining the labels of all patches produces the final segmentation. Lastly, the segmented image is registered to the template image to make the spatial coordinates in the 3D models consistent between all samples.

Using a surface rendering algorithm allows for the extraction of information from a segmentation volume such as the left ventricular myocardium into a surface representation. Through such an algorithm, the wall thickness, curvature and fractional wall thickening at 27,623 positions in the left ventricle were extracted for each individual (figure 7.22).

To assess the reproducibility of the phenotyping approach, one individual was scanned eight times and the images segmented as described above. These repeat scans allowed for the quantification of variation in the segmentation by the coefficient of variation (CV). The CV is a standardised measure of dispersion and is defined as the ratio of the standard deviation to the mean value. I computed the CV for each

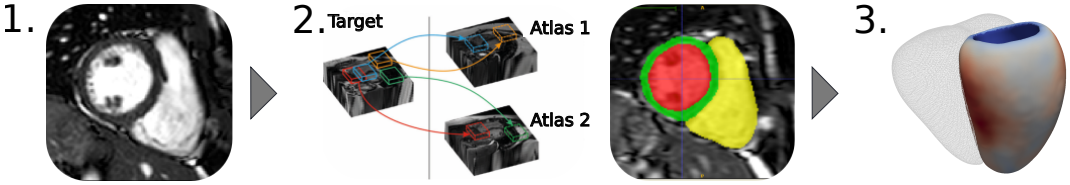

Figure 7.2: Cardiac phenotyping based on cardiac magnetic resonance images. 1. Detailed 3D images of the heart were acquired in the left ventricular short axis plane from base to apex. 2. The images were segmented into left ventricular myocardium (green), left ventricular blood pool (red) and right ventricular blood pool (yellow) and registered to a common template image via a multi atlas-based technique. 3. Through a surface rendering algorithm of the registered segmentation, a 3D model of the heart was generated and wall thickness measurements derived at 27,623 positions of the left ventricle. The left ventricle is shown in solid colors, with the color scheme representing average wall thickness, increasing from light to darker colors. As a point of reference, the right ventricle is depicted as a mesh.

of the 27,623 positions in the 3D heart model across the eight scans and projected the results onto the template image (figure 7.3). Overall, the dispersion is very low i.e. the reproducibility high. Only at the base of the left ventricle in proximity to the right ventricle can a slight increase in dispersion be observed (figure 7.3, red area). The low dispersion shows the accuracy of the segmentation and surface rendering methods. Based on this result and further quality control criteria such as the comparison between the segmentations and manually labelled images (details in [de Marvao & al., 2014] the wall thickness measurements were considered reliable phenotypes for subsequent analyses.

Wall thickness measurements were successfully extracted for 1,185 of the 1,207 individuals that passed the genotyping quality control.

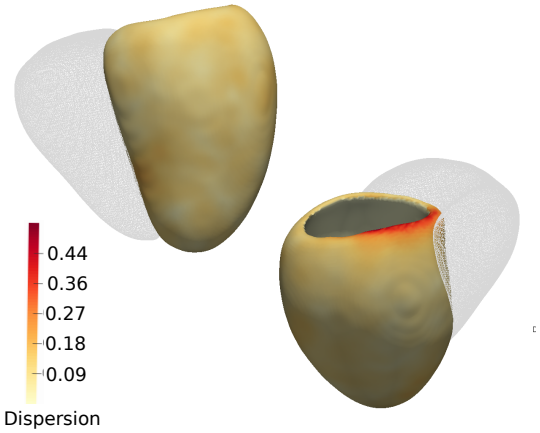

Figure 7.3: Phenotype reproducibility. The dispersion in left ventricular wall thickness at 27,623 positions was computed as the standard deviation over the mean across eight segmentation derived from independent scans of one individual. The right ventricle is shown as a point of reference (mesh structure).

7.2. Dimensionality reduction yields stable low-dimensional phenotype representations

The detailed 3D models of the heart structure offer a rich dataset for investigating spatially-resolved genetic associations on cardiac morphology. By extracting the relevant features from the cardiac magnetic resonance images, the phenotype space has been reduced from intensity values at 140,000 voxels to wall thickness measurements at about 27,000 3D coordinates. While this processing condensed the original image space into relevant phenotype information, considering each position as a phenotype would still require 2×10^5 single-trait association tests which have to be adjusted for multiple testing and which would not be able to take advantage of correlation structure in the phenotypes. In contrast, a multi-trait association test would be more powerful by modelling the correlated traits jointly, however its test-statistic would be subjected to a 2×10^5 degree of freedom test. To avoid this burden of correcting for the high-dimensionality of the traits while making use of intrinsic structure in the data, I applied the twelve dimensionality reduction methods tested in chapter 6 to the 27,623 heart wall thickness measurements in order to find the best low-dimensional representation of the dataset. The low-dimensional components will then serve as proxy phenotypes in the GWAS.

Before applying the dimensionality reduction methods, I adjusted each of the 27,623 left ventricular wall thickness measures independently for any known cov-

ariates such that the low-dimensional components ideally only reflect structure truly related to the underlying cardiac biology. Any covariates with an assumed linear effect on the wall thickness were used as explanatory variables in a linear model with wall thickness as a response variable. These include the biological covariates sex (643/542, female/male), age (40.3 ± 13.3 years, mean $\pm$ standard deviation), height (170.8 ± 9.3 cm) and weight (71.9 ± 13 kg), and technical covariates MRI operator, date of the image acquisition and date of the image segmentation. Figure 7.4 summarises these covariates by their respective univariate distribution (diagonal) and their dependent distributions across the 1,185 genotyped and phenotyped individuals in the cohort.

Other, more complicated covariance structure could arise due to related individuals in the dataset. In order to avoid confounding of subsequent analyses potentially introduced through high levels of relatedness between a number of individuals, related samples were removed from the analysis based on the quality control of the genotypes (section 7.1.1).

The parameters for the dimensionality reduction were chosen as in table 6.2 and the maximum dimension set to 100. The optimal number of neighbours was estimated as $n = 40$. The dimensionality reduction was performed on the residuals of the linear regression described above. I used the new stability criterion introduced in section 6.4.1 to find the low-dimensional representations that can be reliably recovered in subsets of the dataset. As described for the simulations (section 6.4.1), I split the dataset into subsets of 80% of the samples, computed the dimensionality reduction and repeated this step ten times. For each cross-validation, I computed the trustworthiness (equation (6.2)) and continuity (equation (6.3)). Overall, I used the cross-validation to determine the stability. ICA on this dataset with the fastICA::fastICA function in R was not possible and failed with fortran indexing errors. As the dimensionality reduction with ICA yielded the least stable results in the previous chapter, this dimensionality reduction strategy was not investigated further on the heart data.

An initial look at the number of stable components showed a median of ten stable components across all methods. As a first manual control of the low-dimensional representation, I qualitatively analysed the distribution and pair-wise density of the first ten dimensions for each method. While most methods showed a similar spread and distribution of their components with differing levels of correlation, components from kPCA and DiffusionMaps were clear outliers from this observation. Figure 7.5 shows the pairs-wise comparisons for components from DiffusionMaps and kPCA

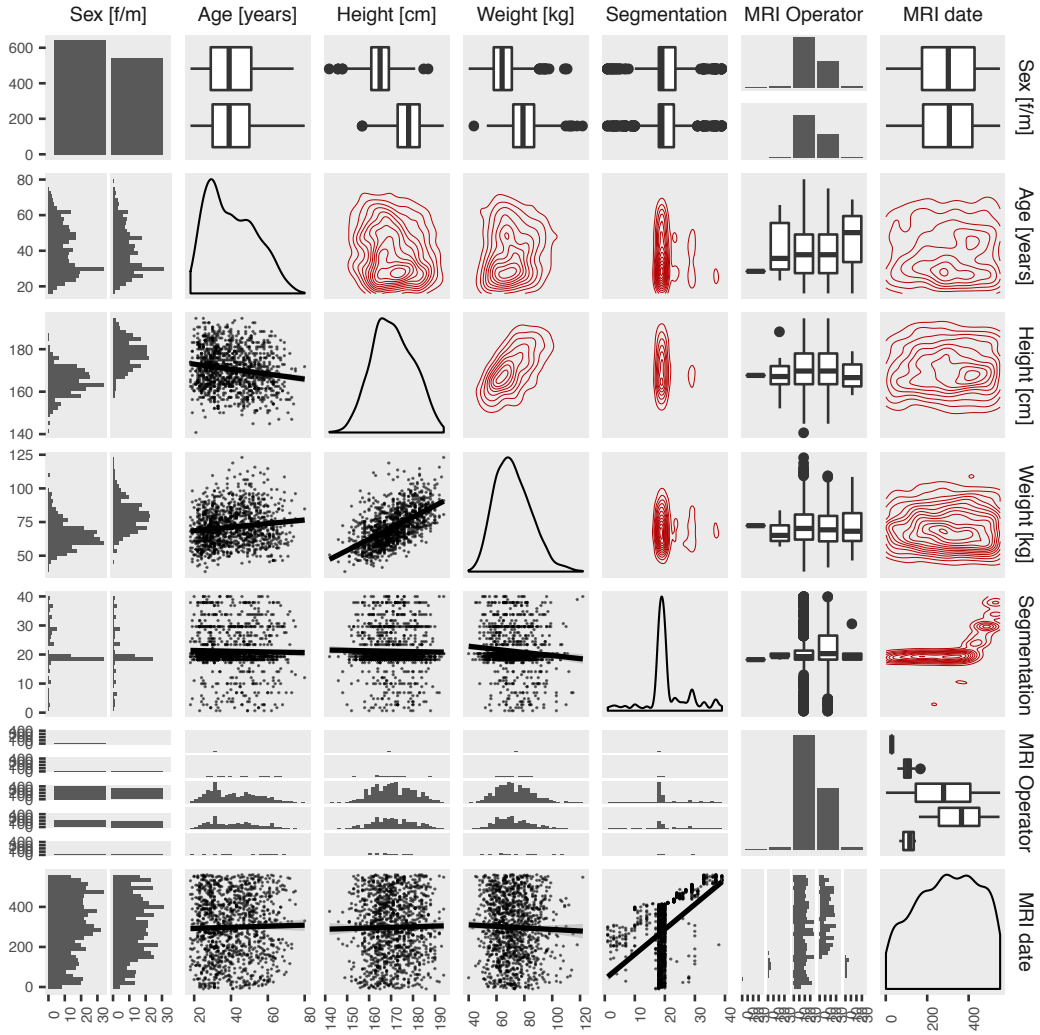

Figure 7.4: Distribution of covariates in 3D heart phenotype cohort. Continuous variables: the univariate-distribution of each variable is depicted on the diagonal. The upper triangular matrix shows the bi-variate distribution while the lower triangular matrix shows the regression line of their linear fit. Categorical variables: Distribution (row) and counts (column) are depicted.

as well as Laplacian Eigenmaps and PCA as references for well-behaved methods. For the PCA data (figure 7.5B), components show the widely uncorrelated behaviour expected of orthogonal vectors (section 6.1). Components derived from Laplacian Eigenmaps (figure 7.5D) display different levels of correlation, from mostly uncorrelated (DR6 vs DR10) to strong non-linear correlation (DR1 vs DR2). DiffusionMaps and kPCA show very little spread in their data, with the distribution of each component spiking at a particular value and zero elsewhere (Figure 7.5A,C, diagonal). Similar plots for the other, well-behaved methods can be found in figure B.3 in the appendix. Based on these observations and without a clear indication as to why these results were observed (i.e. no warnings or error messages in the computation), components from DiffusionMaps and kPCA were not considered in further analyses.

For the majority of methods, the low-dimensional representation has a high level of trustworthiness, with seven methods above 90% for each cross-validation steps and the full dataset (figure 7.6A, boxplots and diamond shape). Only PEER does not reach that threshold. The same result is observed for continuity, with exception for LLE, whose low-dimensional representation of the full dataset does not lie above 90% (figure 7.6B). To provide a consistent *a priori* selection of methods, I only considered stable components reliable if their continuity and trustworthiness measures were above 90% for both the full dataset and the cross-validations. Based on these criteria, components retrieved from DRR (ten), MDS (ten), Isomap (four), Laplacian Eigenmaps (five), PCA (ten) and nMDS (ten) were considered for further analyses.

As demonstrated in chapter 6 and figure 6.2, there is a considerable degree of similarity in the low-dimensional representations for some of the methods tested, especially the linear and PCA-based methods. I analysed the extend of similarities between the stable components from the six methods passing the trustworthiness/continuity threshold by computing the pair-wise Pearson correlation based on the absolute value of their components. The stable components from PCA, MDS and nMDS show perfect correlation, as expected given the strong mathematical similarity of these methods when using Euclidean distance as the distance measure. Isomap, which builds the bridge between the linear and non-linear models as it is based on MDS and kernel-eigenmaps (section 6.1) shows weaker but still strong correlation to the first three methods. Components derived from Laplacian Eigenmaps are only weakly correlated with those from any other method. However, chapter 6 and figures 6.2 and 6.4 also demonstrated the differences in low-dimensional representation for the other methods, in particular between linear and non-linear methods.

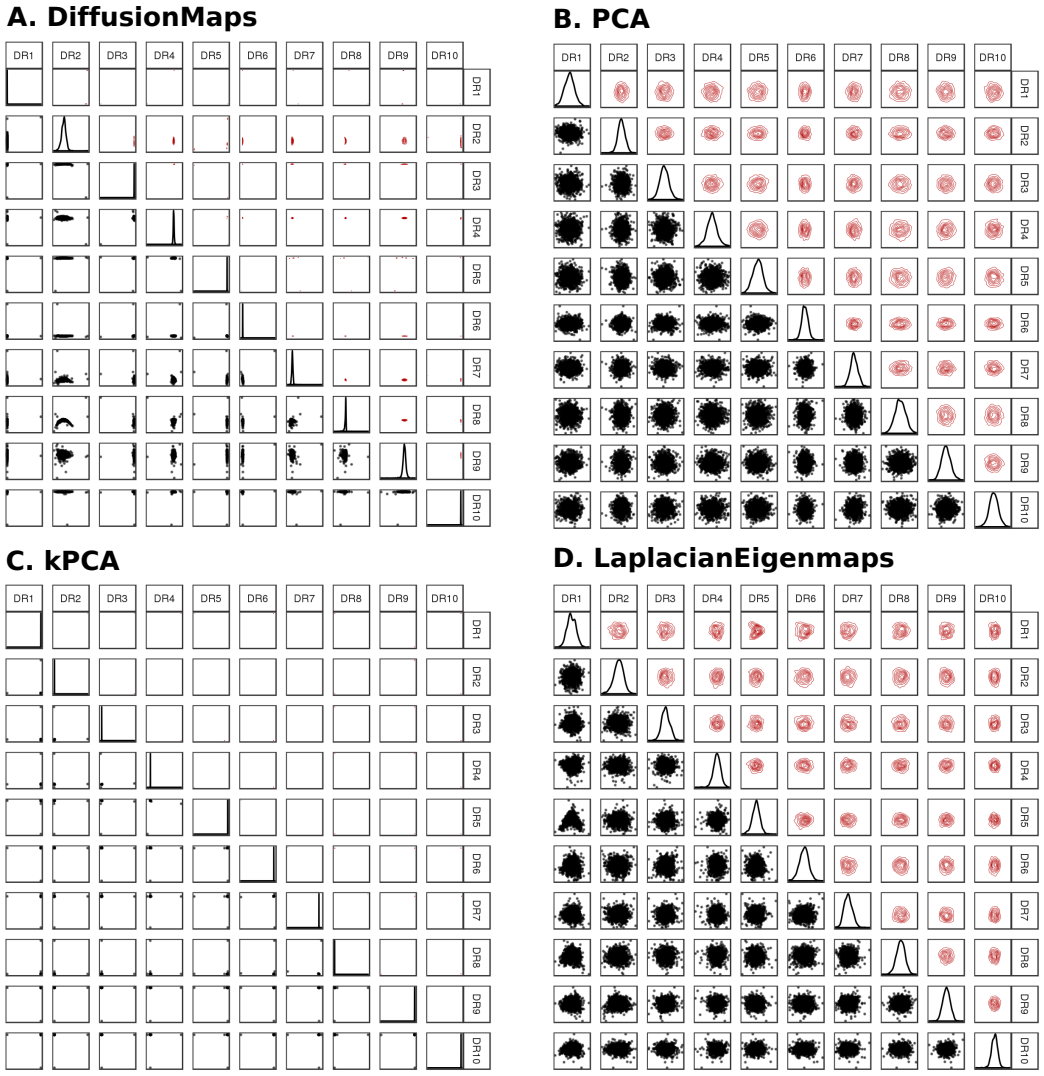

Figure 7.5: Pair-wise scatterplots of low-dimensional components derived from left-ventricular wall thickness. For components from DiffusionMaps (A), PCA (B), kPCA (C) and LaplacianEigenmaps (D), pairwise scatter plots of the components (lower triangle) and density plots (upper triangle) are depicted. The diagonal of each plot shows the distribution of the respective component. Row and column labels specify the rank of the component out of the 100 low-dimensional components. Before plotting, each component was mean-centred and divided by its standard deviation in order to have comparable axis dimensions. Given the normalised scale of the data, and the purpose of qualitative comparison, axis ticks were omitted for a cleaner visualisation.

Without prior knowledge about the true biological features, i.e. the "real" low-dimensional manifold of the left ventricular wall thickness measurements, it is not possible to know which methods will be most suitable in capturing this manifold

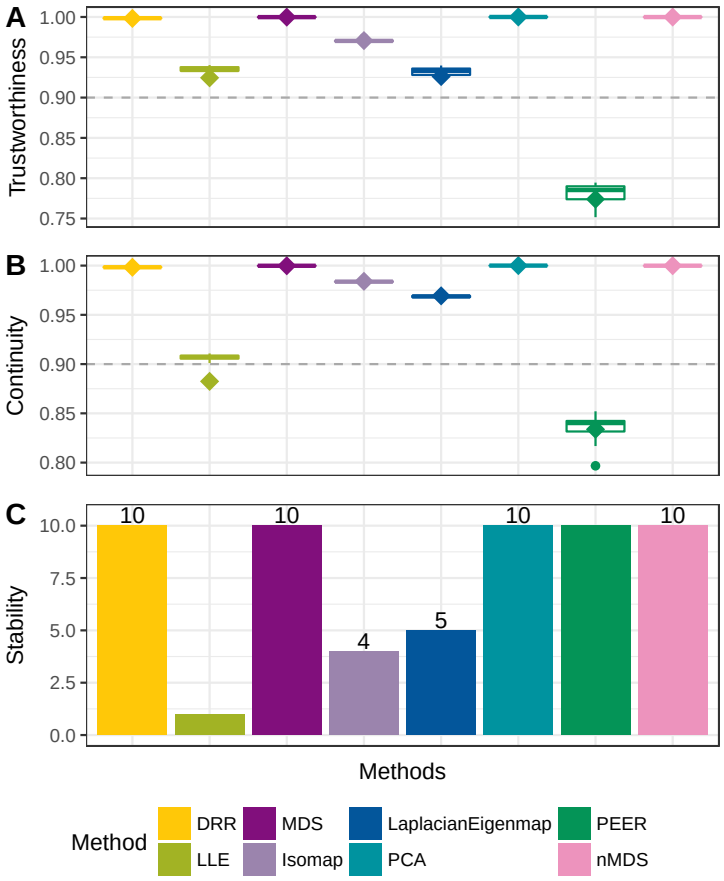

Figure 7.6: Dimensionality reduction of 3D heart phenotypes. The boxplots in A. and B. show the maximum trustworthiness and continuity across neighbourhood sizes ranging from 1 to 3% of the samples for the ten cross-validation sets for each method. The diamonds show the respective measures for the full dataset. Dotted lines are drawn at 0.9, the threshold chosen here at which a projection is considered a good representation of the original space. C. The number of traits passing the stability criterion. For methods that passed the threshold for both continuity and trustworthiness in the full dataset, the number of stable traits is printed above the bar chart. The corresponding traits are taken as input for the multi-trait GWAS.

structure. Instead of choosing a single method to find components to represent the manifold, I combined all components from the models above that pass the stability criterion. From the group of highly to perfectly correlated methods (DRR and PCA, MDS, nMDS), I choose the components from PCA as it has no parameters to specify. Thus, the final low-dimensional representation of the 27,623 left-ventricular wall thickness measurements is comprised of ten stable components from PCA, four from Isomap and five from Laplacian Eigenmaps, a total of 19 dimensions.

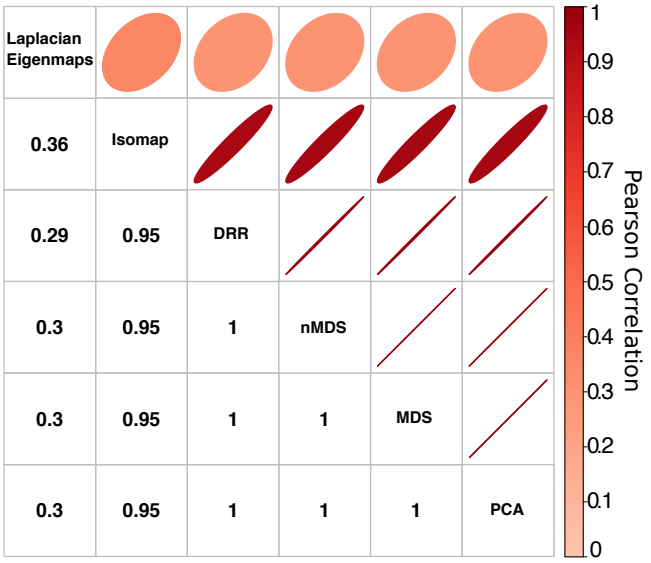

Figure 7.7: Correlation of low-dimensional components across methods. The Pearson correlation coefficients across the stable components of methods that passed the continuity and trustworthiness criteria were computed. The ellipses above show the mean strength of the absolute value of the correlation across all components. For the comparison of PCA, nMDS, MDS, DRR and DiffusionMaps (ten components each) to Isomap and Laplacian Eigenmaps (three and five components), the first or three five components were chosen for comparison.

7.3. Multi-trait GWAS detects three loci associated with heart wall thickness

Treating the 19 components as proxies for the true phenotypes, I was then able to conduct a mtGWAS to capture the genetics of left ventricular wall thickness. Based on previous studies [Price & al., 2006; Patterson & al., 2006] and results obtained in section 4.6 and figure 4.5, we know that mtGWAS is well calibrated in cohorts with little population structure and no relatedness. In order to avoid confounding relationship structure in the dimensionality reduction step, I had already removed related individuals and individuals that were not of European ancestry (section 7.1.1). Given this genotype structure, I used a simple linear model with components as response and genotypes as explanatory variables for the mtGWAS of the low-dimensional heart phenotypes. As there are no prior assumptions about the genotype effects, I modelled the SNP effects based on an any effect design matrix (section 1.7.8).

The results of the mtGWAS are depicted in figure 7.8, with three loci that pass the genome-wide significance level of 5×10^{-8}. The qq-plot in figure 7.9 shows a

well-calibrated test statistic.

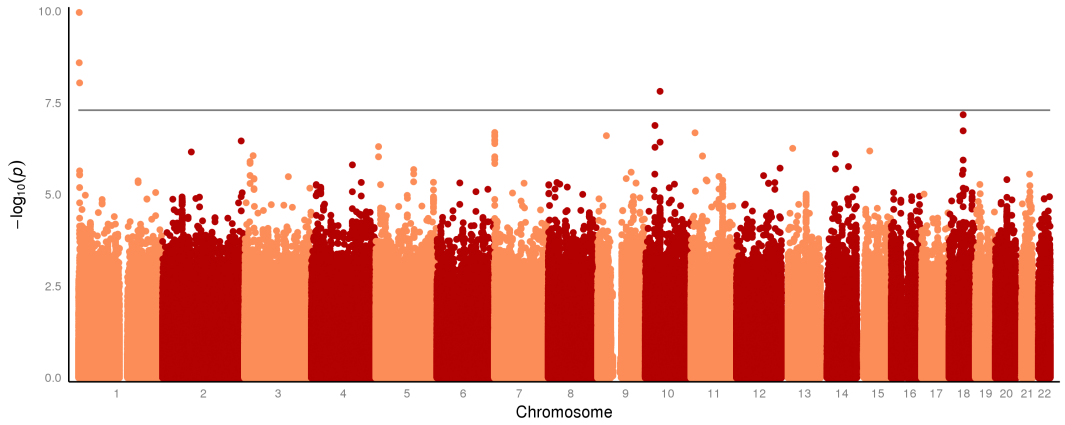

Figure 7.8: Manhattan plot of the multi-trait GWAS on 3D heart phenotypes. The 19 stable components derived from PCA, Isomap and Laplacian Eigenmap were modelled jointly in an any effect mtGWAS. The p-values of all genome-wide SNPs are depicted. The horizontal grey line is drawn at the level of genome-wide significance: $p = 5 \times 10^{-8}$. Two loci on chromosome 1 and one locus on chromosome 10 pass the genome-wide significance level.

Table 7.2 summarises the chromosomal location, p-values and SNP information of the most strongly associated SNPs per locus. Their genomic context is displayed in figure 7.10. The locus with the strongest association is located in a regulatory region of a gene-rich area between the *SKI* gene on the forward and the *MORN1* gene on the reverse strand (figure 7.10, upper panel; figure 7.11). *SKI* is developmental gene where *de novo* mutations are associated with a complex early developmental syndrome (Shprintzen-Goldberg syndrome) with cranofacial, bone development and cardiovascular phenotypes [Greally, 1993]. Zebrafish knockdown models of *SKI* orthologs give rise to complex developmental changes, including cardiac phenotypes [Doyle & al., 2012]. In addition, a non-developmental phenotype for altered expression of a *SKI* orthologues was observed in rat cardiomyocytes. In this system, the overexpression of the rat *SKI* orthologue leads to a decrease in fibroblast-to-myofibroblast phenoconversion, the main mechanisms for fibrotic heart disease [Cunnington & al., 2010; Cunnington & al., 2014; Zeglinski & al., 2016]. Taken together, these studies show an involvement of *SKI* genes in a variety of cardiac phenotypes across different tissues stages. The other gene in proximity to rs139971383, the *MORN1* gene, is relatively unstudied.

The second locus on chromosome 1 lies within intron nine of the *MEGF6* gene (figure 7.10, middle panel), which encodes for a secreted, calcium-iron binding protein [Nakayama & al., 1998]. It is also in proximity to the *PRDM16* gene, wherein

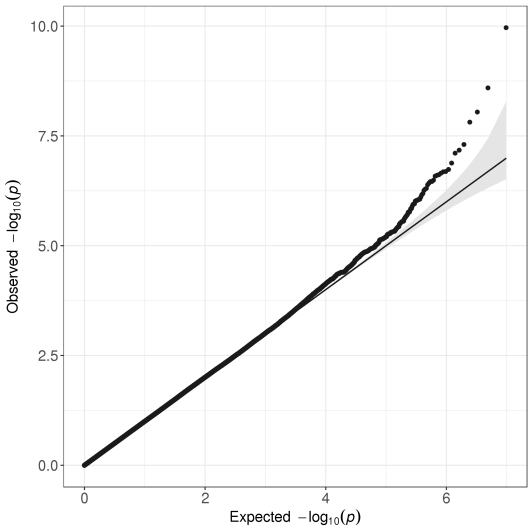

Figure 7.9: Quantile-quantile plot of the multi-trait GWAS on 3D heart phenotypes. The observed genome-wide p-values are plotted against p-values drawn from a uniform distribution in $[0, 1]$ of the same sample size (expected p-values). The diagonal line starts at the origin and has slope one.

deletions and mutations were shown to be implicated in two types of cardiomyopathies, left ventricular non-compaction (section 8.1) and dilated cardiomyopathy (section 2.4) [Arndt & al., 2013]. Based on zebrafish models of the observed human genotypes, the authors propose that *PRDM16* mutations lead to a decreased proliferative capacity during cardiogenesis. Interestingly, the study also found a link between the *SKI* and *PRDM16* genes, suggesting a functional synergy that leads to decreased cardiac output in zebrafish models with knock-down phenotypes of *SKI* and *PRDM16*. rs143266802 is located downstream of the zinc finger protein-encoding gene *ZNF487* (figure 7.10, lower panel), which has no associated phenotypes in human (GRCh38.p10, ensembl release 90, [Aken & al., 2016]).

A database search of the GWAS catalogue [MacArthur & al., 2017] (based on entries in the GWAS catalogue, 0.7.08.2018) and the Global Biobank engine, a resource for estimated genetic effects on cancers, autoimmune diseases, psychiatric, neurological, and cardiometabolic diseases [GBE, 2017] did not yield any other phenotypes that these SNPs were associated with.

