Faculty of Medicine
University of Helsinki
Finland

FINEMAP – a statistical method for identifying causal genetic variants

Christian Benner

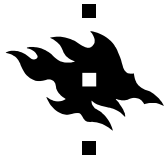
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Helsinki, Finland

and

Department of Public Health, Faculty of Medicine,
University of Helsinki, Helsinki, Finland

ACADEMIC DISSERTATION

To be presented with the permission of the Faculty of Medicine of the University of Helsinki for public examination in lecture hall 3 / Biomedicum 1 on 18.01.2019 at 2 o'clock.



UNIVERSITY OF HELSINKI

ISBN 978-951-51-4805-6 (paperback)
ISBN 978-951-51-4806-3 (PDF)
ISSN 2342-3161 (Print)
ISSN 2342-317X (Online)

Helsinki University Printing House
Helsinki 2019

Supervisors Dr Matti Pirinen

Institute for Molecular Medicine Finland (FIMM)
University of Helsinki
Helsinki, Finland

Department of Public Health
University of Helsinki
Helsinki, Finland

Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics
University of Helsinki
Helsinki, Finland

Prof Samuli Ripatti

Institute for Molecular Medicine Finland (FIMM)
University of Helsinki
Helsinki, Finland

Department of Public Health
University of Helsinki
Helsinki, Finland

Broad Institute of MIT and Harvard
Cambridge, USA

Reviewers Docent Sangita Kulathinal

National Institute for Health and Welfare (THL)
Helsinki, Finland

Dr Harri Lähdesmäki

Aalto University
Espoo, Finland

Turku Center for Biotechnology
Turku University
Turku, Finland

Opponent Dr Zoltán Kutalik

Institute of Social and Preventive Medicine (IUMSP)
Lausanne University Hospital
Lausanne, Switzerland

Swiss Institute of Bioinformatics
Lausanne, Switzerland

To Yuri

ABSTRACT

The explosion of genomic data during the last ten years and the advent of Genome-Wide Association Studies (GWAS) have led to robust statistical associations between thousands of genomic regions and hundreds of phenotypes. However, any one associated genomic region can harbor thousands of correlated genetic variants, complicating the understanding of the underlying biological mechanisms that led to these associations. To address this problem, this doctoral thesis presents the development of the FINEMAP software for fine-mapping causal variants in these regions.

In 2016, we solved the existing issue with the computationally expensive exhaustive search strategy of existing fine-mapping methods by implementing a Bayesian regression model and an ultrafast stochastic search algorithm in the FINEMAP software. We demonstrated that FINEMAP opens up completely new opportunities by fine-mapping the High Density Lipoprotein (HDL) cholesterol association to the *LIPC* locus with 20,000 variants in less than 90 seconds, while exhaustive search would require many years. With extensive simulations we further showed that FINEMAP is as accurate as exhaustive search when the latter can be completed and achieves even higher accuracy when the latter must be restricted due to computational reasons. Thus, FINEMAP is a promising tool for future fine-mapping analyses.

Fine-mapping methods that use GWAS results also require Linkage Disequilibrium (LD) information as input in the form of estimates of pairwise correlations between variants. Motivated by feedback from FINEMAP users, we investigated in 2017 the consequences of misspecification of LD that could happen when publicly available reference genomes are used. We demonstrated both empirically and theoretically that the size of the reference panel needs to scale with the GWAS sample size to produce accurate results and we provided the LDstore software to help share LD estimates. This finding has important consequences for the application of all fine-mapping methods using GWAS results from GWAS consortia in which accurate LD estimates from each