The mvLM per SNP yields individual effect size estimates for each trait that is jointly modelled. There are two ways by which these effect size estimates can be helpful in understanding the genotype-phenotype association. Firstly, traits driving the association with the SNP are expected to have high effect size estimates.

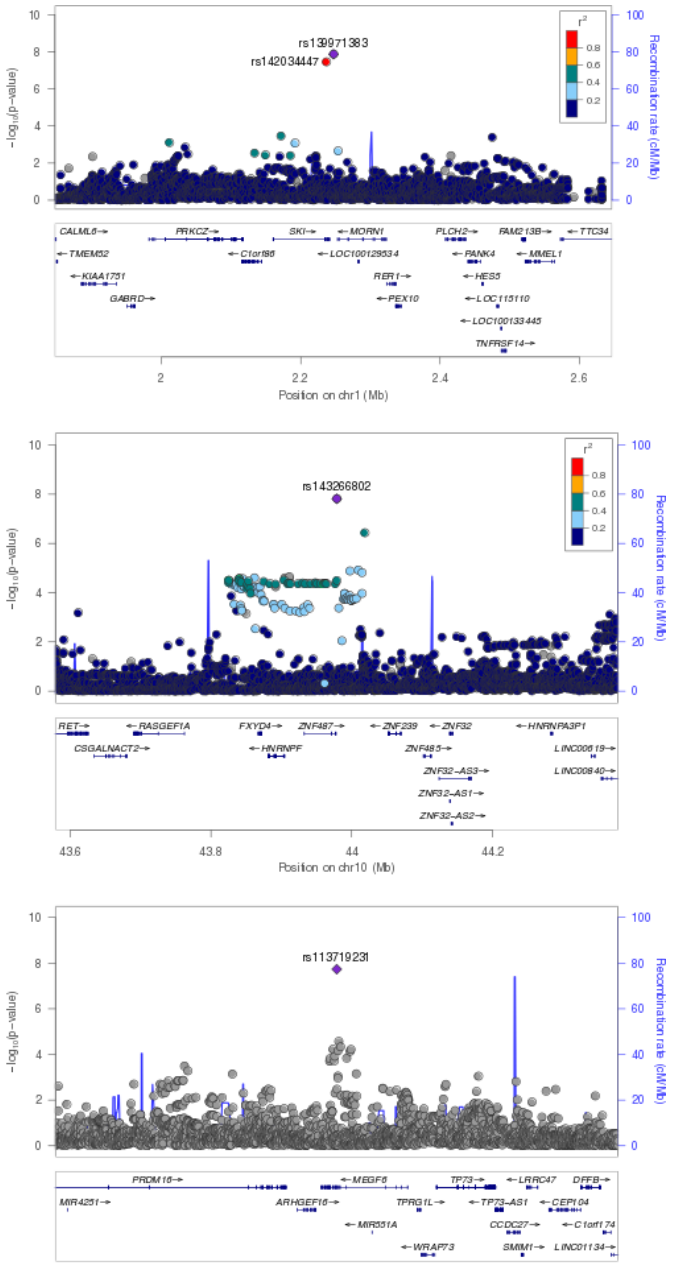

Figure 7.10: Genomic context of loci associated loci with 3D heart phenotypes . The p-values and genomic location of the three associated loci from the mtGWAS on the stable components from PCA, Laplacian Eigenmaps and Isomap are shown in relation to the p-values of surrounding genotypic markers. Markers are coloured by the level of LD they share with the SNP of interest. There was no LD information available on LocusZoom for the locus depicted in the bottom panel. Generated with LocusZoom [Pruim & al., 2010].

Table 7.2: Strongest genotype-phenotype association per locus for 3D heart GWAS. For each locus, the p-values for SNPs in LD with an $r^2 > 0.8$ in a 50kb window were compared and only the SNP with smallest p-value per locus listed below. Gene: gene in proximity to SNP and described in detail in the text above. M: major allele, m: minor allele, MAF: minor allele frequency.

SNP	Gene	Chr	Position	P-value	M/m allele	MAF
rs139971383	*SKI*	1	2,246,921	1.09×10^{-10}	C/G	0.013
rs113719231	*PRDM16*	1	3,427,138	9.04×10^{-9}	C/T	0.11
rs143266802	*ZNF487*	10	43,978,849	1.54×10^{-8}	C/T	0.022

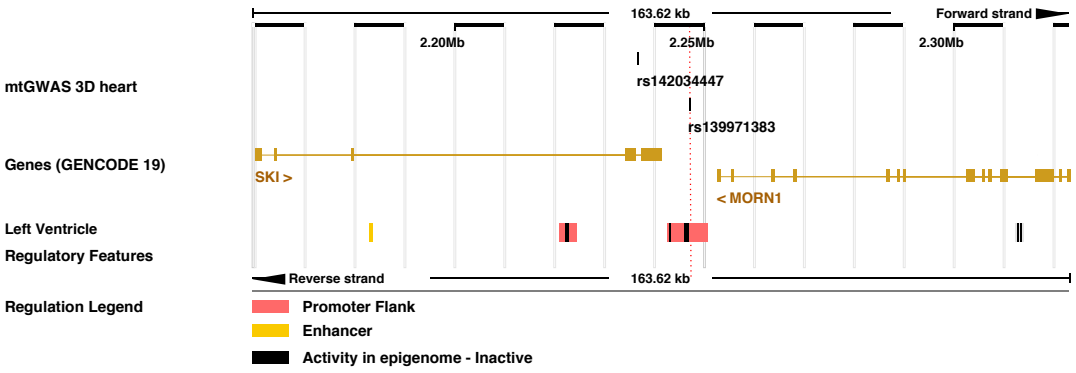

Figure 7.11: Regulatory context of locus with strongest association. The SNP with the strongest association (rs139971383) in the mtGWAS lies in a promoter flanking region epigenetically active in myocytes from the left ventricle (Ensembl, Human Regulatory Features, GRCh37.p13).

Secondly, traits that are similarly affected by the SNPs will have similar effect size estimates. In figure 7.12, I show the effect sizes for each of the 19 components per SNP clustered by their Euclidean distance. For the locus with the strongest associated with the 19 proxy phenotypes of wall thickness, there are two clusters of high effect size estimates (figure 7.12A, rs139971383). While one of them contains components from one method only (LaplacianEigenmaps1, 2, and 6), the other cluster contains two components from different methods, Isomap1 and PCA1. Similarily, the association of rs143266802 seems to be driven by a combination of components from all three methods (PCA2, Isomap2 and LaplacianEigenmap3). These results demonstrate the strength of this analysis approach, where different aspects of phenotype morphology are captured by different methods that can then jointly represent a wider aspect of the phenotype structure. The corresponding trait correlations are shown in figure 7.12B. A number of effect size clusters can seemingly be explained by the strong correlation of their respective traits (indicated by coloured boxes). In contrast, independent analysis of components from a single method, could not detect these strong signals (figure B.8). Only the locus situated in the regulatory region between *MORN*1 and *SKI* was detected in a mtGWAS with the components from Laplacian Eigenmaps alone (figure B.8A); p-value: 1.36×10^{-8}), confirming the effect size cluster structure observed for this locus, with large effect size for LaplacianEigenmaps1, 2, and 6. Additional signal for this independent analysis was overall weaker than the one for the combined analyses. GWAS with components from Isomap and PCA alone did not yield any associations (figure B.8B and C in the appendix). The proxy phenotypes are critical for the discovery of the genetic association but do not necessarily represent a biologically meaningful conformation. In order to understand the effect on the underlying biology without mediation via the dimensionality reduction methods, I linked the SNPs back to the original heart phenotypes.

In a first, simple approach, I used the discovered SNPs as explanatory variables in a simple linear model with left ventricular mass as the phenotype and sex, age, height and weight as additional covariates. None of the three SNP discovered with the mtGWAS shows association with left ventricular mass (rs139971383: $p = 0.89$, rs11371923: $p = 0.22$, rs143266802: $p = 0.68$). This result is not discouraging, however, since the hypothesis was that stable components capture regional variation in left ventricular wall thickness. Summarising wall thickness variation in a single scalar value such as left ventricular mass might not be able to capture these regional changes in mass. In order to analyse if the discovered SNPs show association

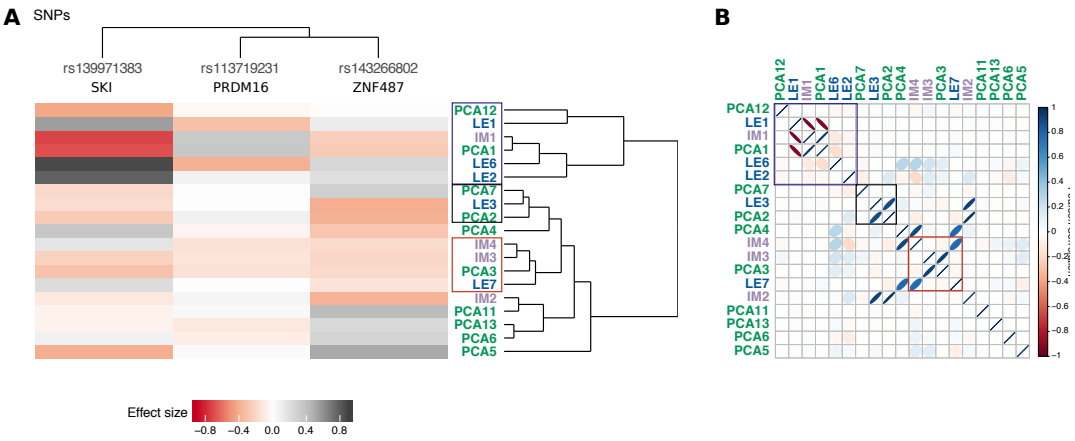

Figure 7.12: Effect size estimates and trait correlation from the 3D heart GWAS. A. The effect size estimates from the most strongly associated SNPs at each locus were clustered across components and SNPs by average-linkage hierarchical clustering of their Euclidean distances. The dendrogram of the components is labelled based on the methods used to generate the low-dimensional representation. The numbering indicates the position of the component as returned from the algorithm, i.e. for PCA the ordering based on the amount of variance explained. LE: Laplacian Eigenmaps; IM: Isomap. B. Trait-trait correlations of the corresponding low-dimensional representations used as response variable in the GWAS. For comparison, the order of the traits was matched to the clustering of the effect sizes. Effect size clustered where strong trait-trait correlations are observed are indicated by colour-matched boxes.

specific to certain regions in the left ventricle, I evaluated the relationship between the genotypes of the strongest associated SNP and the original, spatially-resolved left ventricular wall thickness measurements. For each of the 27,623 positions, I conducted a simple linear model with covariate-adjusted wall thickness measurements (data identical to input data for dimensionality reduction, section 7.2) as the response variable and the genotype of rs139971383 as the explanatory variable. Figure 7.13 shows these associations with the SNP in relation to their location on the left ventricle. Importantly, although none of these associations would be likely to survive the large multiple testing burden if used for discovery, they do show a specific localisation to the left ventricle which is affected by this SNP.

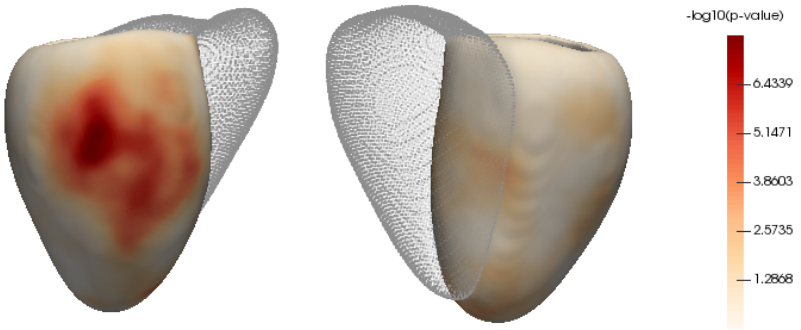

Figure 7.13: Association of rs139971383 with left ventricular wall thickness. The 27,623 covariate-adjusted wall thickness measurements in the left ventricle were used as the response variable in a simple linear model with the genotype of rs139971383 as the explanatory variable. The $-log10(\text{p-value})$ of the association of each models is projected onto its corresponding 3D position. Darker colors indicate stronger associations. Generated with ParaView.

7.4. Successful imaging genetics of cardiac phenotypes

In this chapter, I have described the step-wise feature extraction from high-dimensional cardiac magnetic resonance images of 140,000 voxels to a low-dimensional representation comprised of 19 components from linear and non-linear dimensionality reduction methods. The initial, atlas-based image segmentation of the original cardiac magnetic resonance images yielded reliable cardiac phenotypes at more than 27,000 positions in the left ventricle. One such phenotype was the left ventricular wall thickness, which I transformed into a significantly lower-dimensional component space by applying a variety of dimensionality reduction methods with different properties and consequently different low-dimensional representations. Using the three measures I introduced in chapter 6, I was able to make a principled decision about which low dimensional features to retain for further investigations into the genetics. Combining all stable and trustworthy components from different dimensionality reduction methods provided a robustness to the phenotypes which allowed for qualitatively different, latent cardiac structures to be represented in the final phenotype. I successfully mapped genotypes to these 19 phenotypes in a mtGWAS that detected three significantly associated loci. In order to link these genetic associations back to the observed wall thickness phenotypes, I associated the strongest genetic link with each wall thickness measurement and discovered a region highly associated with this SNP.

These results are promising for genetic association studies of very high-dimensional and correlated phenotypes, as well as for this specific study on cardiac morphology. In the emerging field of imaging genetics [Ge & al., 2014], the phenotype space ranges from simple photographs of face morphology [Liu & al., 2012; Shaffer & al., 2016] to functional MRI scans of brain activity [Stein & al., 2010; Hibar & al., 2015]. While each of these phenotypes are generated by different methods and will be subjected to different challenges in acquisition and quality control, the ultimate challenge lies in handling the high dimensionality of the phenotypes. The dimensionality reduction methods tested on the simulated data in chapter 6 and the 3D heart dataset in this chapter are all publicly available and can be readily applied to any fully phenotyped dataset.

As well as a practical example of the dimensionality reduction methods, the results of this specific combination of dimensionality reduction with GWAS are of great interest to my cardiac biology collaborators. The pre-existing cardiac related phenotypes of *SKI* and *PRDM16* and their interaction in experimental rodent systems

is very reassuring. However, before committing to further studies and publication, I will need to undertake additional manual quality control of the genotypes and ideally would formulate additional ways to ensure the soundness of the result. Although a stringent quality control has been applied both to the actual genotypes and the imputations, poor genotype calling can lead to faulty imputations [Morris & al., 2010]. I have already manually checked the genotype calling quality of 11,377 genotypes of the Sanger12 batch, but manual quality control of the other datasets, re-imputation and potential direct genotyping of the associated SNPs should be conducted. To ensure the soundness of the result, dimensionality reduction and GWAS of 3D heart phenotypes from an independent dataset would be the ideal scenario. Unfortunately, high resolution MRI scans are not routine and even the UK Biobank MRI scans are not directly equivalent. Other possibilities include investigating the specific biology behind these loci or the specific molecular biology of the regulatory elements to provide additional evidence for the biological correctness of these associations. We are also planning to investigate these genetic loci and the spatially confined association signal in the left ventricle (figure 7.13) in cohorts of patients suffering from cardiomyopathies (section 2.4).

As a first step towards understanding the spatial association signal, we investigated additional image analysis approaches for phenotype extraction of left ventricular phenotypes. The extraction of ventricular trabeculation phenotypes and their genetic associations are described in the following chapter.

8

GWAS of left ventricular trabeculation

In addition to the unsupervised phenotype selection through dimensionality reduction, the raw cardiac magnetic resonance images also provide the opportunity for a guided phenotype extraction. From a combined clinical and research point of view, phenotypes which are implicated in diseases but which also show strong natural variation are of special interest. Trabeculation phenotypes of the left ventricle fit this description.

8.1. Left ventricular trabeculation

Trabeculation is the formation of small irregular muscle protrusions from the inside of the heart wall and has its origin in early heart development. As described in section 2.3, the chambers of the human heart develop through the looping of the early cardiac tube. During this process, the compartmentalisation of the heart begins and the composition of the cardiac tissue changes, especially in the ventricles. At this stage, the ventricular myocardium can be described as a loose, "spongy" network of myocardial fibres that form sheet-like protrusions (trabeculae) towards the cardiac lumen. The formation of these structures supports the oxygen and nutrient exchange in the heart [Chen & al., 2009] by blood flowing through the intertrabecular spaces [Zambrano & al., 2002]. Later in development, the myocardium starts to become

more compact and thicker and the large protrusions into the heart lumen flatten or disappear [Yousef & al., 2009]. This compaction process progresses from the base of the heart towards the apex and from epicardium to endocardium [Zambrano & al., 2002]

Failure of the myocardial compaction process leads to persistence of ventricular hypertrabeculation. Clinically, the majority of hypertrabeculation phenotypes are observed in the left ventricle and are referred to as left ventricular non-compaction (LVNC) [Zambrano & al., 2002]. It is still unknown if LVNC constitutes a distinct disease or is a shared characteristic of different cardiomyopathies [Captur & al., 2013]. Linkage studies and targeted sequencing of associated regions have revealed a number of genes implicated in familial cases of LVNC [Bleyl & al., 1997; Klaassen & al., 2008; Moric-Janiszewska & Markiewicz-Łoskot, 2008], with a wide range of functions of the encoded proteins. These include cardiac muscle α actin [Monserrat & al., 2007], β-Myosin Heavy Chain [Budde & al., 2007] as well as cytoskeletal-associated proteins like α-dystrobrevin [Ichida & al., 2001] and Cypher/ZASP [Vatta & al., 2003]. Knock-out studies of genes regulating cardiovascular development have contributed to a molecular understanding of clinically relevant LVNC phenotypes [Chen & al., 2009; Mysliwiec & al., 2011]. However, the genetics of sporadic LVNC remain largely unknown [Zambrano & al., 2002].

In addition to LVNC as a clinical phenotype, variation in trabeculation pattern and strength have also been observed in healthy volunteers. Several studies have analysed the range of natural and diseased non-compaction phenotypes with respect to clinical and demographic parameters [Petersen & al., 2005; Captur & al., 2014]. In particular, two independent studies have observed an increase in the ratio of non-compacted to compacted myocardium (NC:C) in individuals of African-American and Hispanic descent compared to Caucasian individuals. The lowest NC:C ratios were observed for individuals of Chinese descent [Kawel & al., 2012; Captur & al., 2015]. The genetics of this natural variation and clinically observed sporadic phenotypes in humans are still poorly understood.

In this chapter, I analyse natural genetic variation driving left ventricular trabeculation phenotypes in healthy volunteers. Trabeculation phenotypes were extracted automatically via fractal analysis from the cardiac magnetic resonance images of the healthy volunteers by my collaborators. Based on these phenotypes, I conducted a GWAS of left ventricular trabeculation.

8.2. Image acquisition and phenotyping

The cohort used for the genetic association study of left ventricular trabeculation consists of the samples with European ancestry that passed the genotyping and imputation quality control described in section 7.1.1. Since there is no ground to suspect confounding of the phenotype processing based on the relatedness of samples as it was the case in the previous chapter (chapter 7), related samples were included in the cohort. For each of the 1,207 samples, the level of trabeculation was measured at six to ten positions in the heart. Trabeculation was quantified via fractal analysis, a technique which allows to measure the complexity of patterns [Eke & al., 2002]. Fractal analysis yields a unit-less measure, the fractal dimension (FD), which quantifies the complexity of the analysed structure. The higher the FD measure, the higher the complexity of the structure i.e. the more trabeculation is observed in the left ventricular wall.

The pipeline for the automatic detection and quantification of trabeculation in the left ventricle was developed by Jiashen Cai and Pawel Tokarczuk. In the following paragraph, I briefly describe the image acquisition and phenotype extraction procedure.

2D cardiac magnetic resonance imaging was conducted at the Hammersmith Hospital, London. The fractal dimensions were derived from standard left ventricular short axis 2D cardiac magnetic resonance images in the plane from base to apex. Each section had a thickness of 8 mm with a 2 mm gap between sections. A more detailed description of the imaging parameters can be found in [de Marvao & al., 2014]. Fractal analysis was automated according to the protocol proposed by Captur and colleagues [2013]. First, the images (figure 8.1, 1) were binarised into blood pool and myocardium (figure 8.1, 2) and the endomyocardial border extracted via edge detection (figure 8.1, 3). The FD was determined by placing grids with known spacing (scale) of increasing size (i.e. increasing number of edges) on the image and subsequent counting of the number of boxes with non-zero pixels, i.e. how many boxes contain at least one pixel of the border (figure 8.1, 4). The slope of the linear regression of the ln-transformed scale versus the ln-transformed counts corresponds to the FD (figure 8.1, 5) [Captur & al., 2013].

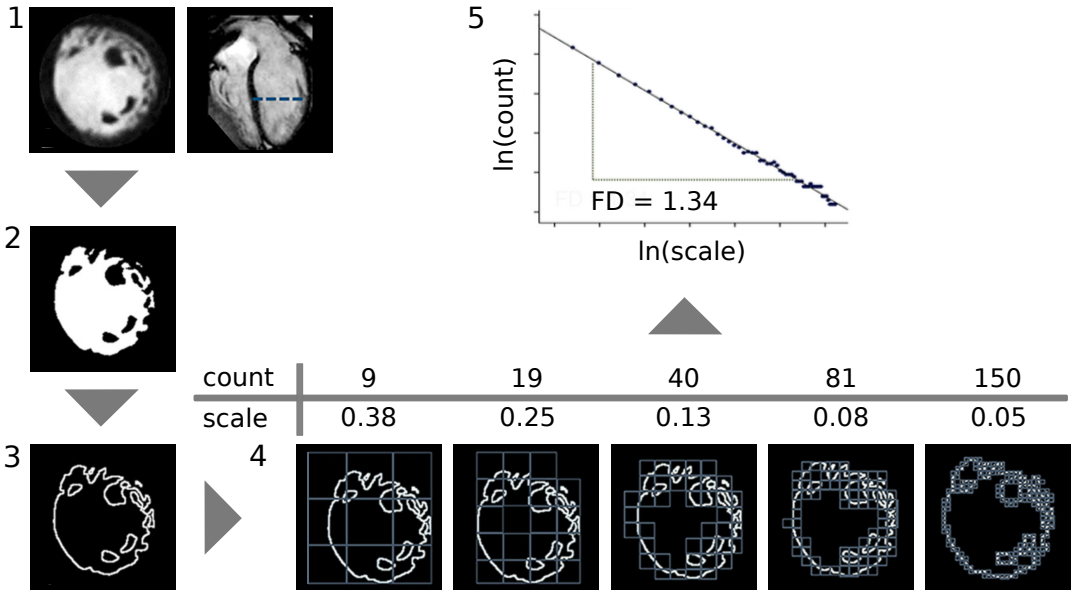

Figure 8.1: FD phenotyping scheme. FD is determined for each of the left ventricular short axis slices derived from standard 2D cardiac magnetic resonance images. 1. An example left ventricular short axis slice is depicted on the left, its location in the heart is indicated by the dashed line of the heart image on the right. 2. The image is binarised into blood pool (white) and other structures (black). 3. The border between the white and the black background is the endocardial wall, which can be extracted via edge detection. 4. A standard box-counting method is applied to the image of the extracted border, where grids of known spacing (scale) are placed on top of the image and boxes containing at least one pixel of endocardial borders are counted. 5. The slope of the regression of the ln-transformed scale versus the ln-transformed count is the FD. Adapted from [Captur & al., 2013].

8.3. The complexity of trabeculation shows a consistent base to apex pattern

For 1,192 out of the 1,207 genotyped samples, FD measurements could successfully be computed at each slice. Their distribution from base to apex is depicted in figure 8.2. Both at the tip of the apex and the end of the basal zone, FD is generally lowest and increases towards the mid-section of the heart. Similar results have been observed by [Kawel & al., 2012] and [Captur & al., 2014]. The latter have shown that most variation between healthy and diseased individuals exists in FD measurements derived from the apical slices of the heart (figure 8.2A) and used the maximal FD value observed in these slices as their final phenotype. I followed the strategy of dividing the measurements into apical and basal (figure 8.2B) and used the maximum FD observed in each group as final phenotypes. For individuals with uneven numbers of slices, the center slice was not considered for the computation of the maximum values.

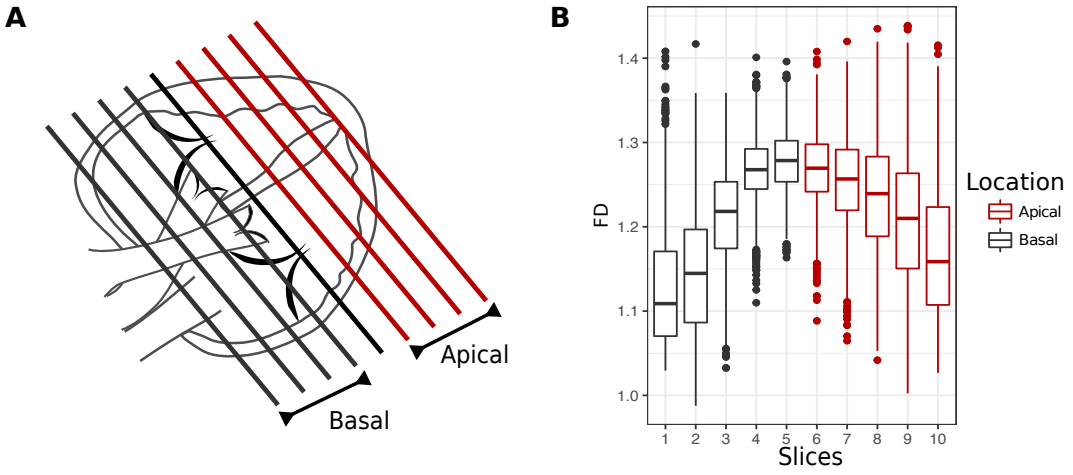

Figure 8.2: FD measurements from base to apex. A. Location of the 2D cardiac magnetic resonance image slices and their classification into apical and basal. B. FD measurements for all samples were interpolated via a cubic spline function to the maximum number of 10 slices for easier visualisation. Subsequent analyses were done based on the original, non-interpolated FD measurements.

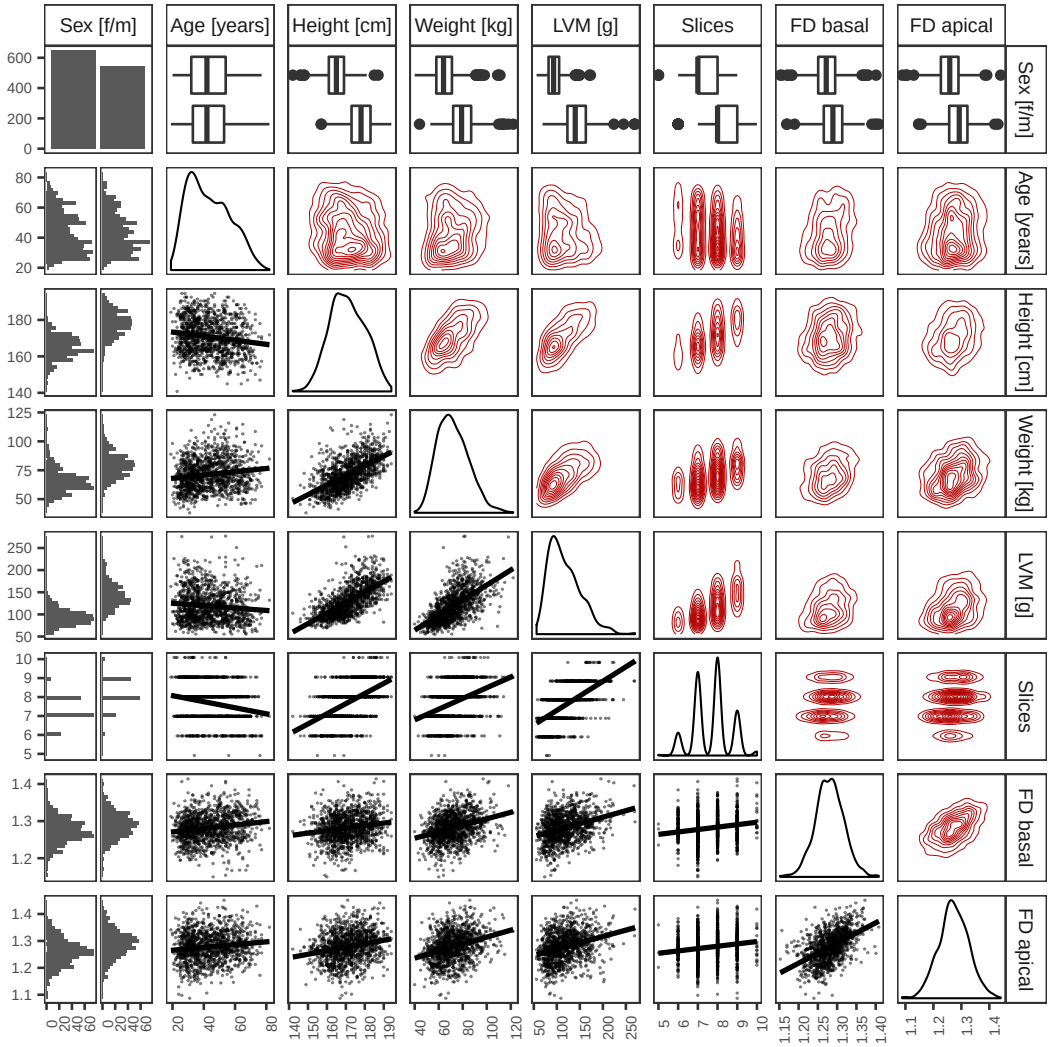

Figure 8.3: Relationship between FD measures and covariates. Continuous variables: the univariate-distribution of each variable is depicted on the diagonal. The upper triangular matrix shows the bi-variate distribution while the lower triangular matrix shows the regression line of their linear fit. Categorical variables (sex): Distribution (first row) and counts (first column) are depicted.

8.4. Relationship between trabeculation phenotypes and covariates

I analysed the distribution of the 2 FD measurements FD_{max}^{basal} and FD_{max}^{apical} in relation to biological and cardiac covariates (figure 8.3). Both FD measurements show correlation with age, weight and left ventricular mass. FD_{max}^{basal} is additionally associated with height, while FD_{max}^{apical} also shows correlation with sex (table 8.1). The association of LVM and FD_{max}^{apical} confirms the findings of the study by Captur and colleagues [Captur & al., 2014], who found increased FD measures for individuals with increased LVM. However, the causality of the relationship has not been determined yet. All associated covariates except for LVM, as the causal relationship to FD measurements is unclear, were used as covariates in the GWAS.

Table 8.1: Association of FD_{max}^{basal} and FD_{max}^{apical} with covariates. Association was determined based on a simple linear model for each FD measurement with all covariates as explanatory variables without interaction effects.

	FD_{max}^{basal}	FD_{max}^{apical}
Sex	5.47×10^{-1}	4.96×10^{-3}
Age	3.04×10^{-8}	2.87×10^{-4}
Height	4.33×10^{-2}	4.43×10^{-1}
Weight	1.25×10^{-4}	4.55×10^{-5}
LVM	1.21×10^{-12}	2.62×10^{-3}
Slices	8.02×10^{-1}	3.89×10^{-1}

8.5. Left ventricular trabeculation is associated with two genomic loci

The extraction of FD measurements from the 2D cardiac magnetic resonance images yields quantitative phenotypes capturing the complexity of trabeculation in the left ventricle. I used the two summary measures FD_{max}^{basal} and FD_{max}^{apical} described above as the response variables in a mtGWAS with the genetic marker and sex, age, height and weight as covariates. Since the dataset contained related individuals, I extended to model used in section 7.3 to a LMM by including an additional random genetic effect based on the RRM of the samples. The RRM was estimated from the samples' genotypes as described in section 1.7.6. The manhattan and qq-plots for the joint analysis of FD_{max}^{basal} and FD_{max}^{apical} are depicted in figure 8.4 and figure 8.5, showing two loci that reach genome-wide significance. As a comparison, stGWAS of FD_{max}^{basal} and

FD_{max}^{apical} only discovered the association on chromosome 2 (with response variable FD_{max}^{apical}; figure B.9), demonstrating the power of the multi-trait approach. A sum-

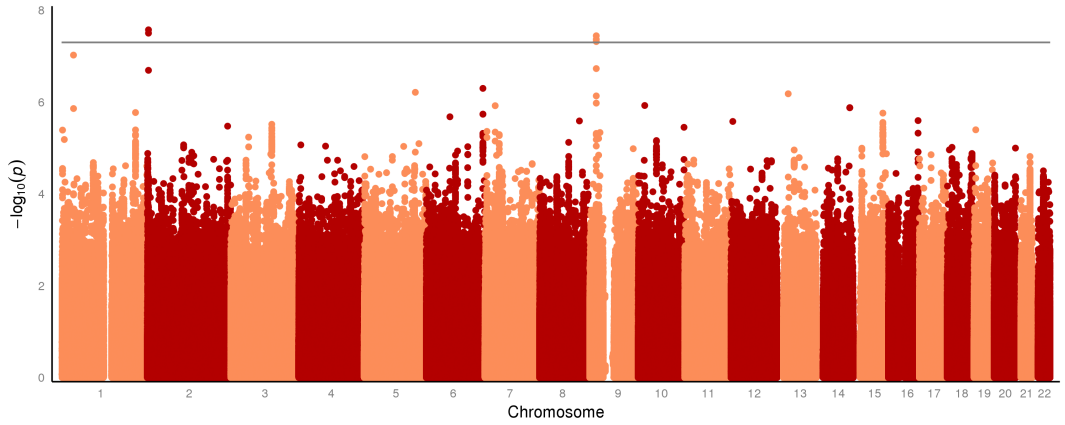

Figure 8.4: Manhattan plot of multi-trait GWAS on left ventricular trabeculation. The maximal apical and basal FD were modelled jointly in an any effect mtGWAS. The p-values of all genome-wide SNPs are depicted. The horizontal grey line is drawn at the level of genome-wide significance: $p = 5 \times 10^{-8}$.

mary of the two loci that reach genome-wide significance is shown in table 8.2 and figure 8.6. The locus on chromosome 2 lies within an intron of a long intergenic noncoding RNAs (lincRNA) of unknown function (figure 8.6, upper panel). The second associated locus is positioned in intron 24 of the *ADAMTSL1* gene (figure 8.6, lower panel). *ADAMTSL1* is also known as *Punctin* and two of its intronic and intergenic variants (rs7869627: intron 17; rs1411242: intergenic between *SH3GL2* and *ADAMTSL1*) have been found associated with blood pressure phenotypes [Sabatti & al., 2009]. rs7855681 is in weak LD with rs7869627 ($r^2 = 0.119$).

Table 8.2: SNPs with strongest association in left ventricular trabeculation GWAS. For each locus, the p-values for SNPs in LD with an $r^2 > 0.8$ in a 50kb window were compared and only the SNP with smallest p-value per locus listed below. M allele: major allele, m allele: minor allele, MAF: minor allele frequency.

SNP	Chr	Position	P-value	M/m allele	MAF
rs7603133	2	3,103,708	3.23×10^{-8}	A/G	0.09
rs7855681	9	18,855,498	3.46×10^{-8}	A/C	0.32

Punctin is a secreted glycoprotein that can be detected in contacts between cells and components of the extra-cellular matrix, but that has not been observed in cell-cell contacts [Hirohata & al., 2002]. It is part of the ADAMTS-like protein family which lack the proteolytic activity of their name-lending metalloprotease pro-

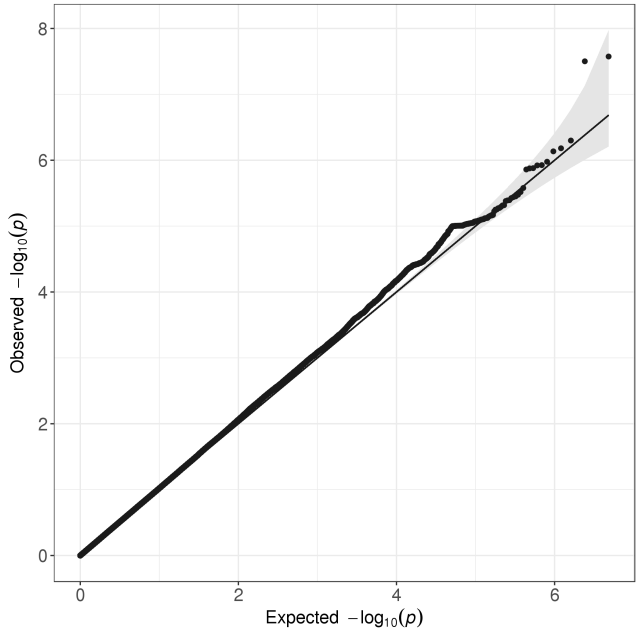

Figure 8.5: Quantile-quantile plot of multi-trait GWAS on left ventricular trabeculation. The observed genome-wide p-values of the multi-trait FD GWAS are plotted against equally spaced values in $[0, 1]$ of the same sample size (expected p-values). The diagonal line starts at the origin and has slope one.

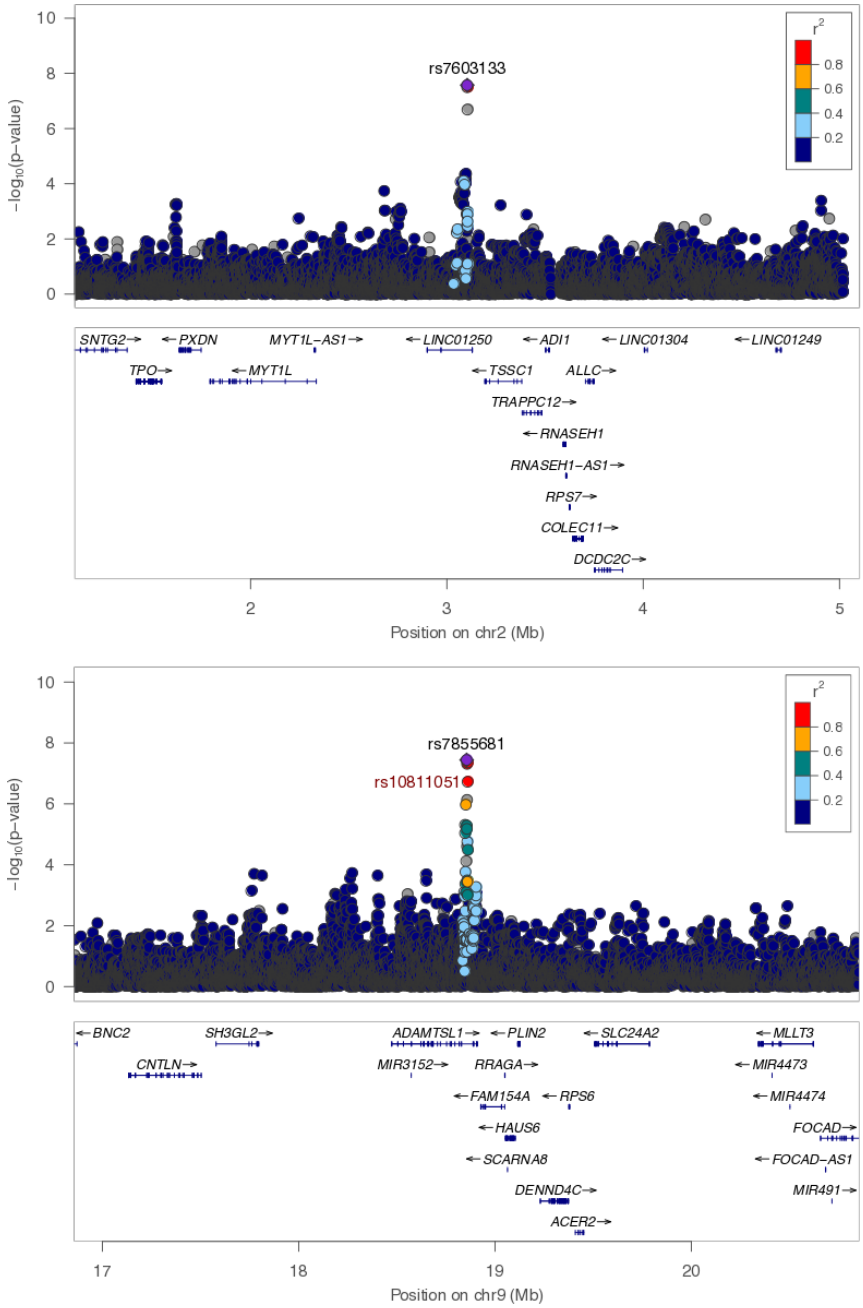

Figure 8.6: Genomic context of loci associated loci with left ventricular trabeculation. The p-values and genomic location of the two loci reaching genome-wide significance are shown in relation to the p-values of surrounding genotypic markers. Markers are coloured by the level of LD they share with the SNP of interest. For both loci, all SNPs that are associated were imputed. For the locus on chromosome 9 (lower panel), an additional SNP which was directly genotypes but has not passed the genome-wide significant level has been marked in red. Generated with LocusZoom [Pruim & al., 2010].

tein family. While other proteins of the ADAMTS-like family have been shown to be associated with connective tissue disorders and affecting the formation of the extra-cellular matrix [Ahram & al., 2009; Hubmacher & Apte, 2015], the function of punctin remains unknown. However, progress has been made in understanding the regulation of its secretion through post-translational modification of its tryptophane 42 residue [Wang & al., 2009]. A recently published study shows a strong systemic phenotype for the mutation of this tryptophane residue, inhibiting the secretion of the protein. However, no further advances in understanding the mechanisms or finding binding partners of ADAMTSL1 could be made [Hendee & al., 2017].

The locus on chromosome 1 (SNP: rs113719231) discovered in section 7.3 is located near the *PRDM16* gene which has been associated with LVNC [Arndt & al., 2013]. A linear model with the rs113719231 genotypes, sex, age, height and weight as explanatory variables and $FD_{max}^{basal}/FD_{max}^{apical}$ as response variables did not show any association, even without the burden of the genome-wide significance threshold ($p = 0.78/p = 0.77$).

The clinical phenotype of left ventricular non-compaction has been found associated with a number of other cardiac and cardiovascular phenotypes such as conduction abnormalities [Yousef & al., 2009], arrhythmias [Ritter & al., 1997; Oechslin & al., 2000; Yousef & al., 2009], coronary artery disease [Ritter & al., 1997; Junga & al., 1999; Jenni & al., 2002; Soler & al., 2002] and myocardial infarction [Swinkels & al., 2007; Toufan & al., 2012; Güvenç & al., 2012]. In addition, a study on population variation of left ventricular trabeculation found associations between the increase in left ventricular trabeculation and prevalence of hypertension, left ventricular mass and wall thickness [Captur & al., 2015]. For the majority of these phenotypes, original GWAS and meta-analysis of GWAS have been conducted including atrial fibrillation [Gudbjartsson & al., 2007; Christophersen & al., 2017], atrioventricular conduction [Denny & al., 2010], coronary heart disease [Schunkert & al., 2011; Lee & al., 2013; Nikpay & al., 2015], myocardial infarction [Kathiresan & al., 2009; Hirokawa & al., 2015; Nikpay & al., 2015; Dehghan & al., 2016] and blood pressure phenotypes [Ehret & al., 2011; Wain & al., 2011]. For studies where the summary statistics of the genome-wide associations were made publicly available (blood pressure phenotypes [Ehret & al., 2011; Wain & al., 2011], coronary artery disease [Schunkert & al., 2011] and myocardial infarction [Nikpay & al., 2015]), I collected the effect size estimate (continuous traits) and odds ratios (case-control setting) for the associated loci on chromosome 2 and 9. The SNP with the highest association on chromosome 9 (rs7855681) was contained in all available studies. For the locus on chromosome 2,

the SNP with the highest association was not contained in any of the studies, however rs6758505 which is in strong LD with the discovered SNP in Europeans ($r^2 = 1$) was found in one of the studies. Figure 8.7 depicts the effect size estimates and odds ratios for both SNPs estimated for different blood pressure measurements, coronary artery disease and myocardial infarction. For all phenotypes, the confidence intervals of effect size/odds ratio estimates contain zero and one, respectively and thus show no effect of the SNPs on these phenotypes. A database search of the GWAS catalogue [MacArthur & al., 2017] for associated SNPs and SNPs in LD (based on entries in the GWAS catalogue, 0.7.08.2018) and the Global Biobank engine [GBE, 2017] did not yield associations with any other phenotype.

8.6. Summary

In this chapter, I used phenotypes derived from a guided feature extraction method to map naturally occurring genetic variation in healthy individuals to a clinically relevant phenotype. The association of the FD phenotypes as a quantification of left ventricular trabeculation detected two loci that are linked on a genome-wide significant level. Both loci lie in intronic regions and have no direct protein-coding consequences. Loci in proximity to the association detected within the *ADAMTSL1* gene have been implicated in cardiac phenotypes such a blood pressure. However, the absence of any effect for this locus in well-powered published GWAS of blood pressure phenotypes points towards a blood pressure-independent effect on left ventricular trabeculation.

For quantitative, continuous phenotypes and additive genotype effects, understanding naturally occurring variation can give insights into the genetic architecture of the traits and might help to understand more extreme disease phenotypes. In order to extend this study and confirm results in a larger cohort, we applied for access to the UK Biobank a "large, population-based prospective study, established to allow detailed investigations of the genetic and non-genetic determinants of the diseases of middle and old age" [Sudlow & al., 2015]. Within this project, 500,000 individuals have been genotyped and phenotyped for wide array of traits, including 2D cardiac magnetic resonance imaging scans on an expected 100,000 individuals. In contrast to the 3D heart phenotypes investigated in chapter 7, the FD phenotypes can be automatically extracted from these images. Upon access to the data, phenotype extraction and a mtGWAS with the same model and parameters as described in this chapter will be conducted.

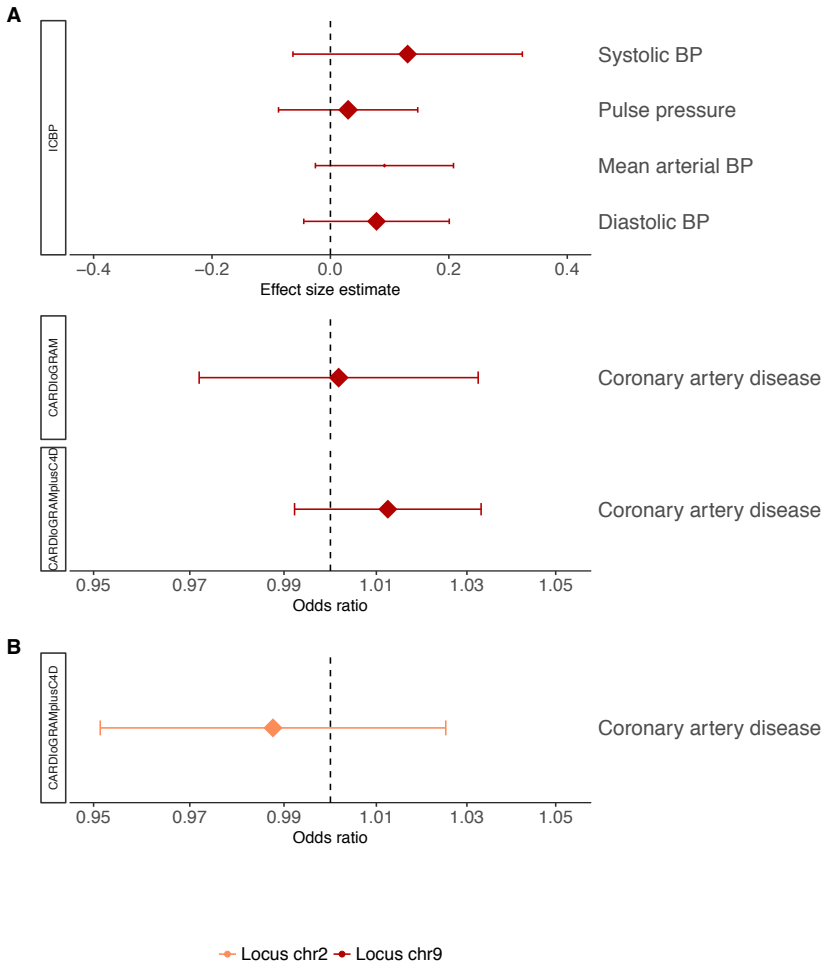

Figure 8.7: Effect estimates of associated FD SNPs with other cardiovascular phenotypes. Effect size estimates and odds rations for the SNPs associated with FD were derived from previous published studies on blood pressure (BP) phenotypes and risk for coronary artery diseases and myocardial infarction. The diamond indicates the effect estimates, the error bars their confidence interval. The size of the diamond represents the sample size of the study and is normalised to the largest study size (pulse pressure: N =71,663). All studies were conducted as meta-analyses in the scope of large consortia (faceting labels). The dashed vertical line indicates the value of no effect.

In addition to this replication study, investigating the genetic variation driving the healthy phenotype differences in individuals of different ethnicities [Kawel & al., 2012; Captur & al., 2014] will be of great interest. While the cohort in this study only contained a minority of non-European samples, a more diverse cohort structure might be observed in the UK Biobank cohort, enabling this analysis.

9
Concluding remarks

Initially, GWAS used seemingly simple case-control designs to map genotypes to a variety of disease phenotypes. In subsequent years, existing models were discovered for their application in GWAS [Korte & al., 2012] and novel techniques developed, enabling the analysis in cohorts with complex structure [Yu & al., 2006; Kang & al., 2010], the effect estimation of sets of genotypes [Wu & al., 2010; Casale & al., 2015] or gene-environment interaction in the context of GWAS [Casale & al., 2017]. While sophisticated methods for the analysis of multiple traits existed [Korte & al., 2012; Zhou & Stephens, 2012; Casale & al., 2015], they were mainly limited to moderate trait numbers due to their computational complexity. LiMMBo (chapter 4) fills this gap by enabling the joint analysis of hundreds of phenotypes. Its performance on simulated data demonstrated its power even when only a moderate number of observed phenotypes is governed by the same genetic factors. The application to a dataset for yeast growth traits did not only show its usefulness on real data, but also demonstrated its value for investigating and generating biologically relevant hypotheses such as pleiotropy of traits and complex trait structures.

I provide the phenotype simulation framework (chapter 3) and LiMMBo as open-source software packages: *PhenotypeSimulator* (chapter 3) is accessible via the Comprehensive R Archive Network [Meyer, 2017] and LiMMBo is implemented in a python module which can be used in combination with the publicly available LIMIX

suit for flexible linear mixed model designs [Lippert & al., 2014].

For very high-dimensional datasets, one is often interested in applying *a priori* dimensionality reduction method to the data to extract information relevant for the biological question of interest. In the biological literature, PCA is standardly employed for this task [Avery & al., 2011; Liu & al., 2012; Zhang & al., 2012]. However, there exist a growing number of dimensionality reduction techniques based on different statistical methods and assumptions about the hidden data structures. Twelve of these publicly available dimensionality reduction techniques were explored for their ability to find a robust representation of the input data (chapter 6). I used *PhenotypeSimulator* to generate datasets of different sizes and underlying structures and introduced stability as a new measure to determine the dimensions of a robust low-dimensional representation. I was able to show that dimensionality reduction techniques are valuable for genotype-phenotype mapping studies of very high-dimensional datasets as the simulated genetic effects could be discovered in genetic association studies with the stable low-dimensional representations as phenotypes.

I directly applied these insights to a clinically interesting dataset of spatially-resolved three-dimensional human heart phenotypes. Based on the hypothesis that there are genetic factors that influence the heart morphology in a spatially-confined manner, I extracted low-dimensional representations of the left-ventricular wall thickness measurements and used these in a genome-wide association study. Associated SNPs did not only show a regional-confined effect but have also been implicated in cardiac phenotypes in model organisms. While further studies are needed to confirm these findings, the results demonstrate the power of this approach to investigate biologically and clinically relevant questions.

In the feature extraction approach used for this GWAS, I combined the stable low-dimensional representations from a variety of different dimensionality reduction approaches, with the underlying hypotheses that different methods capture different aspects of the morphology and a combination of the methods will yield a comprehensive representation. Alternatively, models which are more tailored to the specific structure of the dataset could be employed. The spatially-resolved heart wall thickness measurements in this study are part of a larger class of data structures, where measurements on a two-dimensional surface are embedded in a three-dimensional space. Similar data has been observed for 3D structural MRI or 4D functional MRI studies in the brain [Van Essen & al., 2012; Glasser & al., 2013]. Novel feature extraction methods for neuroscience data can take *a priori* knowledge about spatial correlation of the input data into account. For instance, functionalPCA com-

bines approaches from PCA and DRR and incorporates additional sparsity priors into the model, which act on the underlying three-dimensional model of the data [Lila & al., 2016]. Similar extensions could be envisaged for the Bayesian factor analysis model PEER [Stegle & al., 2012], where the spatial coordinates could be build into the model as priors.

In addition to the wall thickness measurements, the phenotyping approach developed by my collaborators also provides spatially-resolved measurement for heart wall curvature and fractional wall thickness i.e. wall thickness changes between diastole and systole. In molecular phenotyping of different tissues or conditions the simple, albeit high-dimensional genotype-phenotype mapping is extended from the two dimensional "sample by phenotype" space into the higher-dimensional "sample by phenotype by condition/tissue/etc." space. Novel methods have been developed for the task of jointly analysing such datasets [Hore & al., 2016]. These approaches could be applied to extend this study and find stable phenotype components representing a more comprehensive cardiac phenotype based on wall thickness, curvature or fractional wall thickening.

In a second genetic association study with heart morphology, I discovered SNP-associations with a trabeculation phenotype from a supervised feature extraction approach on the raw MRI data. The implicated SNPs are located in proximity of a gene important in the developmental process of this trabeculation and follow-up studies are underway to confirm these results.

Improved diagnosis and interventional strategies in the past two decades have contributed to the general improvements in fighting cardiovascular diseases. While these improvements were mainly based on large-scale clinical trials, there is a call now for more personalised approaches to further improve the management of cardiovascular diseases [Meder & al., 2016]. The proposed strategies ask for a stronger interaction between clinical, molecular and statistical expertise to enhance the characterisation of these diseases. Studies such as the GWAS on cardiac morphology show the feasibility of this proposal, with a strong collaboration between clinical and bioinformatics expertise to investigate the genetic basis of cardiac phenotypes. Follow up studies and further exploration of the data as outlined above can contribute to further characterise the genetics of cardiac structure and function.

Appendix

A

Supplementary tables

A.1. Additional information chapter 2

Table A.1: GWAS catalogue trait descriptions relating to cardiovascular diseases. Out of the 4,148 studies in the GWAS catalogue (accessed 11.08.2017), 159 contain phenotype description related to cardiovascular diseases. For a summary of the studies conducted, they were broadly summarised into eight groups (Summary name). A graphical overview is shown in figure 2.3.

Summary name	GWAS catalogue trait
Congenital heart disease	Congenital heart disease
	Congenital left-sided heart lesions (maternal effect)
	Congenital left-sided heart lesions
	Conotruncal heart defects
Coronary heart disease	Coronary heart disease
	Myocardial infarction
	Myocardial infarction (early onset)
	Coronary artery disease
	Coronary heart disease event reduction in response to statin therapy (interaction)
	Coronary restenosis
	Myocardial infarction in coronary artery disease
Blood pressure	Blood pressure
	Hypertension
	Systolic blood pressure
	Diastolic blood pressure
	Hypertension (young onset)
	Systolic blood pressure in sickle cell anemia
	Blood pressure (smoking interaction)
	Blood pressure measurement (cold pressor test)
	Blood pressure measurement (high sodium and potassium intervention)
	Blood pressure measurement (low sodium intervention)
	Blood pressure measurement (high sodium intervention)
	Systolic blood pressure (alcohol consumption interaction)
	Diastolic blood pressure (alcohol consumption interaction)
	Mean arterial pressure (alcohol consumption interaction)
	Pulse pressure (alcohol consumption interaction)
	Pulse pressure in young-onset hypertension
	Blood pressure (anthropometric measures interaction)
	Blood pressure (age interaction)

Table A.1: continued

	Ejection fraction in Tripanosoma cruzi seropositivity
Electrocardiographic traits	Atrial fibrillation
	Echocardiographic traits
	Atrial fibrillation/atrial flutter
	QT interval
	Electrocardiographic conduction measures
	Atrioventricular conduction
	QRS duration
	Cardiac repolarization
	QT interval (interaction)
	P wave duration
	PR segment
	PR interval in Tripanosoma cruzi seropositivity
	QT interval in Tripanosoma cruzi seropositivity
	QRS duration in Tripanosoma cruzi seropositivity
	Heart rate variability traits
	PR interval
	Resting heart rate
	RR interval (heart rate)
Morphological traits	Left ventricular mass
	Cardiac structure and function
	Cardiac muscle measurement
	Cardiac hypertrophy
	Dilated cardiomyopathy
	Chagas cardiomyopathy in Tripanosoma cruzi seropositivity
Heart failure	Heart failure
	Sudden cardiac arrest
	Mortality in heart failure
Others	Cardiac Troponin-T levels
	Cardiovascular disease risk factors

A.2. Additional results chapter 7

Table A.2: Number of SNPs after imputation, imputation QC and filtering for deviation from HWE and low MAF. Every batch was imputed independently (columns "SNPs after imputation"). SNPs that had an IMPUTE2 "info" metric of > 0.4 in all of the batches were combined and subsequently filtered for SNPs deviating from Hardy-Weinberg equilibrium ($p < 0.001$) and with low MAF (< 0.008), corresponding to a minor allele count of less than 20.

Chr	SNPs after Imputation			INFO > 0.4	HWE and MAF
	Sanger12	Duke-NUS12	Duke-NUS3		
1	3,196,692	3,197,145	3,196,563	1,251,157	719,882
2	3,515,670	3,515,861	3,515,602	1,360,182	780,152
3	2,941,265	2,941,468	2,941,223	1,156,243	665,038
4	2,900,679	2,900,786	2,900,634	1,154,742	684,602
5	2,688,219	2,688,348	2,688,174	1,049,671	606,951
6	2,581,500	2,581,851	2,581,410	1,058,844	635,257
7	2,359,370	2,359,598	2,359,319	932,726	551,744
8	2,323,181	2,323,290	2,323,144	890,407	514,803
9	1,752,242	1,752,363	1,752,199	698,510	398,777
10	2,003,743	2,003,881	2,003,694	812,616	474,686
11	2,013,331	2,013,535	2,013,273	794,587	481,479
12	1,947,915	1,948,107	1,947,865	767,854	452,193
13	1,458,325	1,458,401	1,458,308	590,863	348,525
14	1,333,919	1,333,973	1,333,901	524,391	309,825
15	1,194,294	1,194,406	1,194,264	458,617	266,813
16	1,289,127	1,289,335	1,289,074	497,688	286,620
17	1,118,587	1,118,772	1,118,528	434,724	252,227
18	1,153,963	1,154,034	1,153,942	457,454	268,986
19	877,689	877,866	877,645	361,419	222,264
20	912,602	912,721	912,574	357,156	210,128
21	546,390	546,414	546,381	216,911	131,079
22	531,437	531,528	531,416	215,547	129,771
genome	42,989,377	42,993,178	42,988,308	16,042,309	9,391,802

B

Supplementary Figures

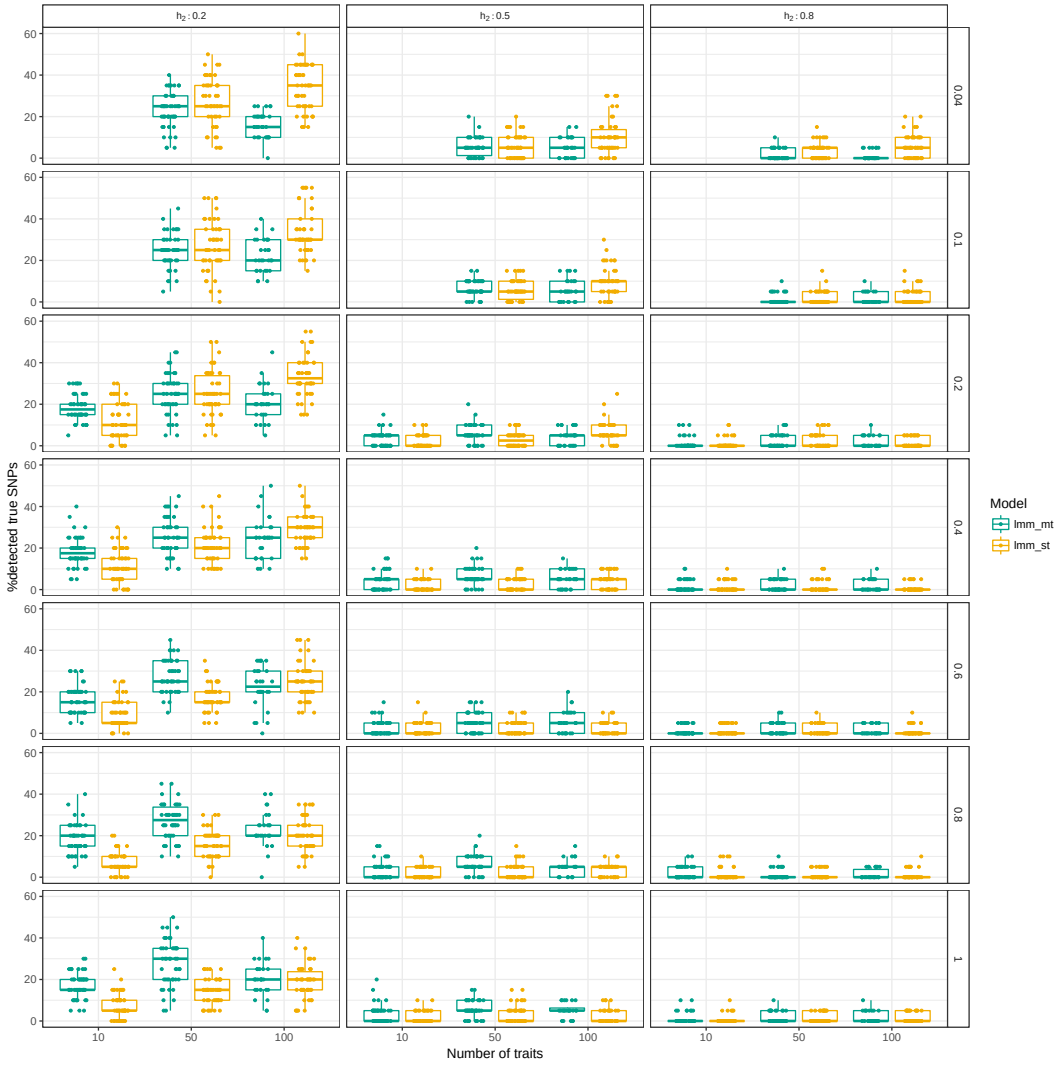

Figure B.1: All parameter combinations of power comparison for multivariate and univariate LMMs of high-dimensional phenotypes. Each panel shows the influence of two simulation parameters on the power to detect the causal SNPs.

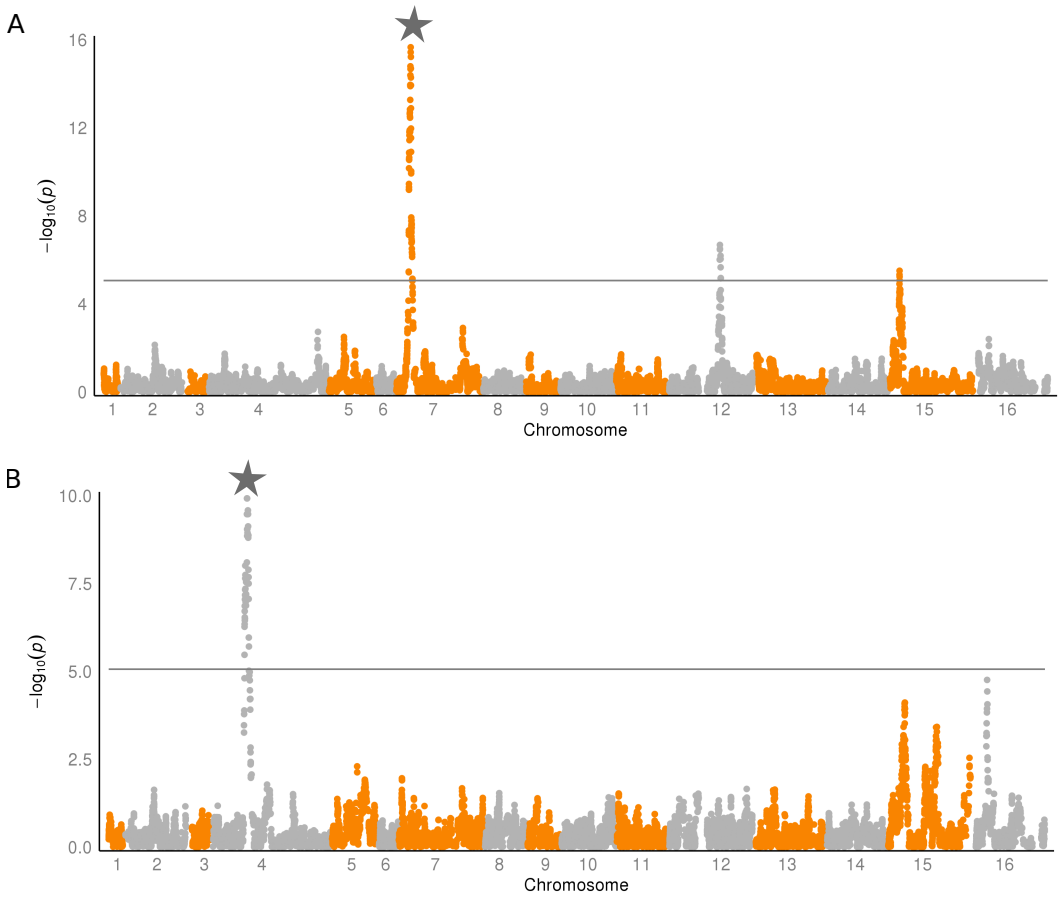

Figure B.2: Manhattan plot of traits with strong single-trait associations. Single-trait GWAS of A. magnesium sulfate and B. hydroquinone. The loci marked with a grey star are only found for these two traits and cannot be detected in the mtGWAS (figure 5.6), pointing to purely single-trait association that is burdened by the multi-trait testing based on 41 degrees of freedom. The p-values were adjusted for multiple testing by the effective number of tests ($M_{\text{eff}} = 33$). The significance line is drawn at the empirical $\text{FDR}_{\text{stGWAS}} = 8.6 \times 10^{-6}$.

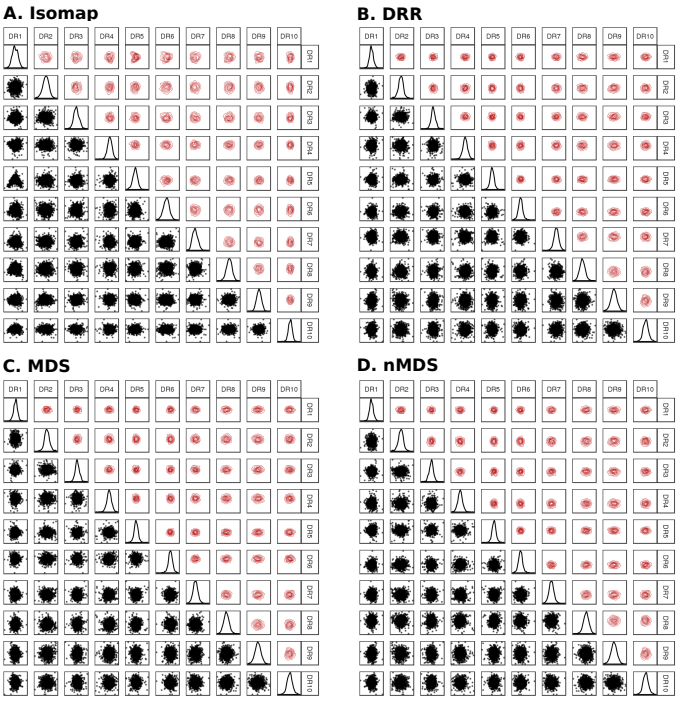

Figure B.3: Additional scatterplots for visual assessment of low-dimensional components derived from left-ventricular wall thickness. Pairwise scatter plots of the components (lower triangle) and density plots (upper triangle) are depicted. The diagonal of each plot shows the distribution of the respective component. Row and column labels specify the rank of the component out of the 100 low-dimensional components. Before plotting, each component was mean-centred and divided by its standard deviation in order to have comparable axis dimensions. Given the normalised scale of the data, and the purpose of qualitative comparison, axis ticks were omitted for a cleaner visualisation.

B.4. Additional results chapter 7

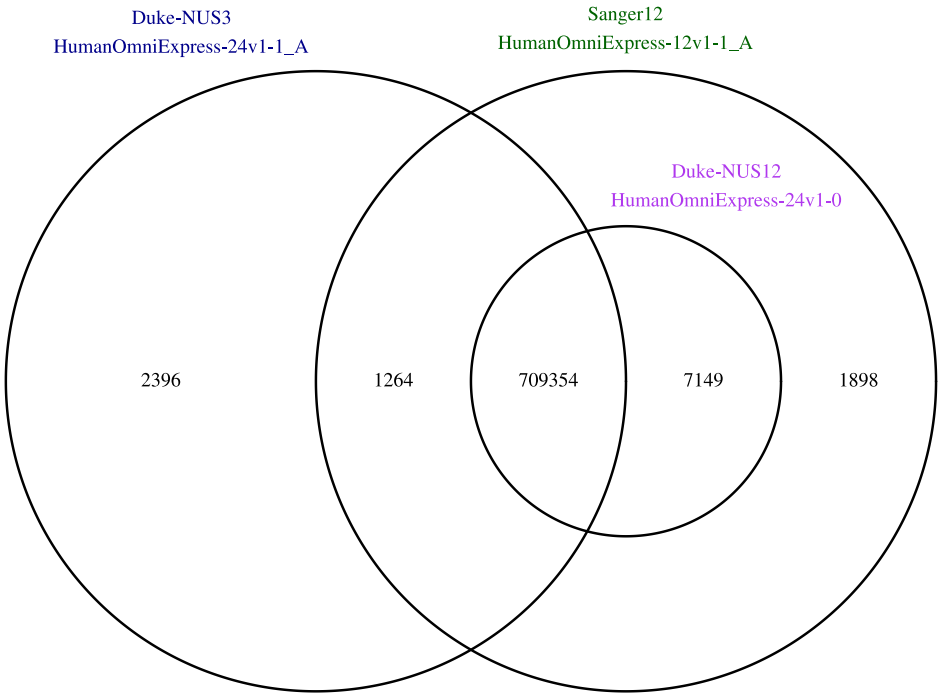

Figure B.4: Number of DNA probes on the different genotyping chips and their overlap. For the genotyping of the individuals in the Digital Heart project three different Illumina HumanOmniExpress genotyping chips were used (24v1-1_A, 12v1-1_A, 24v1-0), differing in the number of probes on the chip (numbers inside Venn diagram) and the number of samples that can be genotyped (12 and 24; indicated in name of chip).

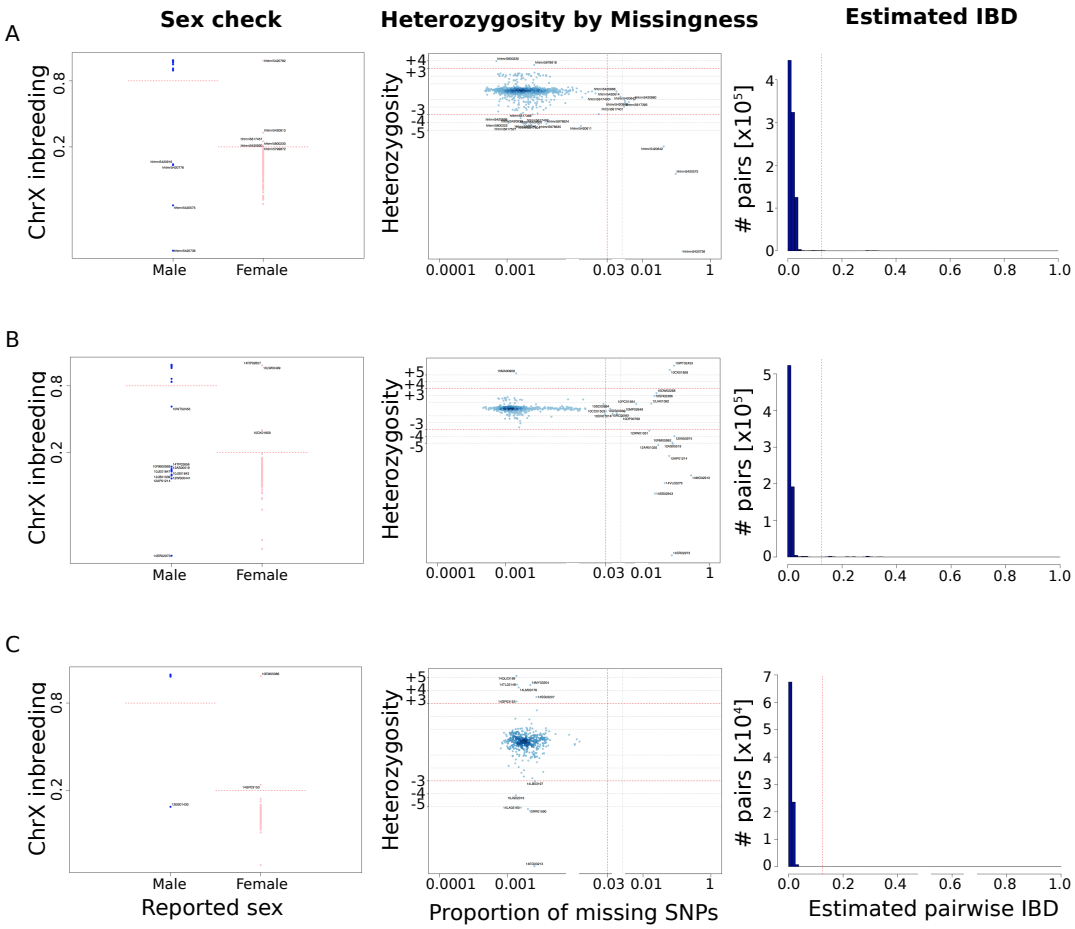

Figure B.5: Genotyping quality control per sample. A. Sanger12. B. Duke-NUS12. C. Duke-NUS3. Supplementary plots for genotyping QC described in section 7.1.1.

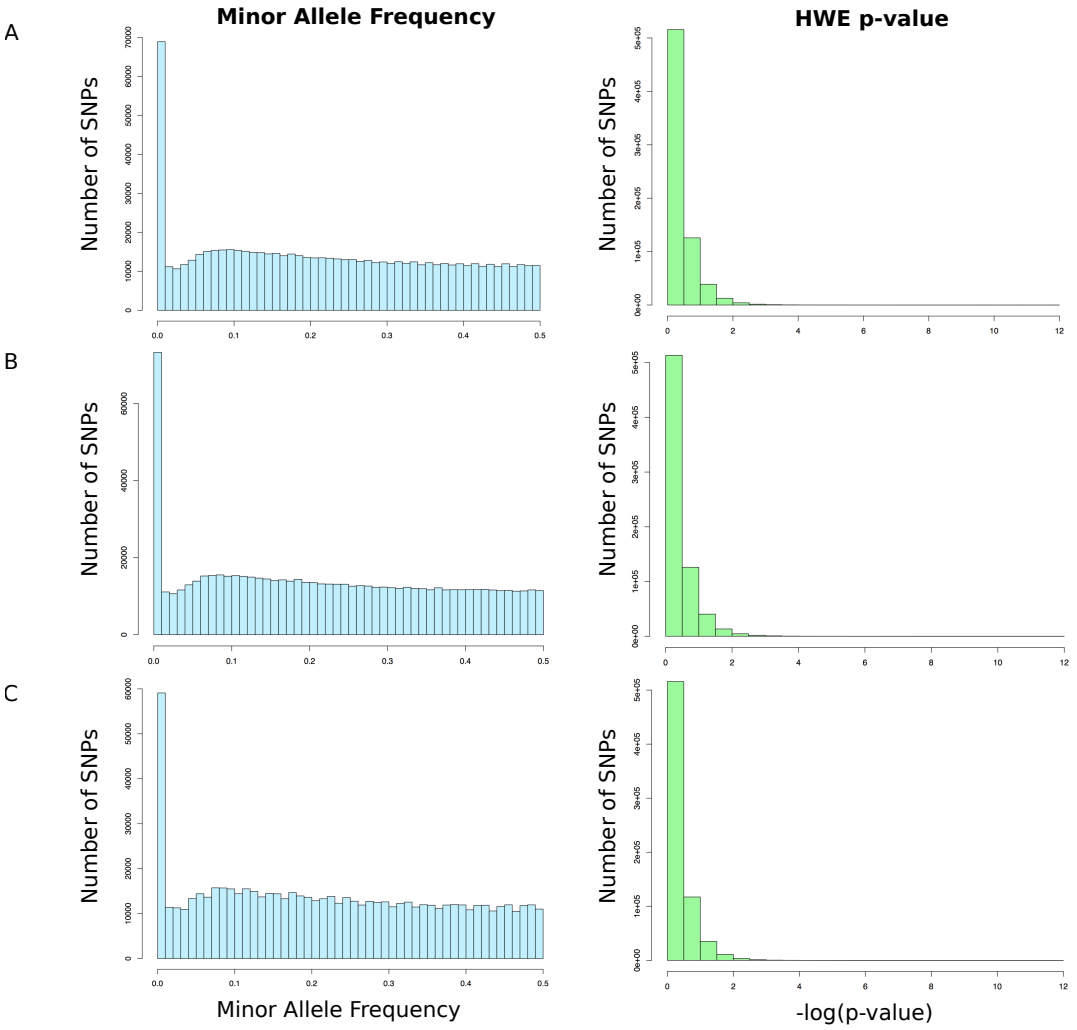

Figure B.6: Genotyping quality control per SNP. A. Sanger12. B. Duke-NUS12. C. Duke-NUS3. Supplementary plots for genotyping QC described in section 7.1.1.

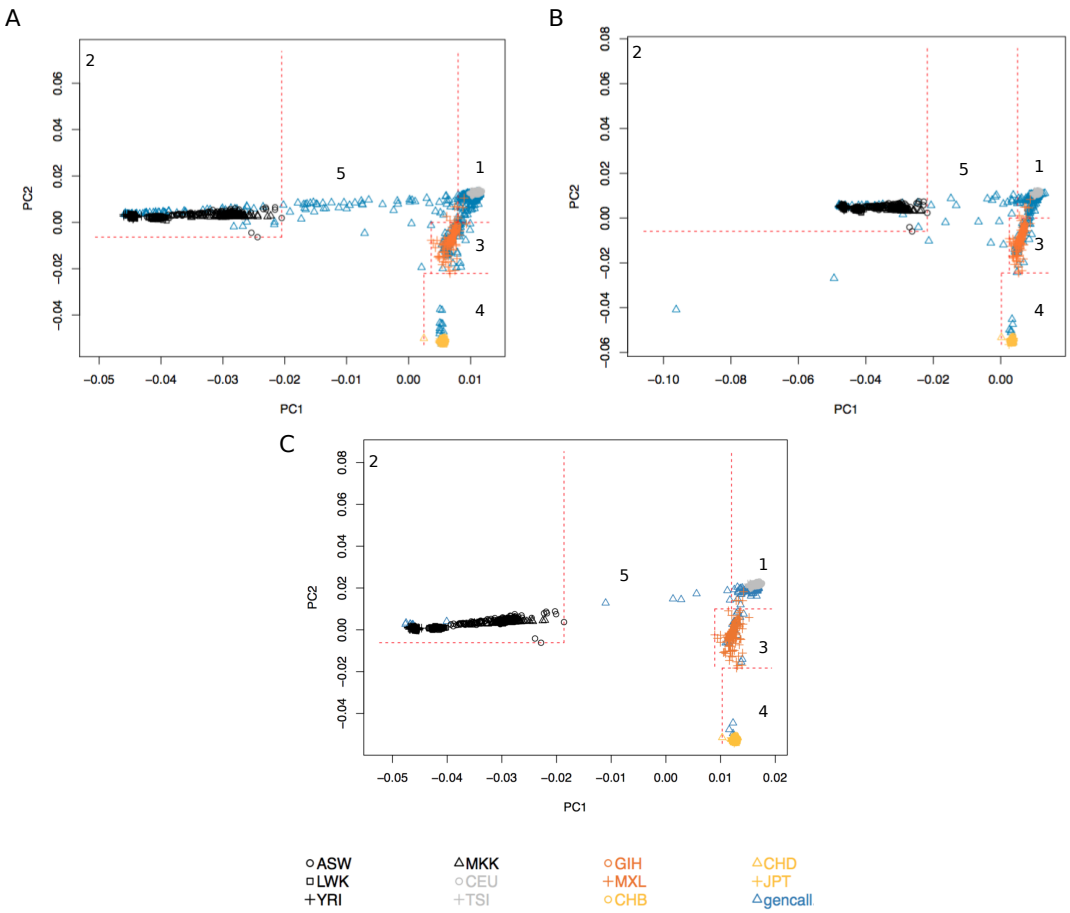

Figure B.7: Ethnicity of samples within the Digital Heart project. A. Sanger12. B. Duke-NUS12. C. Duke-NUS3. PCA was conducted on the SNP genotypes of the samples within the Digital Heart project (gencall) and genotypes of four greater ethnicities of the HapMap project (black: African, orange:Mexican/Native American, grey: European, yellow: Asian) [The International HapMap Consortium, 2005; The International HapMap Consortium, 2007]. The clustering of the samples based on the first and second PCs are depicted. Red dotted lines indicate borders considered to separate ancestries: 1. European, 2: African, 3: Mexican/Native American, 4. Asian, 5: Mixed ancestry. Gencall samples within the first group were used in chapters 7 and 8. A description of the analysis is described in section 7.1.1.

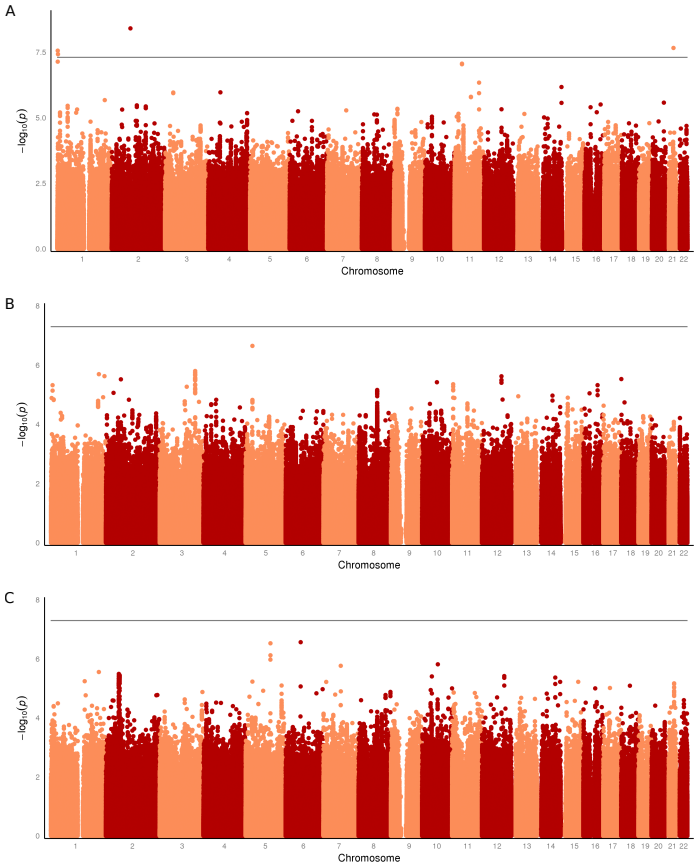

Figure B.8: Manhattan plots for GWAS on stable components from a single dimensionality reduction method. The five stable components derived from Laplacian Eigenmaps (A), four from Isomap (B) and ten from PCA (C) were used as the response variables in three independent any effect mtGWAS. Their p-values were adjusted for the effective number of test conducted, estimated via equation (5.4) based on the correlation across their components (figure 7.7): $M_{eff} = 2.04$. The horizontal grey line is drawn at the level of genome-wide significance: $p = 5 \times 10^{-8}$. Only the locus on chromosome 1 which was detected in the combined analyses (figure 7.8) could also be detected via components from Laplacian Eigenmaps alone.

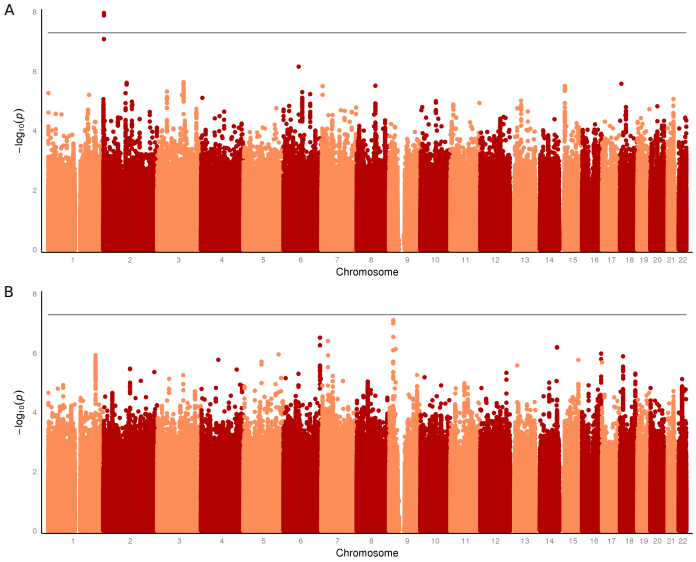

Figure B.9: Manhattan plot of two single-trait GWAS on left ventricular trabeculation The maximal apical (A) and basal FD (B) were used as the response variable in a stGWAS. Their p-values were adjusted for the effective number of test conducted, estimated via equation (5.4) based on their correlation: $M_{eff} = 1.86$. The p-values of all genome-wide SNPs are depicted. The horizontal grey line is drawn at the level of genome-wide significance: $p = 5 \times 10^{-8}$.

C

Derivations

The following section describes the derivation of the simulation scheme for the infinitesimal genetic effects in section 3.2. A suitable model for simulating the infinitesimal genetic effect $\mathbf{G} \in \mathcal{R}^{N,\,P}$ with known $N \times N$ sample (row) covariance is a matrix-normally distributed random variable, defined by its mean $\mathbf{M} \in \mathcal{R}^{N,\,P}$, its row covariance $\mathbf{D} \in \mathcal{R}^{N,\,N}$ and its column covariance $\mathbf{C} \in \mathcal{R}^{P,\,P}$:

$$\mathbf{G} \sim \mathcal{MN}_{N,P}(\,\mathbf{M}\,,\ \mathbf{D}\,,\ \mathbf{C}\,)\,. \tag{C.1}$$

With the $N \times N$ sample-by-sample covariance captured in R and $\mathbf{M} = 0$, the component of $\mathbf{G}$ which has to be simulated is the trait-by-trait covariance $\mathbf{C}$:

$$\mathbf{G} \sim \mathcal{MN}_{N,P}(\,\mathbf{0}\,,\ \mathbf{R}\,,\ \mathbf{C}\,) \tag{C.2}$$

The structure of $\mathbf{C}$ depends on the design of the covariance effect. In order to simulate $\mathbf{C}$, $\mathbf{G}$ is first expressed in terms of a multivariate normal distribution

$$\mathrm{vec}(\mathbf{G}) \sim \mathcal{N}_{N\times P}(\,\mathbf{0}\,,\ \mathbf{C}\otimes\mathbf{R}\,)\,. \tag{C.3}$$

With the Cholesky decomposition of $\mathbf{R}$ and $\mathbf{C}$ into $\mathbf{R} = \mathbf{B}\mathbf{B}^T$ and $\mathbf{C} = \mathbf{A}\mathbf{A}^T$

$$\mathrm{vec}(\mathbf{G}) \sim \mathcal{N}_{N\times P}\Big(\,\mathbf{0}\,,\ \mathbf{A}\mathbf{A}^T\otimes\mathbf{B}\mathbf{B}^T\Big)\,, \tag{C.4}$$

which can be rearranged as

$$\text{vec}(\mathbf{G}) \sim \mathcal{N}_{N\times P}(\,\mathbf{0}\,,\ (\mathbf{A}\otimes\mathbf{B})\mathbf{I}(\mathbf{A}\otimes\mathbf{B})^T)\,)\,. \tag{C.5}$$

$\mathbf{I}$ is the identity matrix. Using the property of a normally distributed random variable $\mathbf{Y}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

$$w\mathbf{Y} \sim \mathcal{N}\,(\,w\boldsymbol{\mu}\,,\ w\boldsymbol{\Sigma}w^T\,)\,, \tag{C.6}$$

we can let $\text{vec}(\mathbf{G}) = (\mathbf{A}\otimes\mathbf{B})\text{vec}(\mathbf{Y})$ and $\mathbf{Y} \sim \mathcal{N}_{N\times P}(\,\mathbf{0}\,,\ \mathbf{I}\,)$ such that

$$(\mathbf{A}\otimes\mathbf{B})\text{vec}(\mathbf{Y}) \sim \mathcal{N}_{N\times P}(\,\mathbf{0}\,,\ (\mathbf{A}\otimes\mathbf{B})\mathbf{I}(\mathbf{A}\otimes\mathbf{B})^T\,) \tag{C.7}$$

Using [Horn & Johnson, 1985]: Lemma 4.3.1, we get

$$(\mathbf{A}\otimes\mathbf{B})\text{vec}(\mathbf{Y}) = \text{vec}(\mathbf{B}\mathbf{Y}\mathbf{A}^T) = \text{vec}(\mathbf{G}). \tag{C.8}$$

References

1000 Genomes Project Consortium (2011) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073 (cit. on p. 36).

1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422): 56–65 (cit. on pp. 36, 82).

1000 Genomes Project Consortium (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* 526: 75–81 (cit. on pp. 36, 47, 73, 157, 159).

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., & al. (2000) The genome sequence of Drosophila melanogaster. *Science* 287(5461): 2185–95 (cit. on p. 33).

Adhikari, K., Fontanil, T., Cal, S., Mendoza-Revilla, J., Fuentes-Guajardo, M., Chacón-Duque, J.-C., Al-Saadi, F., Johansson, J. A., Quinto-Sanchez, M., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Lozano, R. B., Pérez, G. M., Gómez-Valdés, J., Villamil-Ramírez, H., Hunemeier, T., Ramallo, V., Silva De Cerqueira, C. C., Hurtado, M., & al. (2016) A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nature Communications* 7: 1–12 (cit. on p. 38).

Agopian, A. J., Mitchell, L. E., Glessner, J., Bhalla, A. D., Sewda, A., Hakonarson, H., & Goldmuntz, E. (2014) Genome-Wide Association Study of Maternal and Inherited Loci for Conotruncal Heart Defects. *PLoS ONE* 9(5): e96057 (cit. on p. 69).

Ahram, D., Sato, T. S., Kohilan, A., Tayeh, M., Chen, S., Leal, S., Al-Salem, M., & El-Shanti, H. (2009) A homozygous mutation in ADAMTSL4 causes autosomal-recessive isolated ectopia lentis. *The American Journal of Human Genetics* 84(2): 274–8 (cit. on p. 189).

Eu-Ahsunthornwattana, J., Miller, N. E., Fakiola, M., Wellcome Trust Case Control Consortium, Jeronimo, S. M. B., Blackwell, J. M., & Cordell, H. J. (2014) Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. *PLoS Genetics* 10(7): e1004445 (cit. on pp. 52, 54).

Aird, I., Bentall, H. H., Mehigan, J. A., & Roberts, J. A. F. (1954) The blood groups in relation to peptic ulceration and carcinoma of colon, rectum, breast, and bronchus; an association between the ABO groups and peptic ulceration. *British Medical Journal* 2(4883): 315–21 (cit. on p. 34).

Aird, I., Bentall, H. H., & Roberts, J. A. F. (1953) A relationship between cancer of stomach and the ABO blood groups. *British Medical Journal* 1(4814): 799–801 (cit. on p. 34).

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F. J., Murphy, D. N., Nag, R., Ruffier, M., Schuster, M., Tang, Y. A., Vogel, J.-H., & al. (2016) The Ensembl gene annotation system. *Database* 2016 (cit. on pp. 119, 171).

Allen, G. E. (1968) Thomas Hunt Morgan and the Problem of Natural Selection. *Journal of the History of Biology* 1: 113–139 (cit. on p. 29).

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010) Data quality control in genetic case-control association studies. *Nature Protocols* 5(9): 1564–73 (cit. on pp. 158, 159).

Anderson, R. H., Yanni, J., Boyett, M. R., Chandler, N. J., & Dobrzynski, H. (2009) The Anatomy of the Cardiac Conduction System. *Clinical Anatomy* 22: 99–113 (cit. on p. 64).

Anderson, S. (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research* 9(13): 3015–27 (cit. on p. 32).

Anttila, V., Stefansson, H., Kallela, M., Todt, U., Terwindt, G. M., Calafato, M. S., Nyholt, D. R., Dimas, A. S., Freilinger, T., Müller-Myhsok, B., Artto, V., Inouye, M., Alakurtti, K., Kaunisto, M. a., Hämäläinen, E., de Vries, B., Stam, A. H., Weller, C. M., Heinze, A., Heinze-Kuhn, K., & al. (2010) Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. *Nature Genetics* 42(10): 869–873 (cit. on p. 43).

Arndt, A.-K., Schafer, S., Drenckhahn, J.-D., Sabeh, M. K., Plovie, E. R., Caliebe, A., Klopocki, E., Musso, G., Werdich, A. A., Kalwa, H., Heinig, M., Padera, R. F.,

Wassilew, K., Bluhm, J., Harnack, C., Martitz, J., Barton, P. J., Greutmann, M., Berger, F., Hubner, N., & al. (2013) Fine Mapping of the 1p36 Deletion Syndrome Identifies Mutation of PRDM16 as a Cause of Cardiomyopathy. *The American Journal of Human Genetics* 93(1): 67–77 (cit. on pp. 171, 189).

Arnett, D. K., Meyers, K. J., Devereux, R. B., Tiwari, H. K., Gu, C. C., Vaughan, L. K., Perry, R. T., Patki, A., Claas, S. A., Sun, Y. V., Broeckel, U., & Kardia, S. L. (2011) Genetic Variation in NCAM1 Contributes to Left Ventricular Wall Thickness in Hypertensive Families. *Circulation Research* 108(3): 279–283 (cit. on p. 69).

Arnett, D. K., Li, N., Tang, W., Rao, D. C., Devereux, R. B., Claas, S. A., Kraemer, R., & Broeckel, U. (2009) Genome-wide association study identifies single-nucleotide polymorphism in KCNB1 associated with left ventricular mass in humans: the HyperGEN Study. *BMC Medical Genetics* 10: 43 (cit. on p. 155).

Aschard, H., Vilhjálmsson, B. J., Greliche, N., Morange, P.-E., Trégouët, D.-A., & Kraft, P. (2014) Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics* 94(5): 662–76 (cit. on pp. 54, 146).

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., Lambourne, J. J., Sivapalaratnam, S., Downes, K., Kundu, K., Bomba, L., Berentsen, K., Bradley, J. R., Daugherty, L. C., Delaneau, O., Freson, K., & al. (2016) The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* 167(5): 1415–1429.e19 (cit. on p. 147).

Astle, W. & Balding, D. J. (2009) Population Structure and Cryptic Relatedness in Genetic Association Studies. EN. *Statistical Science* 24(4): 451–471 (cit. on p. 54).

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., & al. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* 465(7298): 627–31 (cit. on pp. 54, 104, 105).

Avery, C. L., He, Q., North, K. E., Ambite, J. L., Boerwinkle, E., Fornage, M., Hindorff, L. A., Kooperberg, C., Meigs, J. B., Pankow, J. S., Pendergrass, S. A., Psaty, B. M., Ritchie, M. D., Rotter, J. I., Taylor, K. D., Wilkens, L. R., Heiss, G., & Lin, D. Y. (2011) A Phenomics-Based Strategy Identifies Loci on APOC1, BRAP, and PLCG1 Associated with Metabolic Syndrome Phenotype Domains. *PLoS Genetics* 7(10): e1002322 (cit. on pp. 144, 194).

Avery, O. T., Macleod, C. M., & McCarty, M. (1944) Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. *The Journal of Experimental Medicine* 79(2): 137–58 (cit. on p. 31).

Babisch, W. (2014) Updated exposure-response relationship between road traffic noise and coronary heart diseases: A meta-analysis. *Noise and Health* 16(68): 1 (cit. on p. 67).

Bacanu, S.-A., Devlin, B., & Roeder, K. (2002) Association studies for quantitative traits in structured populations. *Genetic Epidemiology* 22(1): 78–93 (cit. on p. 48).

Balding, D. J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7(10): 781–791 (cit. on pp. 44, 47).

Baron, M., Risch, N., Hamburger, R., Mandel, B., Kushner, S., Newman, M., Drumer, D., & Belmaker, R. H. (1987) Genetic linkage between X-chromosome markers and bipolar affective illness. *Nature* 326(6110): 289–292 (cit. on p. 34).

Bateson, W. (1902) *Mendel's Principles of Heredity: A Defense*. Ed. by C. J. Clay and Sons. Cambridge, UK: Cambridge University Press: 1–212 (cit. on pp. 26, 29).

Bateson, W. (1909) *Mendel's Principles of Heredity*. Cambridge: Cambridge University Press (cit. on p. 26).

Bateson, W., Saunders, E. R., & Punnett, R. C. (1905) Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society* 2: 1–55, 80–99 (cit. on p. 28).

Beaujean, A. A. (2015) *R Package: BaylorEdPsych* (cit. on p. 110).

Belkin, M. & Niyogi, P. (2003) Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation* 15: 1373–1396 (cit. on pp. 130, 134).

Benjamini, Y. & Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1): 289–300 (cit. on pp. 46, 115).

Bernstein, F. (1930) Ueber die Erblichkeit der Blutgruppen. *Zeitschrift für induktive Abstammungs- und Vererbungslehre* 54: 400–426 (cit. on pp. 33, 34).

Betts, G. J., Desaix, P., Johnson, E., Johnson, J. E., Korol, O., Kruse, D., Poe, B., Wise, J. A., Womble, M., & Young, K. A. (2013) *Anatomy and Physiology*. Houston: OpenStax: 1420 (cit. on pp. 61, 63).

Bhatnagar, A. (2004) Cardiovascular pathophysiology of environmental pollutants. *American Journal of Physiology. Heart and Circulatory Physiology* 286(2): H479–85 (cit. on p. 67).

Bickel, P. J. & Levina, E. (2008) Regularized estimation of large covariance matrices. *The Annals of Statistics* 36(1): 199–227 (cit. on p. 104).

Biffi, C., de Marvao, A., Attard, M. I., Dawes, T. J., Whiffin, N., Bai, W., Shi, W., Francis, C., Meyer, H., Buchan, R., Cook, S. A., Rueckert, D., & O'Regan, D. P. (2017) Three-dimensional Cardiovascular Imaging-Genetics: A Mass Univariate Framework. *Bioinformatics* (cit. on pp. 70, 157).

Bird, T. D. (1993) Are linkage studies boring? *Nature Genetics* 4(3): 213–214 (cit. on p. 34).

Bleyl, S. B., Mumford, B. R., Thompson, V., Carey, J. C., Pysher, T. J., Chin, T. K., & Ward, K. (1997) Neonatal, Lethal Noncompaction of the Left Ventricular Myocardium Is Allelic with Barth Syndrome. *The American Journal of Human Genetics* 61: 868–872 (cit. on p. 180).

Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V., & Kruglyak, L. (2013) Finding the sources of missing heritability in a yeast cross. *Nature* 494(7436): 234–7 (cit. on pp. 47, 54, 105, 108, 109).

Bolormaa, S., Pryce, J. E., Reverter, A., Zhang, Y., Barendse, W., Kemper, K., Tier, B., Savin, K., Hayes, B. J., & Goddard, M. E. (2014) A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS genetics* 10(3): e1004198 (cit. on pp. 55, 104).

Bonne, G., Carrier, L., Bercovici, J., Cruaud, C., Richard, P., Hainque, B., Gautel, M., Labeit, S., James, M., Beckmann, J., Weissenbach, J., Vosberg, H.-P., Fiszman, M., Komajda, M., & Schwartz, K. (1995) Cardiac myosin binding protein–C gene splice acceptor site mutation is associated with familial hypertrophic cardiomyopathy. *Nature Genetics* 11(4): 438–440 (cit. on p. 68).

Bonnet, C. (1779) *Oeuvres d'histoire naturelle et de philosophie de Charles Bonnet ...* Neuchatel: Chez S. Fauche: 1–444 (cit. on p. 24).

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *The American Journal of Human Genetics* 32(3): 314–31 (cit. on p. 32).

Bottolo, L., Chadeau-Hyam, M., Hastie, D. I., Zeller, T., Liquet, B., Newcombe, P., Yengo, L., Wild, P. S., Schillert, A., Ziegler, A., Nielsen, S. F., Butterworth, A. S.,

Ho, W. K., Castagné, R., Munzel, T., Tregouet, D., Falchi, M., Cambien, F., Nordestgaard, B. G., Fumeron, F., & al. (2013) GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS genetics* 9(8): e1003657 (cit. on p. 38).

Boveri, T. (1902) Über mehrpolige Mitosen als Mittel zur Analyse des Zellkerns: 67–90 (cit. on p. 28).

Boyden, L. M., Choi, M., Choate, K. A., Nelson-Williams, C. J., Farhi, A., Toka, H. R., Tikhonova, I. R., Bjornson, R., Mane, S. M., Colussi, G., Lebel, M., Gordon, R. D., Semmekrot, B. A., Poujol, A., Välimäki, M. J., De Ferrari, M. E., Sanjad, S. A., Gutkin, M., Karet, F. E., Tucci, J. R., & al. (2012) Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities. *Nature* 482(7383): 98–102 (cit. on p. 68).

Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. (2002) Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* 296(5568) (cit. on pp. 47, 118).

Brem, R. B. & Kruglyak, L. (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America* 102(5): 1572–7 (cit. on p. 118).

Brent, R. P. (1971) An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal* 14(4): 422–425 (cit. on p. 59).

Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., Kaufman, J. D., & American Heart Association Council on Epidemiology and Prevention, Council on the Kidney in Cardiovascular Disease, and Council on Nutrition, Physical Activity and Metabolism (2010) Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement From the American Heart Association. *Circulation* 121(21): 2331–2378 (cit. on p. 67).

Brown, T. A. (2002) *Genomes*. 2nd. Wiley-Liss: 600 (cit. on p. 30).

Browning, S. R. (2008) Estimation of Pairwise Identity by Descent From Dense Genetic Marker Data in a Population Sample of Haplotypes. *Genetics* 178(4) (cit. on p. 54).

Browning, S. R. & Browning, B. L. (2010) High-resolution detection of identity by descent in unrelated individuals. *The American Journal of Human Genetics* 86(4): 526–39 (cit. on p. 52).

Browning, S. R., Browning, B. L., Todd, J., Clayton, D., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H., & Al., E. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* 81(5): 1084–97 (cit. on p. 37).

Broyden, C. G. (1965) A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation* 19: 577–593 (cit. on pp. 51, 59).

Budde, B. S., Binner, P., Waldmüller, S., Höhne, W., Blankenfeldt, W., Hassfeld, S., Brömsen, J., Dermintzoglou, A., Wieczorek, M., May, E., Kirst, E., Selignow, C., Rackebrandt, K., Müller, M., Goody, R. S., Vosberg, H.-P., Nürnberg, P., & Scheffold, T. (2007) Noncompaction of the Ventricular Myocardium Is Associated with a De Novo Mutation in the β-Myosin Heavy Chain Gene. *PLoS ONE* 2(12). Ed. by I. Schrijver: e1362 (cit. on p. 180).

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Consortium, S. W. G. o. t. P. G., Patterson, N., Daly, M. J., Price, A. L., & Neale, B. M. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* advance on(3): 291–295 (cit. on p. 48).

Bulmer, M. G. (2003) *Francis Galton: Pioneer of heredity and biometry*. Baltimore: Johns Hopkins University Press: 357 (cit. on p. 27).

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., Barrett, J. C., Burton, P. R., Davison, D., Donnelly, P., Easton, D., Evans, D., Leung, H.-T., Marchini, J. L., & al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145): 661–678 (cit. on p. 37).

Bush, W. S. & Moore, J. H. (2012) Chapter 11: Genome-wide association studies. *PLoS Computational Biology* 8(12): e1002822 (cit. on p. 42).

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995) A Limited Memory Algorithm for Bound Constrained Optimization. en. *SIAM Journal on Scientific Computing* 16(5): 1190–1208 (cit. on p. 91).

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* 282(5396): 2012–8 (cit. on p. 33).

Cameron, J. R., Loh, E. Y., & Davis, R. W. (1979) Evidence for transposition of dispersed repetitive DNA families in yeast. *Cell* 16(4): 739–751 (cit. on p. 32).

Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., Altshuler, D., Ardlie, K. G., & Hirschhorn, J. N. (2005) Demonstrating stratification in a European American population. *Nature Genetics* 37(8): 868–872 (cit. on p. 48).

Candille, S. I., Absher, D. M., Beleza, S., Bauchet, M., McEvoy, B., Garrison, N. A., Li, J. Z., Myers, R. M., Barsh, G. S., Tang, H., & Shriver, M. D. (2012) Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations. *PloS ONE* 7(10): e48294 (cit. on p. 38).

Cannavò, E., Koelling, N., Harnett, D., Garfield, D., Casale, F. P., Ciglar, L., Gustafson, H. E., Viales, R. R., Marco-Ferreres, R., Degner, J. F., Zhao, B., Stegle, O., Birney, E., & Furlong, E. E. M. (2016) Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature* 541(7637): 402–406 (cit. on p. 90).

Captur, G., Lopes, L. R., Patel, V., Li, C., Bassett, P., Syrris, P., Sado, D. M., Maestrini, V., Mohun, T. J., McKenna, W. J., Muthurangu, V., Elliott, P. M., & Moon, J. C. (2014) Abnormal cardiac formation in hypertrophic cardiomyopathy: fractal analysis of trabeculae and preclinical gene expression. *Circulation. Cardiovascular genetics* 7(3): 241–8 (cit. on pp. 180, 183, 185, 192).

Captur, G., Muthurangu, V., Cook, C., Flett, A. S., Wilson, R., Barison, A., Sado, D. M., Anderson, S., McKenna, W. J., Mohun, T. J., Elliott, P. M., & Moon, J. C. (2013) Quantification of left ventricular trabeculae using fractal analysis. en. *Journal of Cardiovascular Magnetic Resonance* 15(1): 36 (cit. on pp. 180–182).

Captur, G., Zemrak, F., Muthurangu, V., Petersen, S. E., Li, C., Bassett, P., Kawel-Boehm, N., McKenna, W. J., Elliott, P. M., Lima, J. A. C., Bluemke, D. A., & Moon, J. C. (2015) Fractal Analysis of Myocardial Trabeculations in 2547 Study Participants: Multi-Ethnic Study of Atherosclerosis. EN. *Radiology* 277(3): 707–15 (cit. on pp. 180, 189).

Carrier, L., Hengstenberg, C., Beckmann, J. S., Guicheney, P., Dufour, C., Bercovici, J., Dausse, E., Berebbi-Bertrand, I., Wisnewsky, C., Pulvenis, D., Fetler, L., Vignal, A., Weissenbach, J., Hillaire, D., Feingold, J., Bouhour, J.-B., Hagege, A., Desnos, M., Isnard, R., Dubourg, O., & al. (1993) Mapping of a novel gene for familial hypertrophic cardiomyopathy to chromosome 11. *Nature Genetics* 4(3): 311 (cit. on p. 68).

Carvajal-Rodríguez, A. (2008) GENOMEPOP: A program to simulate genomes in populations. *BMC Bioinformatics* 9(1): 223 (cit. on p. 72).

Casale, F. P., Horta, D., Rakitsch, B., & Stegle, O. (2017) Joint genetic analysis using variant sets reveals polygenic gene-context interactions. *PLoS Genetics* 13(4). Ed. by M. P. Epstein: e1006693 (cit. on p. 193).

Casale, F. P., Rakitsch, B., Lippert, C., & Stegle, O. (2015) Efficient set tests for the genetic analysis of correlated traits. *Nature Methods* 12: 755–758 (cit. on pp. 38, 42, 43, 50, 51, 53, 57, 59, 72, 79, 80, 89, 90, 92, 99, 193).

Chakravarti, A. (1999) Population genetics—making sense out of sequence. *Nature Genetics* 21(1 Suppl): 56–60 (cit. on p. 36).

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4(1): 7 (cit. on pp. 76, 117, 158).

Chargaff, E., Lipshitz, R., & Green, C. (1952) Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *The Journal of Biological Chemistry* 195(1): 155–60 (cit. on p. 31).

Chargaff, E., Vischer, E., Doninger, R., Green, C., & Fernanda, M. (1949) The composition of the desoxypentose nucleic acids of thymus and spleen. *The Journal of Biological Chemistry* 177(1): 405–16 (cit. on p. 31).

Chen, H., Zhang, W., Li, D., Cordes, T. M., Mark Payne, R., & Shou, W. (2009) Analysis of Ventricular Hypertrabeculation and Noncompaction Using Genetically Engineered Mouse Models. *Pediatric Cardiology* 30(5): 626–634 (cit. on pp. 179, 180).

Chen, J. & Chien, K. R. (1999) Complexity in simplicity: monogenic disorders and complex cardiomyopathies. *The Journal of Clinical Investigation* 103(11): 1483–5 (cit. on pp. 155, 156).

Christoffels, V. M., Habets, P. E., Franco, D., Campione, M., de Jong, F., Lamers, W. H., Bao, Z. Z., Palmer, S., Biben, C., Harvey, R. P., & Moorman, A. F. (2000) Chamber formation and morphogenesis in the developing mammalian heart. *Developmental Biology* 223(2): 266–78 (cit. on p. 64).

Christoffels, V. M. & Moorman, A. F. M. (2009) Development of the cardiac conduction system: why are some regions of the heart more arrhythmogenic than others? *Circulation: Arrhythmia and electrophysiology* 2(2): 195–207 (cit. on p. 66).

Christophersen, I. E., Rienstra, M., Roselli, C., Yin, X., Geelhoed, B., Barnard, J., Lin, H., Arking, D. E., Smith, A. V., Albert, C. M., Chaffin, M., Tucker, N. R., Li, M., Klarin, D., Bihlmeyer, N. A., Low, S.-K., Weeke, P. E., Müller-Nurasyid, M., Gustav

Smith, J., Brody, J. A., & al. (2017) Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation (cit. on p. 189).

Claes, P., Liberton, D. K., Daniels, K., Rosana, K. M., Quillen, E. E., Pearson, L. N., McEvoy, B., Bauchet, M., Zaidi, A. A., Yao, W., Tang, H., Barsh, G. S., Absher, D. M., Puts, D. A., Rocha, J., Beleza, S., Pereira, R. W., Baynam, G., Suetens, P., Vandermeulen, D., & al. (2014) Modeling 3D Facial Shape from DNA. *PLoS Genetics* 10(3). Ed. by D. Luquetti: e1004224 (cit. on p. 54).

Cohen, J. (1992) A power primer. *Psychological Bulletin* 112(1): 155–9 (cit. on pp. 83, 147).

Cohen, S. N. (1973) Recircularization and Autonomous Replication of a Sheared R-Factor DNA Segment in Escherichia coli Transformants. *Proceedings of the National Academy of Sciences of the United States of America* 70(5): 1293–1297 (cit. on p. 32).

Coifman, R. R. & Lafon, S. (2006) Diffusion maps. *Applied Computational Harmonic Analysis* 21: 5–30 (cit. on pp. 130, 131).

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., & Zucker, S. W. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps (cit. on p. 130).

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 6(4): 330–51 (cit. on p. 114).

Comon, P. (1994) Independent component analysis. *Signal Processing Comon* 36(36): 28–314 (cit. on p. 128).

Comte de Buffon, G.-L. L. (1749) *Oeuvres d'Histoire Naturelle*. Volume 8. Imprimerie royale (cit. on p. 24).

Cook, S. & O'Regan, D. (2010) *Digital Heart Project* (cit. on p. 156).

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., Majeti, R., & Chang, H. Y. (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics* 48(10): 1193–1203 (cit. on p. 132).

Correns, C. (1900) G. Mendel's Regel über das Verhalten der Nachkommenschaft der Rassenbastarde. *Berichte der deutschen botanischen Gesellschaft.* 18: 158–168 (cit. on p. 26).

Cosselman, K. E., Navas-Acien, A., & Kaufman, J. D. (2015) Environmental factors in cardiovascular disease. *Nature Reviews Cardiology* 12(11): 627–642 (cit. on p. 67).

Crick, F. H. (1958) On protein synthesis. *Symposia of the Society for Experimental Biology* 12: 138–63 (cit. on p. 32).

Crick, F. H., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961) General nature of the genetic code for proteins. *Nature* 192: 1227–32 (cit. on p. 32).

Crowley, J. J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I. K., Kim, Y., Wang, J. R., Morgan, A. P., Calaway, J. D., Aylor, D. L., Yun, Z., Bell, T. A., Buus, R. J., Calaway, M. E., Didion, J. P., Gooch, T. J., Hansen, S. D., Robinson, N. N., Shaw, G. D., Spence, J. S., & al. (2015) Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genetics* 47(4): 353–360 (cit. on p. 132).

Cunnington, R. H., Northcott, J. M., Ghavami, S., Filomeno, K. L., Jahan, F., Kavosh, M. S., Davies, J. J. L., Wigle, J. T., & Dixon, I. M. C. (2014) The Ski–Zeb2–Meox2 pathway provides a novel mechanism for regulation of the cardiac myofibroblast phenotype. *Journal of Cell Science* 127(1) (cit. on p. 170).

Cunnington, R. H., Wang, B., Ghavami, S., Bathe, K. L., Rattan, S. G., & Dixon, I. M. C. (2010) Antifibrotic properties of c-Ski and its regulation of cardiac myofibroblast phenotype and contractility. *American Journal of Physiology - Cell Physiology* 300(1) (cit. on p. 170).

Cupples, L. A., Arruda, H. T., Benjamin, E. J., D'Agostino, R. B., Demissie, S., DeStefano, A. L., Dupuis, J., Falls, K. M., Fox, C. S., Gottlieb, D. J., Govindaraju, D. R., Guo, C.-Y., Heard-Costa, N. L., Hwang, S.-J., Kathiresan, S., Kiel, D. P., Laramie, J. M., Larson, M. G., Levy, D., Liu, C.-Y., & al. (2007) The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Medical Genetics* 8 Suppl 1(Suppl 1): S1 (cit. on pp. 38, 69).

Darwin, C. R. (1859) *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life*. London: John Murray: 556 (cit. on p. 24).

Darwin, C. R. (1868) *The variation of animals and plants under domestication.* 1st ed. London: John Murray (cit. on p. 24).

Davies, M. & McKenna, W. (1995) Hypertrophic cardiomyopathy: pathology and pathogenesis. *Histopathology* 26(6): 493–500 (cit. on pp. 155, 156).

Dawber, T. R., Meadors, G. F., & Moore, F. E. J. (1951) Epidemiological approaches to heart disease: the Framingham Study. *American Journal of Public Health and the Nation's Health* 41(3): 279–81 (cit. on p. 68).

De Vries, H. (1900) Sur la loi de disjonction des hybrides. Comptes Rendus de l'Academie des Sciences (Paris). *Comptes Rendus de l'Academie des Sciences* 130: 845–847 (cit. on p. 26).

De Jong, F., Opthof, T., Wilde, A. A., Janse, M. J., Charles, R., Lamers, W. H., & Moorman, A. F. (1992) Persisting zones of slow impulse conduction in developing chicken hearts. *Circulation Research* 71(2) (cit. on p. 64).

De Marvao, A., Dawes, T. J., Shi, W., Minas, C., Keenan, N. G., Diamond, T., Durighel, G., Montana, G., Rueckert, D., Cook, S. A., & O'Regan, D. P. (2014) Population-based studies of myocardial hypertrophy: high resolution cardiovascular magnetic resonance atlases improve statistical power. *Journal of Cardiovascular Magnetic Resonance* 16(1): 16 (cit. on pp. 156, 160, 162, 181).

De Ridder, D. & Duin, R. (2002) "Locally linear embedding for classification". *Technical Report PH-2002-01*. Delft University of Technology. Delft (cit. on p. 134).

Dehghan, A., Bis, J. C., White, C. C., Smith, A. V., Morrison, A. C., Cupples, L. A., Trompet, S., Chasman, D. I., Lumley, T., Völker, U., Buckley, B. M., Ding, J., Jensen, M. K., Folsom, A. R., Kritchevsky, S. B., Girman, C. J., Ford, I., Dörr, M., Salomaa, V., Uitterlinden, A. G., & al. (2016) Genome-Wide Association Study for Incident Myocardial Infarction and Coronary Heart Disease in Prospective Cohort Studies: The CHARGE Consortium. *PLoS ONE* 11(3). Ed. by M.-P. Dubé: e0144997 (cit. on p. 189).

Delaneau, O., Marchini, J., & Zagury, J.-F. (2012) A linear complexity phasing method for thousands of genomes. *Nature Methods* 9(2): 179–81 (cit. on p. 158).

Delaneau, O., Zagury, J.-F., & Marchini, J. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* 10(1): 5–6 (cit. on p. 158).

Deng, Q., Ramsköld, D., Reinius, B., & Sandberg, R. (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343(6167): 193–6 (cit. on p. 132).

Denny, J. C., Ritchie, M. D., Crawford, D. C., Schildcrout, J. S., Ramirez, A. H., Pulley, J. M., Basford, M. A., Masys, D. R., Haines, J. L., & Roden, D. M. (2010) Identification of genomic predictors of atrioventricular conduction: Using electronic med-

ical records as a tool for genome science. *Circulation* 122(20): 2016–2021 (cit. on p. 189).

Devlin, B., Bacanu, S.-A., & Roeder, K. (2004) Genomic Control to the extreme. *Nature Genetics* 36(11): 1129–1130, author reply 1131 (cit. on p. 48).

Devlin, B. & Roeder, K. (1999) Genomic control for association studies. *Biometrics* 55(4): 997–1004 (cit. on p. 48).

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T. P., Bowden, D. W., Smith, D. F., Lander, E. S., Botstein, D., Powers, J. A., Watt, D. E., Kauffman, E. R., Bricker, A., Phipps, P., Muller-Kahle, H., Fulton, T. R., Ng, S., Schumm, J. W., & al. (1987) A Genetic Linkage Map of the Human Genome. *Cell* 51(0): 319–337 (cit. on p. 32).

Donoho, D. & Jin, J. (2006) Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *The Annals of Statistics* 34(6): 2980–3018 (cit. on p. 46).

Doyle, A. J., Doyle, J. J., Bessling, S. L., Maragh, S., Lindsay, M. E., Schepers, D., Gillis, E., Mortier, G., Homfray, T., Sauls, K., Norris, R. A., Huso, N. D., Leahy, D., Mohr, D. W., Caulfield, M. J., Scott, A. F., Destrée, A., Hennekam, R. C., Arn, P. H., Curry, C. J., & al. (2012) Mutations in the TGF-β repressor SKI cause Shprintzen-Goldberg syndrome with aortic aneurysm. *Nature Genetics* 44(11): 1249–1254 (cit. on p. 170).

Dunn, O. J. (1961) Multiple comparisons among means. *Journal of the American Statistical Association* 56(293): 52–64 (cit. on pp. 46, 118).

Dunwell, J. M. (2007) 100 years on: a century of genetics. *Nature Reviews Genetics* 8(3): 231–235 (cit. on p. 29).

Durham, D. & Worthley, L. I. G. (2002) Cardiac arrhythmias: diagnosis and management. The tachycardias. *Critical Care and Resuscitation* 4(1): 35–53 (cit. on p. 66).

East, E. M. (1910) *A Mendelian Interpretation of Variation that is Apparently Continuous* (cit. on p. 29).

Eden, T. & Fisher, R. A. (1929) Studies in crop variation: VI. Experiments on the response of the potato to potash and nitrogen. *The Journal of Agricultural Science* 19(02): 201 (cit. on p. 31).

Edwards, A. W. F. (2013) Robert Heath Lock and his textbook of genetics, 1906. *Genetics* 194(3): 529–37 (cit. on pp. 26, 28).

Ehrenreich, I. M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J. A., Gresham, D., Caudy, A. A., & Kruglyak, L. (2010) Dissection of genetically complex traits with

extremely large pools of yeast segregants. *Nature* 464(7291): 1039–1042 (cit. on pp. 47, 118).

Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., Chasman, D. I., Smith, A. V., Tobin, M. D., Verwoert, G. C., Hwang, S.-J., Pihur, V., Vollenweider, P., O'Reilly, P. F., Amin, N., Bragg-Gresham, J. L., Teumer, A., Glazer, N. L., Launer, L., Zhao, J. H., Aulchenko, Y., & al. (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. en. *Nature* 478(7367): 103–9 (cit. on pp. 69, 189).

Eke, A., Herman, P., Kocsis, L., & Kozak, L. R. (2002) Fractal characterization of complexity in temporal physiological signals. *Physiological Measurement* 23(1): R1–38 (cit. on p. 181).

Elledge, S. J. & Davis, R. W. (1990) Two genes differentially regulated in the cell cycle and by DNA-damaging agents encode alternative regulatory subunits of ribonucleotide reductase. *Genes & Development* 4(5): 740–51 (cit. on p. 120).

Eriksson, N., Macpherson, J. M., Tung, J. Y., Hon, L. S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I., & Mountain, J. (2010) Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits. *PLoS Genetics* 6(6). Ed. by G. Gibson: e1000993 (cit. on p. 38).

Etzel, C. J., Shete, S., Beasley, T. M., Fernandez, J. R., Allison, D. B., & Amos, C. I. (2003) Effect of Box-Cox transformation on power of Haseman-Elston and maximum-likelihood variance components tests to detect quantitative trait Loci. *Human Heredity* 55(2-3): 108–16 (cit. on p. 43).

Ewens, W. J. & Spielman, R. S. (1995) The Transmission/Disequilibrium Test: History, Subdivision, and Admixture. *The American Journal of Human Genetics* 57: 455–464 (cit. on p. 48).

Ewing, G. & Hermisson, J. (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* 26(16): 2064–5 (cit. on p. 72).

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016) The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics* 24: 1202–1205 (cit. on pp. 47, 83).

Ferreira, M. A. R. & Purcell, S. M. (2009) A multivariate test of association. *Bioinformatics* 25(1): 132–3 (cit. on p. 55).

Filippini, N., MacIntosh, B. J., Hough, M. G., Goodwin, G. M., Frisoni, G. B., Smith, S. M., Matthews, P. M., Beckmann, C. F., & Mackay, C. E. (2009) Distinct patterns of brain activity in young carriers of the APOE-epsilon4 allele. *Proceedings of the National Academy of Sciences of the United States of America* 106(17): 7209–14 (cit. on p. 156).

Fisher, R. A. (1912) On an Absolute Criterion for Fitting Frequency Curves. *Messenger of Mathematics* 41: 155–160 (cit. on p. 30).

Fisher, R. A. (1918) The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Philosophical transactions of the Royal Society of Edinburgh* 52: 399–433 (cit. on p. 31).

Fisher, R. A. (1921) Studies in Crop Variation. I. An examination of the yield of dressed grain from Broadbalk. *The Journal of Agricultural Science* 11(02): 107 (cit. on p. 31).

Fisher, R. A. (1922a) On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society* 85(1): 87 (cit. on p. 30).

Fisher, R. A. (1922b) On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London A* 222(594-604) (cit. on p. 30).

Fisher, R. A. (1922c) The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients. *Journal of the Royal Statistical Society* 85(4): 597 (cit. on p. 30).

Fisher, R. A. (1922d) The Systematic Location of Genes by Means of Crossover Observations. *The American Naturalist* 56: 406–411 (cit. on p. 30).

Fisher, R. A. (1924a) "On a distribution yielding the error functions of several well known statistics". *Proceedings of the International Congress of Mathematicians*. Toronto: 10 (cit. on p. 31).

Fisher, R. A. (1924b) The Distribution of the Partial Correlation Coefficient. *Metron* 3: 329–332 (cit. on p. 30).

Fisher, R. A. (1928) The General Sampling Distribution of the Multiple Correlation Coefficient. *Proceedings of the Royal Society of London. Series A* 121(788): 654–673 (cit. on p. 30).

Fisher, R. A. (1930) *The Genetical Theory of Natural Selection*. 2nd ed. Oxford: Clarendon Press (cit. on p. 31).

Fisher, R. A. & Mackenzie, W. A. (1923) Studies in crop variation. II. The manurial response of different potato varieties. *The Journal of Agricultural Science* 13(03): 311 (cit. on p. 31).

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., & Merrick, J. M. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269(5223): 496–512 (cit. on p. 33).

Flemming, W. (1878) Zur Kenntniss der Zelle und ihrer Theilungs-Erscheinungen. *Schriften des Naturwissenschaftlichen Vereins für Schleswig-Holstein* 3: 23–27 (cit. on p. 27).

Florian, A., Masci, P. G., De Buck, S., Aquaro, G. D., Claus, P., Todiere, G., Van Cleemput, J., Lombardi, M., & Bogaert, J. (2012) Geometric Assessment of Asymmetric Septal Hypertrophic Cardiomyopathy by CMR. *JACC: Cardiovascular Imaging* 5(7): 702–711 (cit. on p. 155).

Foiani, M., Cigan, A. M., Paddon, C. J., Harashima, S., Hinnebusch, A. G., Pavitt, G., Ashe, M., Grant, C., Cyert, M., Hughes, T., Boone, C., Andrews, B., Chua, G., Friesen, H., Goldberg, D., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., & al. (1991) GCD2, a translational repressor of the GCN4 gene, has a general function in the initiation of protein synthesis in Saccharomyces cerevisiae. *Molecular and Cellular Biology* 11(6): 3203–3216 (cit. on p. 120).

Ford, E. S., Ajani, U. A., Croft, J. B., Critchley, J. A., Labarthe, D. R., Kottke, T. E., Giles, W. H., & Capewell, S. (2007) Explaining the Decrease in U.S. Deaths from Coronary Disease, 1980–2000. *New England Journal of Medicine* 356(23): 2388–2398 (cit. on p. 67).

Fox, E. R., Musani, S. K., Barbalic, M., Lin, H., Yu, B., Ogunyankin, K. O., Smith, N. L., Kutlar, A., Glazer, N. L., Post, W. S., Paltoo, D. N., Dries, D. L., Farlow, D. N., Duarte, C. W., Kardia, S. L., Meyers, K. J., Sun, Y. V., Arnett, D. K., Patki, A. A., Sha, J., & al. (2013) Genome-wide association study of cardiac structure and systolic function in African Americans: the Candidate Gene Association Resource (CARe) study. *Circulation. Cardiovascular genetics* 6(1): 37–46 (cit. on p. 155).

Franceschini, N., Fox, E., Zhang, Z., Edwards, T. L., Nalls, M. A., Sung, Y. J., Tayo, B. O., Sun, Y. V., Gottesman, O., Adeyemo, A., Johnson, A. D., Young, J. H., Rice, K., Duan, Q., Chen, F., Li, Y., Tang, H., Fornage, M., Keene, K. L., Andrews, J. S., & al. (2013) Genome-wide Association Analysis of Blood-Pressure Traits in African-Ancestry Individuals Reveals Common Associated Genes in African and Non-African Populations. *The American Journal of Human Genetics* 93(3): 545–554 (cit. on p. 38).

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R. B., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A.-M., Ness, A. R., Ebrahim, S., Lawlor, D. A., & al. (2007) A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science* 316(5826) (cit. on p. 37).

Furrer, R. & Bengtsson, T. (2007) Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis* 98(2): 227–255 (cit. on p. 104).

Galton, F. (1886) Regression towards mediocrity in heriditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246–263 (cit. on p. 27).

Galton, F. (1889) *Natural inheritance*. London: Macmillan Publishers Limited: 282 (cit. on p. 27).

Galton, F. (1901) Biometry. *Biometrika* 1(1) (cit. on p. 27).

Galwey, N. W. (2009) A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology* 33(7): 559–68 (cit. on p. 117).

Garg, V., Kathiriya, I. S., Barnes, R., Schluterman, M. K., King, I. N., Butler, C. A., Rothrock, C. R., Eapen, R. S., Hirayama-Yamada, K., Joo, K., Matsuoka, R., Cohen, J. C., & Srivastava, D. (2003) GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* 424(6947): 443–447 (cit. on p. 68).

Garson, D. G. (2015) *Missing values analysis and data imputation*. 2nd ed. Asheboro, NC: Statistical Associates Publishing: 113 (cit. on pp. 107, 110, 113).

GBE (2017) *Global Biobank Engine*. Stanford, CA (cit. on pp. 171, 190).

Ge, T., Schumann, G., & Feng, J. (2014) Imaging genetics — towards discovery neuroscience. *Quantitative Biology* 1(4): 227–245 (cit. on pp. 156, 177).

Geisterfer-Lowrance, A. T., Kass, S., Tanigawa, G., W&erg, H.-P., Mckenna, W., Seidman, C. E., & Seldmant, J. G. (1990) A Molecular Basis for Familial Hypertrophic Cardiomyopathy: A p Cardiac Myosin Heavy Chain Gene Missense Mutation. *Cell* 62: 999–1006 (cit. on p. 68).

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995) Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* 51(4): 1440 (cit. on p. 59).

Gittenberger-de Groot, A. C., Bartelings, M. M., Deruiter, M. C., & Poelmann, R. E. (2005) Basics of cardiac development for the understanding of congenital heart malformations. *Pediatric Research* 57(2): 169–76 (cit. on p. 65).

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., Jenkinson, M., & WU-Minn HCP Consortium (2013) The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* 80: 105–124 (cit. on p. 194).

Glazner, C. & Thompson, E. A. (2012) Improving pedigree-based linkage analysis by estimating coancestry among families. *Statistical Applications in Genetics and Molecular Biology* 11(2) (cit. on p. 52).

Glover, M., Ware, J. S., Henry, A., Wolley, M., Walsh, R., Wain, L. V., Xu, S., Van ', W. G., Hoff, T., Tobin, M. D., Hall, I. P., Cook, S., Gordon, R. D., Stowasser, M., & O 'shaughnessy, K. M. (2014) Detection of mutations in KLHL3 and CUL3 in families with FHHt (familial hyperkalaemic hypertension or Gordon's syndrome). *Clinical Science* 126: 721–726 (cit. on p. 68).

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., & Oliver, S. G. (1996) Life with 6000 genes. *Science* 274(5287): 546, 563–7 (cit. on p. 33).

Goodman, H. M., Olson, M. V., & Hall, B. D. (1977) Nucleotide sequence of a mutant eukaryotic gene: the yeast tyrosine-inserting ochre suppressor SUP4-o. *Proceedings of the National Academy of Sciences of the United States of America* 74(12): 5453–7 (cit. on p. 32).

Gormley, P., Anttila, V., Winsvold, B. S., Palta, P., Esko, T., Pers, T. H., Farh, K.-H., Cuenca-Leon, E., Muona, M., Furlotte, N. A., Kurth, T., Ingason, A., McMahon, G., Ligthart, L., Terwindt, G. M., Kallela, M., Freilinger, T. M., Ran, C., Gordon, S. G., Stam, A. H., & al. (2016) Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nature Genetics* 48(8): 856–866 (cit. on p. 38).

Goss, C. M. (1938) The first contractions of the heart in rat embryos. *The Anatomical Record* 70(5): 505–524 (cit. on p. 64).

Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53: 325 (cit. on pp. 125, 126, 134).

Greally, M. T. (1993) *Shprintzen-Goldberg Syndrome*. University of Washington, Seattle (cit. on p. 170).

Grodzicker, T., Williams, J., Sharp, P., & Sambrook, J. G. (1974) Physical mapping of temperature- sensitive mutations of adenoviruses. *Cold Spring Harbor Symp Quant Biol* 39: 439–446 (cit. on p. 32).

Guan, Y. & Stephens, M. (2008) Practical issues in imputation-based association mapping. *PLoS Genetics* 4(12): e1000279 (cit. on p. 43).

Gudbjartsson, D. F., Arnar, D. O., Helgadottir, A., Gretarsdottir, S., Holm, H., Sigurdsson, A., Jonasdottir, A., Baker, A., Thorleifsson, G., Kristjansson, K., Palsson, A., Blondal, T., Sulem, P., Backman, V. M., Hardarson, G. A., Palsdottir, E., Helgason, A., Sigurjonsdottir, R., Sverrisson, J. T., Kostulas, K., & al. (2007) Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* 448(7151): 353–7 (cit. on p. 189).

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E., & Martin, J. B. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306(5940): 234–238 (cit. on p. 34).

Güvenç, T. S., Erer, H. B., Altay, S., Ilhan, E., Sayar, N., & Eren, M. (2012) 'Idiopathic' acute myocardial infarction in a young patient with noncompaction cardiomyopathy. *Cardiology Journal* 19(4): 429–33 (cit. on p. 189).

Haider, A. W., Larson, M. G., Benjamin, E. J., & Levy, D. (1998) Increased left ventricular mass and hypertrophy are associated with increased risk for sudden death. *Journal of the American College of Cardiology* 32(5): 1454–9 (cit. on p. 155).

Hald, A. (1999) On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares. *Statistical Science* 14: 214–222 (cit. on p. 30).

Haldane, J. B. S. (1934) Methods for the detection of autosomal linkage in man. *Annals of Eugenics* 6(1): 26–65 (cit. on p. 34).

Haldane, J. B. S. & Smith, C. A. B. (1947) A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics* 14(1): 10–31 (cit. on p. 34).

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015) The fickle P value generates irreproducible results. *Nature Methods* 12(3): 179–185 (cit. on pp. 83, 147).

Hanchard, N. A., Swaminathan, S., Bucasas, K., Furthner, D., Fernbach, S., Azamian, M. S., Wang, X., Lewin, M., Towbin, J. A., D'Alessandro, L. C., Morris, S. A., Dreyer, W., Denfield, S., Ayres, N. A., Franklin, W. J., Justino, H., Lantin-Hermoso, M. R.,

Ocampo, E. C., Santos, A. B., Parekh, D., & al. (2016) A genome-wide association study of congenital cardiovascular left-sided lesions shows association with a locus on chromosome 20. *Human Molecular Genetics* 25(11): 2331–2341 (cit. on p. 69).

Hannah, A. (& De Vries, H. ((1950) Concerning the law of segregation in hybrids. *Genetics* 35(5): 30–32 (cit. on p. 26).

Hansson, J. H., Nelson-Williams, C., Suzuki, H., Schild, L., Shimkets, R., Lu, Y., Canessa, C., Iwasaki, T., Rossier, B., & Lifton, R. P. (1995) Hypertension caused by a truncated epithelial sodium channel γ subunit: genetic heterogeneity of Liddle syndrome. *Nature Genetics* 11(1): 76–82 (cit. on p. 68).

Hästbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., & Lander, E. (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics* 2(3): 204–211 (cit. on p. 34).

Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2009) Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research* 91: 47–60 (cit. on p. 52).

He, L.-N., Liu, Y.-J., Xiao, P., Zhang, L., Guo, Y., Yang, T.-L., Zhao, L.-J., Drees, B., Hamilton, J., Deng, H.-Y., Recker, R. R., & Deng, H.-W. (2008) Genomewide Linkage Scan for Combined Obesity Phenotypes using Principal Component Analysis. *Annals of Human Genetics* 72(3): 319–326 (cit. on p. 54).

Heather, J. M. & Chain, B. (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics* 107: 1–8 (cit. on p. 33).

Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, M. C., Speliotes, E. K., Mägi, R., Workalemahu, T., White, C. C., Bouatia-Naji, N., Harris, T. B., Berndt, S. I., Ingelsson, E., Willer, C. J., Weedon, M. N., Luan, J., Vedantam, S., & al. (2010) Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genetics* 42(11): 949–960 (cit. on p. 38).

Hendee, K., Wang, L. W., Reis, L. M., Rice, G. M., Apte, S. S., & Semina, E. V. (2017) Identification and functional analysis of an ADAMTSL1 variant associated with a complex phenotype including congenital glaucoma, craniofacial, and other systemic features in a three-generation human pedigree. *Human Mutation* (cit. on p. 189).

Henderson, C. R. & Quaas, R. L. (1976) Multiple Trait Evaluation Using Relatives' Records. *Journal of Animal Science* 43(6): 1188 (cit. on p. 57).

Herault, J. & Jutten, C. (1986) "Space or time adaptive signal processing by neural network models". *AIP Conference Proceedings*. Vol. 151. AIP: 206–211 (cit. on p. 128).

Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivières, S., Jahanshad, N., Toro, R., Wittfeld, K., Abramovic, L., Andersson, M., Aribisala, B. S., Armstrong, N. J., Bernard, M., Bohlken, M. M., Boks, M. P., Bralten, J., Brown, A. A., Mallar Chakravarty, M., Chen, Q., Ching, C. R. K., & al. (2015) Common genetic variants influence human subcortical brain structures. *Nature* 520(7546): 224–229 (cit. on pp. 156, 177).

Hirohata, S., Wang, L. W., Miyagi, M., Yan, L., Seldin, M. F., Keene, D. R., Crabb, J. W., & Apte, S. S. (2002) Punctin, a novel ADAMTS-like molecule, ADAMTSL-1, in extracellular matrix. *The Journal of Biological Chemistry* 277(14): 12182–9 (cit. on p. 186).

Hirokawa, M., Morita, H., Tajima, T., Takahashi, A., Ashikawa, K., Miya, F., Shigemizu, D., Ozaki, K., Sakata, Y., Nakatani, D., Suna, S., Imai, Y., Tanaka, T., Tsunoda, T., Matsuda, K., Kadowaki, T., Nakamura, Y., Nagai, R., Komuro, I., & Kubo, M. (2015) A genome-wide association study identifies PLCL2 and AP3D1-DOT1L-SF3A2 as new susceptibility loci for myocardial infarction in Japanese. *European Journal of Human Genetics* 23(3): 374–380 (cit. on p. 189).

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002) A comprehensive review of genetic association studies. *Genetics in Medicine* 4(2): 45–61 (cit. on p. 35).

Ho, A. J., Stein, J. L., Hua, X., Lee, S., Hibar, D. P., Leow, A. D., Dinov, I. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M. J., Craig, D. W., Gerber, J. D., Allen, A. N., Corneveaux, J. J., Stephan, D. A., DeCarli, C. S., DeChairo, B. M., & al. (2010) A commonly carried allele of the obesity-related FTO gene is associated with reduced brain volume in the healthy elderly. *Proceedings of the National Academy of Sciences of the United States of America* 107(18): 8404–9 (cit. on p. 156).

Hoffman, J. I. E. (2005) "Congenital Heart Disease". *Essential Cardiology*. Totowa, NJ: Humana Press: 393–406 (cit. on p. 67).

Hoggart, C. J., Chadeau-Hyam, M., Clark, T. G., Lampariello, R., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2007) Sequence-Level Population Simulations Over Large Genomic Regions. *Genetics* 177(3): 1725–1731 (cit. on p. 72).

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., & Marchini, J. (2016) Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics* 48(9): 1094–1100 (cit. on p. 195).

Horn, R. A. & Johnson, C. R. (1985) *Matrix analysis*. 23rd ed. New York: Cambridge University Press: 561 (cit. on p. 214).

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24(6): 417–441 (cit. on pp. 125, 134).

Howie, B. N., Donnelly, P., & Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* 5(6). Ed. by N. J. Schork: e1000529 (cit. on pp. 37, 159).

Howie, B., Marchini, J., & Stephens, M. (2011) Genotype imputation with thousands of genomes. *G3* 1(6): 457–70 (cit. on p. 160).

Hua, S. S. T., Hernlem, B. J., Yokoyama, W., & Sarreal, S. B. L. (2015) Intracellular trehalose and sorbitol synergistically promoting cell viability of a biocontrol yeast, Pichia anomala, for aflatoxin reduction. *World Journal of Microbiology and Biotechnology* 31(5): 729–734 (cit. on p. 120).

Hubmacher, D. & Apte, S. S. (2015) ADAMTS proteins as modulators of microfibril formation and function. *Matrix Biology* 47: 34–43 (cit. on p. 189).

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2): 337–338 (cit. on p. 72).

Hughes, S. E. (2004) The pathology of hypertrophic cardiomyopathy. *Histopathology* 44(5): 412–427 (cit. on p. 155).

Huisman, S. M. H., van Lew, B., Mahfouz, A., Pezzotti, N., Höllt, T., Michielsen, L., Vilanova, A., Reinders, M. J., & Lelieveldt, B. P. F. (2017) BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome. *Nucleic Acids Research* 45(10): e83 (cit. on p. 132).

Hunkapiller, T., Kaiser, R., Koop, B., & Hood, L. (1991) Large-scale and automated DNA sequence determination. *Science* 254(5028) (cit. on p. 32).

Hunter, D. J. (2005) Gene–environment interactions in human diseases. *Nature Reviews Genetics* 6(4): 287–298 (cit. on p. 35).

Hyvärinen, A. & Oja, E. (2000) Independent component analysis: algorithms and applications. *Neural Networks* 13(4): 411–430 (cit. on pp. 128, 134).

Ichida, F., Tsubata, S., Bowles, K. R., Haneda, N., Uese, K., Miyawaki, T., Dreyer, W. J., Messina, J., Li, H., Bowles, N. E., & Towbin, J. A. (2001) Novel Gene Mutations in Patients With Left Ventricular Noncompaction or Barth Syndrome. *Circulation* 103(9) (cit. on p. 180).

Ingram, V. M. & Stretton, A. O. W. (1959) Genetic Basis of the Thalassæmia Diseases. *Nature* 184(4703): 1903–1909 (cit. on p. 34).

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860–921 (cit. on p. 33).

Jackson, D. A., Symonst, R. H., & Berg, P. (1972) Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America.* 69(10): 2904–2909 (cit. on p. 32).

Jahanshad, N., Kochunov, P. V., Sprooten, E., Mandl, R. C., Nichols, T. E., Almasy, L., Blangero, J., Brouwer, R. M., Curran, J. E., de Zubicaray, G. I., Duggirala, R., Fox, P. T., Hong, L. E., Landman, B. A., Martin, N. G., McMahon, K. L., Medland, S. E., Mitchell, B. D., Olvera, R. L., Peterson, C. P., & al. (2013) Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the ENIGMA–DTI working group. *NeuroImage* 81: 455–469 (cit. on p. 156).

Jeffreys, A. J. (1979) DNA sequence variants in the G gamma-, A gamma-, delta- and beta-globin genes of man. *Cell* 18(1): 1–10 (cit. on p. 32).

Jenni, R., Wyss, C. A., Oechslin, E. N., & Kaufmann, P. A. (2002) Isolated Ventricular Noncompaction Is Associated With Coronary Microcirculatory Dysfunction. *Journal of the American College of Cardiology* 39: 450–454 (cit. on p. 189).

Jiang, C. & Zeng, Z. B. (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140(3): 1111–27 (cit. on pp. 35, 56, 59, 97).

Johannsen, W. (1911) The Genotype Conception of Heredity. *The American Naturalist* 45: 129–159 (cit. on p. 29).

Junga, G., Kneifel, S., Smekal, A. V., Steinert, H., & Bauersfeld, U. (1999) Myocardial ischaemia in children with isolated ventricular non-compaction. *European Heart Journal* 20: 910–916 (cit. on p. 189).

Kan, Y. W. & Dozy, A. M. (1978) Antenatal diagnosis of sickle-cell anaemia by D.N.A. analysis of amniotic-fluid cells. *Lancet* 2(8096): 910–2 (cit. on p. 32).

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* 42(4): 348–54 (cit. on pp. 38, 50, 88, 89, 103, 193).

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178(3): 1709–23 (cit. on pp. 52, 53).

Kannel, W. B. & McGee, D. L. (1979) Diabetes and cardiovascular disease. The Framingham study. *JAMA* 241(19): 2035–8 (cit. on p. 68).

Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törönen, P., & Castrén, E. (2003) Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics* 4(1): 48 (cit. on p. 137).

Kathiresan, S. & Srivastava, D. (2012) Genetics of Human Cardiovascular Disease. *Cell* 148(6): 1242–1257 (cit. on p. 68).

Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., Voight, B. F., Bonnycastle, L. L., Jackson, A. U., Crawford, G., Surti, A., Guiducci, C., Burtt, N. P., Parish, S., Clarke, R., Zelenika, D., & al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics* 41(1): 56–65 (cit. on p. 189).

Kato, N., Takeuchi, F., Tabara, Y., Kelly, T. N., Go, M. J., Sim, X., Tay, W. T., Chen, C.-H., Zhang, Y., Yamamoto, K., Katsuya, T., Yokota, M., Kim, Y. J., Ong, R. T. H., Nabika, T., Gu, D., Chang, L.-c., Kokubo, Y., Huang, W., Ohnaka, K., & al. (2011) Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nature Genetics* 43(6): 531–538 (cit. on p. 38).

Kawel, N., Nacif, M., Arai, A. E., Gomes, A. S., Hundley, W. G., Johnson, W. C., Prince, M. R., Stacey, R. B., Lima, J. A. C., & Bluemke, D. A. (2012) Trabeculated (Noncompacted) and Compact Myocardium in AdultsClinical Perspective. *Circulation: Cardiovascular Imaging* 5(3) (cit. on pp. 180, 183, 192).

Kayo, O. (2006) *Locally Linear Embedding Algorithm Extensions and Applications* (cit. on pp. 133, 142).

Kelleher, J., Etheridge, A. M., & McVean, G. (2016) Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology* 12(5). Ed. by Y. S. Song: e1004842 (cit. on p. 72).

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M., & Tsui, L. C. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* 245(4922): 1073–80 (cit. on p. 34).

Keynes, M. & Cox, T. M. (2008) William Bateson, the rediscoverer of Mendel. *Journal of the Royal Society of Medicine* 101(3): 104 (cit. on p. 26).

Kimura, A., Harada, H., Park, J.-E., Nishi, H., Satoh, M., Takahashi, M., Hiroi, S., Sasaoka, T., Ohbuchi, N., Nakamura, T., Koyanagi, T., Hwang, T.-H., Choo, J.-A., Chung, K.-S., Hasegawa, A., Nagai, R., Okazaki, O., Nakamura, H., Matsuzaki, M., Sakamoto, T., & al. (1997) Mutations in the cardiac troponin I gene associated with hypertrophic cardiomyopathy. *Nature Genetics* 16(4): 379–382 (cit. on p. 68).

Klaassen, S., Probst, S., Oechslin, E., Gerull, B., Krings, G., Schuler, P., Greutmann, M., Hürlimann, D., Yegitbasi, M., Pons, L., Gramlich, M., Drenckhahn, J., Heuser, A., Berger, F., Jenni, R., & Thierfelder, L. (2008) Mutations in Sarcomere Protein Genes in Left Ventricular Noncompaction. *Circulation* 117(22) (cit. on p. 180).

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* 308(5720) (cit. on p. 37).

Knott, S. A. & Haley, C. S. (2000) Multitrait Least Squares for Quantitative Trait Loci Detection. *Genetics* 156: 899–911 (cit. on p. 56).

Korol, A. B., Ronin, Y. I., Itskovich, A. M., Peng, J., & Nevo, E. (2001) Enhanced Efficiency of Quantitative Trait Loci Mapping Analysis Based on Multivariate Complexes of Quantitative Traits. *Genetics* 157: 1789–1803 (cit. on p. 56).

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., & Nordborg, M. (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. en. *Nature Genetics* 44(9): 1066–71 (cit. on pp. 38, 56, 59, 89, 97, 118, 193).

Krol, K., Brozda, I., Skoneczny, M., Bretne, M., Skoneczna, A., & Yoshida, S. (2015) A Genomic Screen Revealing the Importance of Vesicular Trafficking Pathways in Genome Maintenance and Protection against Genotoxic Stress in Diploid Saccharomyces cerevisiae Cells. *PLoS ONE* 10(3). Ed. by M. S.-Y. Huen: e0120702 (cit. on p. 120).

Kruskal, J. B. (1964a) Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika* 29(1): 1–27 (cit. on p. 128).

Kruskal, J. B. (1964b) Nonmetric multidimensional scaling: A numerical method. *Psychometrika* 29(2): 115–129 (cit. on p. 128).

Krzywinski, M. & Altman, N. (2013a) Points of significance: Power and sample size. *Nature Methods* 10(12): 1139–1140 (cit. on pp. 43, 44).

Krzywinski, M. & Altman, N. (2013b) Points of significance: Significance, P values and t-tests. *Nature Methods* 10(11): 1041–1042 (cit. on p. 44).

Kulbrock, M., Lehner, S., Metzger, J., Ohnesorge, B., & Distl, O. (2013) A genome-wide association study identifies risk loci to equine recurrent uveitis in German warmblood horses. *PloS ONE* 8(8): e71619 (cit. on p. 49).

Kullback, S. & Leibler, R. A. (1951) On Information and Sufficiency. *The Annals of Mathematical Statistics* 22(1): 79–86 (cit. on p. 130).

Lafon, S. & Lee, A. (2006) Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parameterization. *EEE Trans. Pattern Anal. and Mach. Intel* 28: 1393–1403 (cit. on p. 134).

Lander, E. S. (1996) The new genomics: global views of biology. *Science* 274(5287): 536–9 (cit. on p. 36).

Lander, E. S. & Schork, N. J. (1994) Genetic dissection of complex traits. *Science* 265(5181): 2037–48 (cit. on pp. 44, 48).

Lango, A. H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., Ferreira, T., Wood, A. R., Weyant, R. J., Segrè, A. V., Speliotes, E. K., Wheeler, E., Soranzo, N., Park, J.-H., Yang, J., Gudbjartsson, D., & al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467(7317): 832–838 (cit. on p. 38).

Laparra, V., Malo, J., & Camps-Valls, G. (2015) Dimensionality Reduction via Regression in Hyperspectral Imagery. *IEEE Journal of Selected Topics in Signal Processing* 9(6): 1026–1036 (cit. on pp. 128, 129, 134).

Laske, T. G. & Iaizzo, P. A. (2005) The Cardiac Conduction System. *Handbook of Cardiac Anatomy, Physiology, and Devices*. Ed. by P. A. Iaizzo. Totowa, NJ: Humana Press: 123–136 (cit. on p. 64).

Ledoit, O. & Wolf, M. (2004) A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88(2): 365–411 (cit. on p. 104).

Lee, J. A. & Verleysen, M. (2009) Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing* 72(7): 1431–1443 (cit. on p. 137).

Lee, S., Yang, J., Goddard, M., Visscher, P., & Wray, N. (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 28(19): 2540–2542 (cit. on p. 88).

Lee, S., Goddard, M. E., Visscher, P. M., & van der Werf, J. H. (2010) Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. *Genetics Selection Evolution* 42(1): 22 (cit. on p. 52).

Lee, J.-Y., Lee, B.-S., Shin, D.-J., Woo Park, K., Shin, Y.-A., Joong Kim, K., Heo, L., Young Lee, J., Kyoung Kim, Y., Jin Kim, Y., Bum Hong, C., Lee, S.-H., Yoon, D., Jung Ku, H., Oh, I.-Y., Kim, B.-J., Lee, J., Park, S.-J., Kim, J., Kawk, H.-k., & al. (2013) A genome-wide association study of a coronary artery disease risk variant. *Journal of Human Genetics* 58(3): 120–126 (cit. on p. 189).

Lettre, G., Lange, C., & Hirschhorn, J. N. (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology* 31(4): 358–362 (cit. on p. 42).

Levy, D., Larson, M. G., Benjamin, E. J., Newton-Cheh, C., Wang, T. J., Hwang, S.-J., Vasan, R. S., & Mitchell, G. F. (2007) Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC Medical Genetics* 8(Suppl 1): S3 (cit. on p. 69).

Lewontin, R. C. (1970) The Units of Selection. *Annual Review of Ecology and Systematics* 1: 1–18 (cit. on p. 24).

Lewontin, R. C. & Kojima, K.-i. (1960) The Evolutionary Dynamics of Complex Polymorphisms. *Evolution* 14(4): 458 (cit. on p. 36).

Li, D., Deogun, J., Spaulding, W., & Shuart, B. (2004) "Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method". *Rough Sets and Current Trends in Computing*. Springer, Berlin, Heidelberg: 573–579 (cit. on pp. 49, 107).

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* 34(8): 816–834 (cit. on p. 37).

Lila, E., Aston, J. A. D., & Sangalli, L. M. (2016) Smooth Principal Component Analysis over two-dimensional manifolds with an application to Neuroimaging. arXiv: 1601.03670 (cit. on p. 195).

Lindgren, C. M., Heid, I. M., Randall, J. C., Lamina, C., Steinthorsdottir, V., Qi, L., Speliotes, E. K., Thorleifsson, G., Willer, C. J., Herrera, B. M., Jackson, A. U., Lim, N., Scheet, P., Soranzo, N., Amin, N., Aulchenko, Y. S., Chambers, J. C., Drong, A., Luan, J., Lyon, H. N., & al. (2009) Genome-Wide Association Scan Meta-Analysis Identifies Three Loci Influencing Adiposity and Fat Distribution. *PLoS Genetics* 5(6). Ed. by D. B. Allison: e1000508 (cit. on p. 38).

Lippert, C., Casale, F. P., Rakitsch, B., & Stegle, O. (2014) LIMIX: genetic analysis of multiple traits. en. *bioRxiv*: 003905 (cit. on pp. 57, 90, 103, 194).

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011) FaST linear mixed models for genome-wide association studies. *Nature Methods* 8(10): 833–837 (cit. on pp. 38, 51, 59, 89).

Lippert, C., Quon, G., Kang, E. Y., Kadie, C. M., Listgarten, J., & Heckerman, D. (2013) The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. en. *Scientific Reports* 3: 1815 (cit. on pp. 53, 72).

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J. T., van Oudenaarden, A., Barton, D. B. H., Bailes, E., Nguyen, A. N., Jones, M., Quail, M. A., & al. (2009) Population genomics of domestic and wild yeasts. *Nature* 458(7236): 337 (cit. on p. 117).

Little, R. J. A. (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Source Journal of the American Statistical Association* 83(404): 1198–1202 (cit. on pp. 107, 110, 113, 114).

Little, R. J. A. & Rubin, D. B. (2002) *Statistical analysis with missing data*. Ed. by D. J. Balding, P. Bloomfield, N. A. C. Cressie, N. I. Fisher, I. M. Johnstone, J. B. Kadane, L. M. Ryan, D. W. Scott, A. F. M. Smith, & J. L. Teugels. 2nd. New Jersey: John Wiley & Sons, Inc: 408 (cit. on pp. 106–108).

Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., Wollstein, A., Lao, O., de Bruijne, M., Ikram, M. A., van der Lugt, A., Rivadeneira, F., Uitterlinden, A. G., Hofman, A., Niessen, W. J., Homuth, G., de Zubicaray, G., McMahon, K. L., Thompson, P. M., Daboul, A., & al. (2012) A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. *PLoS Genetics* 8(9). Ed. by G. Gibson: e1002932 (cit. on pp. 54, 141, 144, 177, 194).

Liu, Y., Athanasiadis, G., & Weale, M. E. (2008) A survey of genetic simulation software for population and epidemiological studies. *Human Genomics* 3(1): 79 (cit. on p. 72).

Lock, R. H. (1906) *Recent progress in the study of variation, heredity, and evolution*. London: J. Murray: 352 (cit. on pp. 26, 28).

Loh, P.-r., Tucker, G., Bulik-sullivan, B. K., & Vilhj, B. J. (2014) Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics* 47(3): 1–79 (cit. on pp. 71, 72, 78, 89).

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., & Hirschhorn, J. N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* 33(2): 177–182 (cit. on p. 35).

Lopes, M. S., Silva, F. F., Harlizius, B., Duijvesteijn, N., Lopes, P. S., Guimarães, S. E., & Knol, E. F. (2013) Improved estimation of inbreeding and kinship in pigs using optimized SNP panels. *BMC Genetics* 14(1): 92 (cit. on p. 52).

Lorell, B. H. & Carabello, B. A. (2000) Left ventricular hypertrophy: pathogenesis, detection, and prognosis. *Circulation* 102(4) (cit. on pp. 67, 155).

Lu, X., Wang, L., Chen, S., He, L., Yang, X., Shi, Y., Cheng, J., Zhang, L., Gu, C. C., Huang, J., Wu, T., Ma, Y., Li, J., Cao, J., Chen, J., Ge, D., Fan, Z., Li, Y., Zhao, L., Li, H., & al. (2012) Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. *Nature Genetics* 44(8): 890–894 (cit. on p. 38).

Lynch, M. & Ritland, K. (1999) Estimation of pairwise relatedness with molecular markers. *Genetics* 152(4): 1753–66 (cit. on p. 49).

Maaten, L. V. D. & Hinton, G. (2008) Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9: 2579–2605 (cit. on pp. 131, 134).

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* 45(D1): D896–D901 (cit. on pp. 38, 68, 171, 190).

Mackay, J., Mensah, G., & A, O. (2004) *The Atlas of Heart Disease And Stroke*. Ed. by A. Haley. 1st ed. World Health Organisation (cit. on p. 66).

Malosetti, M., van der Linden, C. G., Vosman, B., & van Eeuwijk, F. A. (2007) A Mixed-Model Approach to Association Mapping Using Pedigree Information With an Illustration of Resistance to Phytophthora infestans in Potato. *Genetics* 175(2): 879–889 (cit. on p. 50).

Marchini, J., Cardon, L. R., Phillips, M. S., & Donnelly, P. (2004) The effects of human population structure on large genetic association studies. *Nature Genetics* 36(5): 512–7 (cit. on pp. 44, 47).

Marchini, J. & Howie, B. (2010) Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* 11(7): 499–511 (cit. on pp. 37, 159).

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* 39(7): 906–13 (cit. on pp. 76, 158).

Marigorta, U. M. & Gibson, G. (2014) A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects. *Frontiers in Genetics* 5(July): 225 (cit. on pp. 71, 78).

Martinez-Jimenez, C. P., Eling, N., Chen, H.-C., Vallejos, C. A., Kolodziejczyk, A. A., Connor, F., Stojic, L., Rayner, T. F., Stubbington, M. J. T., Teichmann, S. A., de la Roche, M., Marioni, J. C., & Odom, D. T. (2017) Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* 355(6332): 1433–1436 (cit. on p. 132).

Matthaei, H. J., Jones, O. W., Martig, R. G., & W, N. M. (1962) Characteristics and composition of RNA coding units. *Proceedings of the National Academy of Sciences of the United States of America* 48: 666–677 (cit. on p. 32).

Maxam, A. M. & Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* 74(2): 560–4 (cit. on p. 32).

Meder, B., Katus, H. A., & Keller, A. (2016) Computational Cardiology - A New Discipline of Translational Research. *Genomics, Proteomics & Bioinformatics* 14(4): 177–8 (cit. on p. 195).

Mendel, G. (1866) Versuche über Plflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865*: 3–47 (cit. on p. 26).

Mendel, G. (1869) Ueber einige aus künstlicher Befruchtung gewonnenen Hieracium-Bastarde. *Verhandlungen des naturforschenden Vereines in Brünn* 8: 26–31 (cit. on p. 26).

Meyer, H. V. (2017) *R Package: PhenotypeSimulator: Flexible Phenotype Simulation from Different Genetic and Noise Models*. Cambridge (cit. on pp. 72, 193).

Meyer, H. V. & Birney, E. (2018) PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics* (cit. on p. 69).

Meyer, H. V., Casale, F. P., Stegle, O., & Birney, E. (2018) LiMMBo: a simple, scalable approach for linear mixed models in high-dimensional genetic association studies. *bioRxiv*: 255497 (cit. on pp. 69, 89).

Miescher, F. (1871) Ueber die chemische Zusammensetzung der Eiterzellen. *Medizinisch-chemische Untersuchungen* 4: 441–460 (cit. on p. 27).

Minchin, P. R. (1987) An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio* 69: 89–107 (cit. on p. 128).

Mitchell, L. E., Agopian, a. J., Bhalla, a., Glessner, J. T., Kim, C. E., Swartz, M. D., Hakonarson, H., & Goldmuntz, E. (2015) Genome-wide association study of maternal and inherited effects on left-sided cardiac malformations. *Human Molecular Genetics* 24(1): 265–273 (cit. on p. 69).

Mitchell, S. C., Korones, S. B., & Berendes, H. W. (1971) Congenital heart disease in 56,109 births incidence and natural history. *Circulation* 43(3) (cit. on p. 67).

Monaghan, F. V. & Corcos, A. F. (1986) Tschermak: a non-discoverer of Mendelism: I. An historical note. *Journal of Heredity* 77(6): 468–469 (cit. on p. 26).

Monaghan, F. V. & Corcos, A. F. (1987) Tschermak: a non-discoverer of Mendelism II. A critique. *Journal of Heredity* 78(3): 208–210 (cit. on p. 26).

Monserrat, L., Hermida-Prieto, M., Fernandez, X., Rodriguez, I., Dumont, C., Cazon, L., Cuesta, M. G., Gonzalez-Juanatey, C., Peteiro, J., Alvarez, N., Penas-Lado, M., & Castro-Beiras, A. (2007) Mutation in the alpha-cardiac actin gene associated with apical hypertrophic cardiomyopathy, left ventricular non-compaction, and septal defects. *European Heart Journal* 28(16): 1953–1961 (cit. on p. 180).

Moorman, A. F. M. & Lamers, W. H. (1994) Molecular anatomy of the developing heart. *Trends in Cardiovascular Medicine* 4(6): 257–264 (cit. on p. 64).

Morgan, T. H. (1910) Sex limited inheritance in Drosophila. *Science* 32(812): 120–2 (cit. on p. 29).

Morgan, T. H. (1911a) An attempt to analyze the constitution of the chromosomes on the basis of sex-limited inheritance in Drosophila. *Journal of Experimental Zoology* 11(4): 365–413 (cit. on p. 29).

Morgan, T. H. (1911b) Random segregation versus coupling in Mendelian inheritance. *Science* 34(873) (cit. on p. 29).

Morgan, T. H., Sturtevant, A. H., Muller, H. J., & Bridges, C. B. (1915) *The mechanism of Mendelian heredity*. New York: H. Holt & company: 288 (cit. on p. 30).

Moric-Janiszewska, E. & Markiewicz-Łoskot, G. (2008) Genetic Heterogeneity of Left-ventricular Noncompaction Cardiomyopathy. *Clinical Cardiology* 31(5): 201–204 (cit. on p. 180).

Morris, J. A., Randall, J. C., Maller, J. B., & Barrett, J. C. (2010) Evoker: A visualization tool for genotype intensity data. *Bioinformatics* 26(14): 1786–1787 (cit. on p. 178).

Morrow, J. F. & Berg, P. (1972) Cleavage of Simian virus 40 DNA at a unique site by a bacterial restriction enzyme. *Proceedings of the National Academy of Sciences of the United States of America* 69(11): 3365–9 (cit. on p. 32).

Morton, N. E. (1955) Sequential tests for the detection of linkage. *The American Journal of Human Genetics* 7(3): 277–318 (cit. on p. 34).

Mysliwiec, M. R., Bresnick, E. H., & Lee, Y. (2011) Endothelial Jarid2/Jumonji is required for normal cardiac development and proper Notch1 expression. *The Journal of Biological Chemistry* 286(19): 17193–204 (cit. on p. 180).

Nakayama, M., Nakajima, D., Nagase, T., Nomura, N., Seki, N., & Ohara, O. (1998) Identification of high-molecular-weight proteins with multiple EGF-like motifs by motif-trap screening. *Genomics* 51(1): 27–34 (cit. on p. 170).

Naylor, M. G., Lin, X., Weiss, S. T., Raby, B. A., & Lange, C. (2010) Using Canonical Correlation Analysis to Discover Genetic Regulatory Variants. *PLoS ONE* 5(5). Ed. by A. C. Goldberg: e10395 (cit. on p. 55).

Nejati-Javaremi, A., Smith, C., & Gibson, J. P. (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. *Journal of Animal Science* 75(7): 1738–45 (cit. on p. 52).

Newton-Cheh, C., Guo, C.-Y., Wang, T. J., O'donnell, C. J., Levy, D., & Larson, M. G. (2007) Genome-wide association study of electrocardiographic and heart rate variability traits: the Framingham Heart Study. *BMC Medical Genetics* 8 Suppl 1: S7 (cit. on p. 69).

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., Webb, T. R., Zeng, L., Dehghan, A., Alver, M., Armasu, S. M., Auro, K., Bjonnes, A., Chasman, D. I., Chen, S., Ford, I., & al. (2015) A comprehensive 1,000 Genomes-based genome-wide association

meta-analysis of coronary artery disease. *Nature Genetics* 47(10): 1121–30 (cit. on pp. 69, 189).

Nirenberg, M. W. & Matthaei, J. H. (1961) The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America* 47(10): 1588–602 (cit. on p. 32).

Noguchi, E., Sakamoto, H., Hirota, T., Ochiai, K., Imoto, Y., Sakashita, M., Kurosaka, F., Akasawa, A., Yoshihara, S., Kanno, N., Yamada, Y., Shimojo, N., Kohno, Y., Suzuki, Y., Kang, M.-J., Kwon, J.-W., Hong, S.-J., Inoue, K., Goto, Y.-i., Yamashita, F., & al. (2011) Genome-Wide Association Study Identifies HLA-DP as a Susceptibility Gene for Pediatric Asthma in Asian Populations. *PLoS Genetics* 7(7). Ed. by M. I. McCarthy: e1002170 (cit. on p. 38).

O'Brien, P. C. (1984) Procedures for Comparing Samples with Multiple Endpoints. *Biometrics* 40(40): 1079–1087 (cit. on p. 55).

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, M.-R., & Coin, L. J. M. (2012) MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS one* 7(5): e34861 (cit. on p. 76).

O'Toole, T. E., Conklin, D. J., & Bhatnagar, A. (2008) Environmental risk factors for heart disease. *Reviews on Environmental Health* 23(3): 167–202 (cit. on pp. 67, 155).

Oechslin, E. N., Attenhofer Jost, C. H., Rojas, J. R., Kaufmann, P. A., & Jenni, R. (2000) Long-term follow-up of 34 adults with isolated left ventricular noncompaction: a distinct cardiomyopathy with poor prognosis. *Journal of the American College of Cardiology* 36(2): 493–500 (cit. on p. 189).

Oliveira, A. & Seijas-Macias, A. (2012) An Approach to Distribution of the Product of Two Normal Variables. *Discussiones Mathematicae: Probability and Statistics* 32(1-2): 87 (cit. on p. 79).

Ott, J., Wang, J., & Leal, S. M. (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics* 16(5): 275–284 (cit. on p. 36).

Paige, S. L., Plonowska, K., Xu, A., & Wu, S. M. (2015) Molecular regulation of cardiomyocyte differentiation. *Circulation Research* 116(2): 341–53 (cit. on p. 66).

Park, D. & Fishman, G. (2017) Development and Function of the Cardiac Conduction System in Health and Disease. *Journal of Cardiovascular Development and Disease* 4(2): 7 (cit. on p. 66).

Parkhomenko, E., Tritchler, D., & Beyene, J. (2009) Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. *Statistical Applications in Genetics and Molecular Biology* 8(1) (cit. on p. 55).

Paternoster, L., Zhurov, A. I., Toma, A. M., Kemp, J. P., St. Pourcain, B., Timpson, N. J., McMahon, G., McArdle, W., Ring, S. M., Smith, G. D., Richmond, S., & Evans, D. M. (2012) Genome-wide Association Study of Three-Dimensional Facial Morphology Identifies a Variant in PAX3 Associated with Nasion Position. *The American Journal of Human Genetics* 90(3): 478–485 (cit. on p. 38).

Patterson, H. D. & Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58(3): 545–554 (cit. on p. 41).

Patterson, N., Price, A. L., Reich, D., Reich, D., & Daly, M. (2006) Population Structure and Eigenanalysis. *PLoS Genetics* 2(12): e190 (cit. on pp. 53, 169).

Pausova, Z., Paus, T., Abrahamowicz, M., Almerigi, J., Arbour, N., Bernard, M., Gaudet, D., Hanzalek, P., Hamet, P., Evans, A. C., Kramer, M., Laberge, L., Leal, S. M., Leonard, G., Lerner, J., Lerner, R. M., Mathieu, J., Perron, M., Pike, B., Pitiot, A., & al. (2007) Genes, maternal smoking, and the offspring brain and body during adolescence: Design of the Saguenay Youth Study. *Human Brain Mapping* 28(6): 502–518 (cit. on p. 141).

Payne, R. M., Johnson, M. C., Grant, J. W., & Strauss, A. W. (1995) Toward a Molecular Understanding of Congenital Heart Disease. *Circulation* 91(2): 494–504 (cit. on p. 155).

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* 50(302): 157–175 (cit. on p. 27).

Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2: 559–572 (cit. on pp. 27, 125).

Peng, B., Amos, C. I., & Kimmel, M. (2007) Forward-Time Simulations of Human Populations with Complex Diseases. *PLoS Genetics* 3(3): e47 (cit. on p. 72).

Penrose, L. S. (1935) The detection of autosomal linkage in data which consist of airs of brothers and sisters of unspecified parentage. *Annals of Eugenics* 6(2): 133–138 (cit. on pp. 33, 34).

Petersen, S. E., Selvanayagam, J. B., Wiesmann, F., Robson, M. D., Francis, J. M., Anderson, R. H., Watkins, H., & Neubauer, S. (2005) Left Ventricular Non-Compaction. *Journal of the American College of Cardiology* 46(1): 101–105 (cit. on p. 180).

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., & Hinds, D. A. (2016) Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics* advance on (cit. on p. 38).

Piernick, L. K. (& Correns, C. ((1950) G. Mendel's law concerning behavior of progeny of varietal hybrids. *Genetics* 35(5): 33–41 (cit. on p. 26).

Pigott, T. D. (2001) A Review of Methods for Missing Data. *Educational Research and Evaluation* 7(4): 353–383 (cit. on p. 113).

Plate, C. (1910) "Die Schwanzknickblastovariation". *Festschrift für Richard Hertwig, Zweiter Band*. Jena: Gustav Fischer. Chap. Vererbungs: 537–610 (cit. on p. 29).

Porter, H. F. & O'Reilly, P. F. (2017) Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Scientific Reports* 7: 38837 (cit. on pp. 76, 78).

Post, W. S., Larson, M. G., Myers, R. H., Galderisi, M., & Levy, D. (1997) Heritability of Left Ventricular Mass : The Framingham Heart Study. *Hypertension* 30(5): 1025–1028 (cit. on p. 155).

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8): 904–909 (cit. on pp. 49, 169).

Pritchard, J. K., Stephens, M., Rosenberg, N. A., & Donnelly, P. (2000) Association mapping in structured populations. *The American Journal of Human Genetics* 67(1): 170–81 (cit. on p. 49).

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R., & Willer, C. J. (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26(18): 2336–2337 (cit. on pp. 172, 188).

Pulst, S. M. (1999) Genetic Linkage Analysis. *Archives of Neurology* 56(6): 667 (cit. on p. 34).

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81(3): 559–75 (cit. on p. 158).

Rakitsch, B., Lippert, C., Borgwardt, K., & Stegle, O. (2013) *It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals* (cit. on p. 90).

Reich, D. E. & Goldstein, D. B. (2001) Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology* 20(1): 4–16 (cit. on pp. 35, 36, 48, 104).

Ripley, B. D. (1996) *Pattern recognition and neural networks*. 7th. Cambridge: Cambridge University Press: 416 (cit. on p. 134).

Risch, N. & Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273(5281): 1516–7 (cit. on p. 35).

Ritland, K. (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology* 9(9): 1195–204 (cit. on p. 49).

Ritter, M., Oechslin, E., Sütsch, G., Attenhofer, C., Schneider, J., & Jenni, R. (1997) Isolated noncompaction of the myocardium in adults. *Mayo Clinic proceedings* 72(1): 26–31 (cit. on p. 189).

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., & Feldman, M. W. (2005) Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLoS Genetics* 1(6): e70 (cit. on p. 49).

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002) Genetic Structure of Human Populations. *Science* 298(5602) (cit. on p. 49).

Roweis, S. T. & Saul, L. K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500): 2323–2326 (cit. on pp. 129, 142).

Rubin, D. B. (1976) Inference and missing data. *Biometrika* 63(3): 581–92 (cit. on pp. 106, 110).

Rubin, D. B. (1987) *Multiple Imputation for nonresponse in surveys*. 2nd. New York: John Wiley & Sons, Inc (cit. on pp. 108, 113).

Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., Sovio, U., Ruokonen, A., Laitinen, J., Jakkula, E., Coin, L., Hoggart, C., Collins, A., Turunen, H., Gabriel, S., Elliot, P., & al. (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics* 41(1): 35–46 (cit. on p. 186).

Sanger, F., Nicklen, S., & Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74(12): 5463–5467 (cit. on p. 32).

Sankaranarayanan, K. (1998) Ionizing radiation and genetic risks: IX. Estimates of the frequencies of mendelian diseases and spontaneous mutation rates in human populations: a 1998 perspective. *Mutation Research* 411(2): 129–178 (cit. on p. 35).

Sano, M., Kamitsuji, S., Kamatani, N., Tabara, Y., Kawaguchi, T., Matsuda, F., Yamagishi, H., Fukuda, K., & (JPDSC), J. P. D. S. C. (2016) Genome-Wide Association Study of Absolute QRS Voltage Identifies Common Variants of TBX3 as Genetic Determinants of Left Ventricular Mass in a Healthy Japanese Population. *PLoS ONE* 11(5). Ed. by T. Minamino: e0155550 (cit. on p. 155).

Sanoudou, D., Vafiadaki, E., Arvanitis, D. a., Kranias, E., & Kontrogianni-Konstantopoulos, A. (2005) Array lessons from the heart: focus on the genome and transcriptome of cardiomyopathies. *Physiological genomics* 21(2): 131–143 (cit. on p. 155).

Sarkar, S. K. (2007) Stepup procedures and controlling generalized FWER and generalized FDR. *The Annals of Statistics* 35(6): 2405–2420. arXiv: `arXiv:0803.2934v1` (cit. on p. 46).

Schafer, J. L. (1997) *Analysis of incomplete multivariate data*. Chapman & Hall/CRC (cit. on pp. 108, 114).

Schafer, J. L. & Graham, J. W. (2002) Missing data: Our view of the state of the art. *Psychological Methods* 7(2): 147–177 (cit. on p. 110).

Schäfer, J. & Strimmer, K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* 4: Article32 (cit. on pp. 55, 104).

Schoelkopf, B., Smola, A., & Uller, K.-R. (1998) Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* 10: 1299–1319 (cit. on pp. 126, 134).

Schor, I. E., Degner, J. F., Harnett, D., Cannavò, E., Casale, F. P., Shim, H., Garfield, D. A., Birney, E., Stephens, M., Stegle, O., & M Furlong, E. E. (2017) Promoter shape varies across populations and affects promoter evolution and expression noise. *Nature Publishing Group* 49 (cit. on p. 90).

Schott, J.-J., Benson, D. W., Basson, C. T., Pease, W., Silberbach, G. M., Moak, J. P., Maron, B. J., Seidman, C. E., & Seidman, J. G. (1998) Congenital Heart Disease Caused by Mutations in the Transcription Factor NKX2-5. *Science* 281(5373) (cit. on p. 68).

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F. R., Barbalic, M., Gieger, C., Absher, D., Aherrahrou, Z., Allayee, H., Altshuler, D., Anand, S. S., Andersen, K., Anderson, J. L., Ardissino,

D., Ball, S. G., Balmforth, A. J., & al. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* 43(4): 333–338 (cit. on p. 189).

Scuteri, A., Sanna, S., Chen, W.-M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orrú, M., Usala, G., Dei, M., Lai, S., Maschio, A., Busonero, F., Mulas, A., Ehret, G. B., Fink, A. A., Weder, A. B., Cooper, R. S., Galan, P., & al. (2007) Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits. *PLoS Genetics* 3(7): e115 (cit. on pp. 37, 43).

Seidman, J. G. & Seidman, C. (2001) The genetic basis for cardiomyopathy. *Cell* 104(4): 557–567 (cit. on p. 67).

Serre, D. & Pääbo, S. (2004) Evidence for Gradients of Human Genetic Diversity Within and Among Continents. *Genome Research* 14(9): 1679–1685 (cit. on p. 49).

Shaffer, J. R., Orlova, E., Lee, M. K., Leslie, E. J., Raffensperger, Z. D., Heike, C. L., Cunningham, M. L., Hecht, J. T., Kau, C. H., Nidey, N. L., Moreno, L. M., Wehby, G. L., Murray, J. C., Laurie, C. A., Laurie, C. C., Cole, J., Ferrara, T., Santorico, S., Klein, O., Mio, W., & al. (2016) Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. *PLoS Genetics* 12(8). Ed. by G. S. Barsh: e1006149 (cit. on pp. 54, 177).

Shaffer, J. P. (1995) Multiple Hypothesis Testing. *Annual Review of Psychology* 46: 561–84 (cit. on pp. 44, 46).

Shendure, J. & Ji, H. (2008) Next-generation DNA sequencing. *Nature Biotechnology* 26(10): 1135–1145 (cit. on p. 33).

Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., Walter, K., Menni, C., Chen, L., Vasquez, L., Valdes, A. M., Hyde, C. L., Wang, V., Ziemek, D., Roberts, P., Xi, L., & al. (2014) An atlas of genetic influences on human blood metabolites. *Nature Genetics* 46(6): 543–550 (cit. on p. 105).

Shriner, D. (2012) Moving toward System Genetics through Multiple Trait Analysis in Genome-Wide Association Studies. *Frontiers in Genetics* 3: 1 (cit. on pp. 54, 56).

Sigg, D. C., Iaizzo, P. A., Xiao, Y.-F., & Bin, H. (2010) *Cardiac Electrophysiology Methods and Models*. New York: Springer US: 492 (cit. on p. 64).

Smith, H. O. & Welcox, K. W. (1970) A Restriction enzyme from Hemophilus influenzae. *Journal of Molecular Biology* 51(2): 379–391 (cit. on p. 32).

Soler, R., Rodríguez, E., Monserrat, L., & Alvarez, N. (2002) MRI of subendocardial perfusion deficits in isolated left ventricular noncompaction. *Journal of Computer Assisted Tomography* 26(3): 373–5 (cit. on p. 189).

Song, W. T. (2005) Relationships among some univariate distributions. *IIE Transactions* 37: 651–656 (cit. on p. 79).

Southern, E. M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology* 98(3): 503–17 (cit. on p. 32).

Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., Balding, D. J., & Balding, D. J. (2017) Reevaluation of SNP heritability in complex human traits. *Nature Genetics* 49(7): 986–992 (cit. on p. 54).

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Lango Allen, H., Lindgren, C. M., Luan, J., Mägi, R., Randall, J. C., Vedantam, S., Winkler, T. W., Qi, L., Workalemahu, T., Heid, I. M., Steinthorsdottir, V., Stringham, H. M., Weedon, M. N., Wheeler, E., & al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* 42(11): 937–48 (cit. on p. 38).

Spielman, R. S., McGinnis, R. E., & Ewens, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *The American Journal of Human Genetics* 52(3): 506–16 (cit. on pp. 35, 44, 47).

Staden, R. (1979) A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research* 6(7): 2601–10 (cit. on p. 32).

Stegle, O., Parts, L., Durbin, R., & Winn, J. (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. en. *PLoS Computational Biology* 6(5): e1000770 (cit. on pp. 127, 134).

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols* 7(3): 500–7 (cit. on pp. 127, 133, 195).

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M. J., Craig, D. W., Gerber, J. D., Allen, A. N., Corneveaux, J. J., Dechairo, B. M., Potkin, S. G., Weiner, M. W., & Thompson,

P. (2010) Voxelwise genome-wide association study (vGWAS). *NeuroImage* 53(3): 1160–74 (cit. on pp. 54, 177).

Stein, M. B., Campbell-Sills, L., & Gelernter, J. (2009) Genetic variation in 5HTTLPR is associated with emotional resilience. *American journal of medical genetics. Part B, Neuropsychiatric genetics* 150B(7): 900–6 (cit. on p. 49).

Stephens, M. (2013) A unified framework for association analysis with multiple related phenotypes. *PloS one* 8(7): e65245 (cit. on pp. 71, 78).

Stephens, M., Smith, N. J., & Donnelly, P. (2001) A New Statistical Method for Haplotype Reconstruction from Population Data. *The American Journal of Human Genetics* 68(4): 978–989 (cit. on p. 36).

Sterne, J. A., Smith, G. D., & Cox, D. R. (2001) Sifting the evidence—what's wrong with significance tests? *British Medical Journal* 322(7280): 226 (cit. on p. 43).

Storey, J. D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B* 64(3): 479–498 (cit. on p. 46).

Strittmatter, W. J. & Roses, A. D. (1996) Apolipoprotein E and Alzheimer's Disease. *Annual Review of Neuroscience* 19(1): 53–77 (cit. on p. 35).

Sturtevant, A. H. (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* 14(1): 43–59 (cit. on p. 30).

Su, Z., Marchini, J., & Donnelly, P. (2011) HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* 27(16): 2304–2305 (cit. on pp. 72, 73, 82).

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015) UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine* 12(3): e1001779 (cit. on pp. 37, 38, 104, 190).

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S. A., Palsson, A., Thorleifsson, G., Pálsson, S., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., Vermeulen, S. H., Goldstein, A. M., Tucker, M. A., Kiemeney, L. A., & al. (2008) Two newly identified genetic determinants of pigmentation in Europeans. *Nature Genetics* 40(7): 835–837 (cit. on p. 38).

Sutton, W. S. (1903) The chromosomes in heredity. *Biological Bulletin* 4: 231–251 (cit. on p. 28).

Suzuki, R. & Shimodaira, H. (2006) Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22(12): 1540–1542 (cit. on p. 120).

Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M., & Aulchenko, Y. S. (2012) Rapid variance components-based method for whole-genome association analysis. *Nature Genetics* 44(10): 1166–1170 (cit. on pp. 38, 59, 89).

Swinkels, B. M., Boersma, L. V. A., Rensing, B. J., & Jaarsma, W. (2007) Isolated left ventricular noncompaction in a patient presenting with a subacute myocardial infarction. *Netherlands Heart Journal* 15(3): 109–11 (cit. on p. 189).

Tachmazidou, I., Dedoussis, G., Southam, L., Farmaki, A.-E., Ritchie, G. R. S., Xifara, D. K., Matchan, A., Hatzikotoulas, K., Rayner, N. W., Chen, Y., Pollin, T. I., O'Connell, J. R., Yerges-Armstrong, L. M., Kiagiadaki, C., Panoutsopoulou, K., Schwartzentruber, J., Moutsianas, L., UK10K consortium, E., Tsafantakis, E., Tyler-Smith, C., & al. (2013) A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nature Communications* 4: 2872 (cit. on p. 104).

Takeuchi, F., Yokota, M., Yamamoto, K., Nakashima, E., Katsuya, T., Asano, H., Isono, M., Nabika, T., Sugiyama, T., Fujioka, A., Awata, N., Ohnaka, K., Nakatochi, M., Kitajima, H., Rakugi, H., Nakamura, J., Ohkubo, T., Imai, Y., Shimamoto, K., Yamori, Y., & al. (2012) Genome-wide association study of coronary artery disease in the Japanese. *European Journal of Human Genetics* 20(3): 333–340 (cit. on p. 38).

Tang, H., Quertermous, T., Rodriguez, B., Kardia, S. L. R., Zhu, X., Brown, A., Pankow, J. S., Province, M. A., Hunt, S. C., Boerwinkle, E., Schork, N. J., & Risch, N. J. (2005) Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *The American Journal of Human Genetics* 76(2): 268–75 (cit. on p. 49).

Templ, M., Alfons, A., & Filzmoser, P. (2012) Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification* 6(1): 29–47 (cit. on pp. 107, 110, 111).

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000) A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290(5500): 2319–2323 (cit. on pp. 129, 134).

Teng, S. L. & Huang, H. (2009) A Statistical Framework to Infer Functional Gene Relationships From Biologically Interrelated Microarray Experiments. *Journal of the American Statistical Association* 104(486): 465–473 (cit. on p. 104).

Teo, Y. Y., Inouye, M., Small, K. S., Gwilliam, R., Deloukas, P., Kwiatkowski, D. P., & Clark, T. G. (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* 23(20): 2741–6 (cit. on p. 157).

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437(7063): 1299–320 (cit. on pp. 36, 47, 158, 210).

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(18): 851– (cit. on pp. 36, 210).

The International HapMap Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467 (cit. on p. 36).

Thierfelder, L., Watkins, H., MacRae, C., Lamas, R., McKenna, W., Vosberg, H. P., Seidman, J. G., & Seidman, C. E. (1994) Alpha-tropomyosin and cardiac troponin T mutations cause familial hypertrophic cardiomyopathy: a disease of the sarcomere. *Cell* 77(5): 701–12 (cit. on p. 68).

Thomas, S. C. (2005) The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360(1459): 1457–67 (cit. on p. 49).

Tian, C., Gregersen, P. K., & Seldin, M. F. (2008a) Accounting for ancestry: population substructure and genome-wide association studies. *Human Molecular Genetics* 17(R2): R143–R150 (cit. on p. 48).

Tian, C., Plenge, R. M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., Klareskog, L., Pulver, A. E., Qi, L., Gregersen, P. K., & Seldin, M. F. (2008b) Analysis and Application of European Genetic Substructure Using 300 K SNP Information. *PLoS Genetics* 4(1): e4 (cit. on p. 48).

Toufan, M., Shahvalizadeh, R., & Khalili, M. (2012) Myocardial infarction in a patient with left ventricular noncompaction: a case report. *International Journal of General Medicine* 5: 661–5 (cit. on p. 189).

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17(6): 520–525 (cit. on p. 107).

Tschermak, E. (1900) Ueber künstliche Kreuzung bei Pisum sativum. *Plant Biology* 18(6): 232–239 (cit. on p. 26).

Tuan, D., Biro, P. A., DeRiel, J. K., Lazarus, H., & Forget, B. G. (1979) Restriction endonuclease mapping of the human gamma globin gene loci. *Nucleic Acids Research* 6(7): 2519–44 (cit. on p. 32).

UK10K Consortium (2015) The UK10K project identifies rare variants in health and disease. en. *Nature* 526: 82–90 (cit. on pp. 36, 157, 159).

Unal, B., Critchley, J. A., & Capewell, S. (2004) Explaining the Decline in Coronary Heart Disease Mortality in England and Wales Between 1981 and 2000. *Circulation* 109(9) (cit. on p. 67).

Van Buuren, S. & Oudshoorn, K. (1999) Flexible multivariate imputation by MICE (cit. on p. 113).

Van der Merwe, L., Cloete, R., Revera, M., Heradien, M., Goosen, A., Corfield, V. A., Brink, P. A., & Moolman-Smook, J. C. (2008) Genetic variation in angiotensin-converting enzyme 2 gene is associated with extent of left ventricular hypertrophy in hypertrophic cardiomyopathy. *Human Genetics* 124(1): 57–61 (cit. on pp. 155, 156).

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., & al. (2012) The Human Connectome Project: A data acquisition perspective. *NeuroImage* 62(4): 2222–2231 (cit. on p. 194).

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011) mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45(3): 1–67 (cit. on pp. 107, 113).

Vanovschi, V. (2017) *Parallel Python Software* (cit. on p. 103).

Vasan, R. S., Glazer, N. L., Felix, J. F., Lieb, W., Wild, P. S., Felix, S. B., Watzinger, N., Larson, M. G., Smith, N. L., Dehghan, A., Grosshennig, A., Schillert, A., Teumer, A., Schmidt, R., Kathiresan, S., Lumley, T., Aulchenko, Y. S., König, I. R., Zeller, T., Homuth, G., & al. (2009) Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data. *The Journal of the American Medical Association* 302(2): 168–78 (cit. on pp. 69, 155).

Vasan, R. S., Larson, M. G., Aragam, J., Wang, T. J., Mitchell, G. F., Kathiresan, S., Newton-Cheh, C., Vita, J. A., Keyes, M. J., O'Donnell, C. J., Levy, D., & Benjamin, E. J. (2007) Genome-wide association of echocardiographic dimensions, brachial

artery endothelial function and treadmill exercise responses in the Framingham Heart Study. *BMC Medical Genetics* 8 Suppl 1: S2 (cit. on pp. 69, 155).

Vatta, M., Mohapatra, B., Jimenez, S., Sanchez, X., Faulkner, G., Perles, Z., Sinagra, G., Lin, J.-H., Vu, T. M., Zhou, Q., Bowles, K. R., Di Lenarda, A., Schimmenti, L., Fox, M., Chrisco, M. A., Murphy, R. T., McKenna, W., Elliott, P., Bowles, N. E., Chen, J., & al. (2003) Mutations in Cypher/ZASPin patients with dilated cardiomyopathy and left ventricular non-compaction. *Journal of the American College of Cardiology* 42(11): 2014–2027 (cit. on p. 180).

Villanueva, B., Pong-Wong, R., Fernández, J., & Toro, M. A. (2005) Benefits from marker-assisted selection under an additive polygenic genetic model. *Journal of Animal Science* 83(8): 1747 (cit. on p. 52).

Villard, E., Perret, C., Gary, F., Proust, C., Dilanian, G., Hengstenberg, C., Ruppert, V., Arbustini, E., Wichter, T., Germain, M., Dubourg, O., Tavazzi, L., Aumont, M. C., De Groote, P., Fauchier, L., Trochu, J. N., Gibelin, P., Aupetit, J. F., Stark, K., Erdmann, J., & al. (2011) A genome-wide association study identifies two loci associated with heart failure due to dilated cardiomyopathy. *European Heart Journal* 32(9): 1065–1076 (cit. on p. 69).

Vischer, E. & Chargaff, E. (1948) The composition of the pentose nucleic acids of yeast and pancreas. *The Journal of Biological Chemistry* 176(2): 715–34 (cit. on p. 31).

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* 101(1): 5–22 (cit. on p. 38).

Wain, L. V., Verwoert, G. C., O'Reilly, P. F., Shi, G., Johnson, T., Johnson, A. D., Bochud, M., Rice, K. M., Henneman, P., Smith, A. V., Ehret, G. B., Amin, N., Larson, M. G., Mooser, V., Hadley, D., Dörr, M., Bis, J. C., Aspelund, T., Esko, T., Janssens, A. C. J. W., & al. (2011) Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nature Genetics* 43(10): 1005–1012 (cit. on p. 189).

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., & al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280(5366): 1077–82 (cit. on pp. 33, 37).

Wang, L. W., Leonhard-Melief, C., Haltiwanger, R. S., & Apte, S. S. (2009) Post-translational modification of thrombospondin type-1 repeats in ADAMTS-like 1/punctin-1 by C-mannosylation of tryptophan. *The Journal of Biological Chemistry* 284(44): 30004–15 (cit. on p. 189).

Watson, J. D. & Crick, F. H. C. (1953) Genetical Implications of the Structure of Deoxyribonucleic Acid. *Nature* 171(4361): 964–967 (cit. on p. 31).

Weldon, W. F. R. (1890) The variations occurring in certain Decapod Crustacea. *Proceedings of the Royal Society* 47: 445–453 (cit. on p. 27).

Weldon, W. F. R. (1892) Certain correlated variations in Crangon vulgaris. *Proceedings of the Royal Society* 51: 2–21 (cit. on p. 27).

Wild, P. S., Zeller, T., Schillert, A., Szymczak, S., Sinning, C. R., Deiseroth, A., Schnabel, R. B., Lubos, E., Keller, T., Eleftheriadis, M. S., Bickel, C., Rupprecht, H. J., Wilde, S., Rossmann, H., Diemert, P., Cupples, L. A., Perret, C., Erdmann, J., Stark, K., Kleber, M. E., & al. (2011) A Genome-Wide Association Study Identifies LIPA as a Susceptibility Gene for Coronary Artery Disease. *Circulation: Cardiovascular Genetics* 4(4): 403–412 (cit. on p. 38).

Wilks, S. S. (1938) The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics* 9(1): 60–62 (cit. on pp. 43, 99).

Willer, C. J., Speliotes, E. K., Loos, R. J. F., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., Elliott, A. L., Jackson, A. U., Lamina, C., Lettre, G., Lim, N., Lyon, H. N., McCarroll, S. A., Papadakis, K., Qi, L., Randall, J. C., Roccasecca, R. M., Sanna, S., Scheet, P., & al. (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics* 41(1): 25–34 (cit. on p. 38).

Wong, N. D. (2014) Epidemiological studies of CHD and the evolution of preventive cardiology. *Nature Reviews Cardiology* 11(5): 276–289 (cit. on p. 66).

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., Amin, N., Buchkovich, M. L., Croteau-Chonka, D. C., Day, F. R., Duan, Y., Fall, T., Fehrmann, R., Ferreira, T., Jackson, A. U., Karjalainen, J., & al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* 46(11): 1173–1186 (cit. on p. 38).

World Health Organisation (2016) *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. 5th. Geneva (cit. on p. 66).

Wright, F. A., Huang, H., Guan, X., Gamiel, K., Jeffries, C., Barry, W. T., Pardo-Manuel de Villena, F., Sullivan, P. F., Wilhelmsen, K. C., & Zou, F. (2007) Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics* 23(19): 2581–2588 (cit. on pp. 72, 73).

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., & Lin, X. (2010) Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *The American Journal of Human Genetics* 86(6): 929–942 (cit. on pp. 38, 193).

Xu, C., Tachmazidou, I., Walter, K., Ciampi, A., Zeggini, E., Greenwood, C. M. T., & UK10K Consortium (2014) Estimating genome-wide significance for whole-genome sequencing studies. *Genetic Epidemiology* 38(4): 281–90 (cit. on p. 47).

Xu, X., Tian, L., & Wei, L. J. (2003) Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics* 4(2): 223–229 (cit. on p. 55).

Yang, J., Lee, S. H., Goddard, M. E. M., Visscher, P. M. P., Hindorff, L., Sethupathy, P., Junkins, H., Ramos, E., Mehta, J., Collins, F., Manolio, T., Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., Hunter, D., McCarthy, M., Ramos, E., Cardon, L., & al. (2011) GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* 88(1): 76–82 (cit. on pp. 38, 48, 53, 59, 88, 117).

Yang, J., Loos, R. J. F., Powell, J. E., Medland, S. E., Speliotes, E. K., Chasman, D. I., Rose, L. M., Thorleifsson, G., Steinthorsdottir, V., Mägi, R., Waite, L., Smith, A. V., Yerges-Armstrong, L. M., Monda, K. L., Hadley, D., Mahajan, A., Li, G., Kapur, K., Vitart, V., Huffman, J. E., & al. (2012) FTO genotype is associated with phenotypic variability of body mass index. *Nature* 490(7419): 267–72 (cit. on p. 38).

Yang, Q. & Wang, Y. (2012) Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies. *Journal of Probability and Statistics* 2012: 1–13 (cit. on pp. 54–56).

Yang, Q., Wu, H., Guo, C.-Y., & Fox, C. S. (2010) Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic Epidemiology* 34(5): 444–454 (cit. on p. 55).

Yang, R., Yi, N., & Xu, S. (2006) Box–Cox transformation for QTL mapping. *Genetica* 128(1-3): 133–143 (cit. on p. 43).

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., Fang, W., Feng, H., Xie, W., Lian, X., Wang, G., Luo, Q., Zhang, Q., Liu, Q., & Xiong, L. (2014) Combin-

ing high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nature Communications* 5: 5087 (cit. on pp. 56, 88, 104).

Yousef, Z. R., Foley, P. W., Khadjooi, K., Chalil, S., Sandman, H., Mohammed, N. U., & Leyva, F. (2009) Left ventricular non-compaction: clinical features and cardiovascular magnetic resonance imaging. *BMC Cardiovascular Disorders* 9(37) (cit. on pp. 180, 189).

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38(2): 203–208 (cit. on pp. 50, 52, 193).

Yuan, X., Miller, D. J., Zhang, J., Herrington, D., & Wang, Y. (2012) An overview of population genetic data simulation. *Journal of Computational Biology* 19(1): 42–54 (cit. on p. 72).

Zambrano, E., Marshalko, S. J., Jaffe, C. C., Hui, P., Sandman, H., Mohammed, N. U., Leyva, F., Zambrano, E., Marshalko, S., Jaffe, E., Hui, P., Jenni, R., Oechslin, E., Jost, C. A., Kaufmann, P., Ritter, M., Oechslin, E., Siutsch, G., Attenhofer, C., Schneider, J., & al. (2002) Isolated Noncompaction of the Ventricular Myocardium: Clinical and Molecular Aspects of a Rare Cardiomyopathy. *Laboratory Investigation* 82(2): 117–122 (cit. on pp. 64, 179, 180).

Zeglinski, M. R., Davies, J. J. L., Ghavami, S., Rattan, S. G., Halayko, A. J., & Dixon, I. M. C. (2016) Chronic expression of Ski induces apoptosis and represses autophagy in cardiac myofibroblasts. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1863(6): 1261–1268 (cit. on p. 170).

Zhang, F., Guo, X., Wu, S., Han, J., Liu, Y., Shen, H., & Deng, H.-W. (2012) Genome-Wide Pathway Association Studies of Multiple Correlated Quantitative Phenotypes Using Principle Component Analyses. *PLoS ONE* 7(12). Ed. by M. Xiong: e53320 (cit. on pp. 144, 194).

Zhang, M., Song, F., Liang, L., Nan, H., Zhang, J., Liu, H., Wang, L.-E., Wei, Q., Lee, J. E., Amos, C. I., Kraft, P., Qureshi, A. A., & Han, J. (2013) Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. *Human Molecular Genetics* 22(14): 2948–2959 (cit. on p. 38).

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010) Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* 42(4): 355–360 (cit. on pp. 38, 88, 89).

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., & Nordborg, M. (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genetics* 3(1): e4 (cit. on pp. 50, 53).

Zhou, X. & Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44(7) (cit. on p. 193).

Zhou, X. & Stephens, M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods* 11(4): 407–9 (cit. on pp. 57, 71, 76, 78, 79, 81, 82, 89, 90).

Zirkle, C. (1935) *The Inheritance of Acquired Characters and the Provisional Hypothesis of Pangenesis* (cit. on p. 24).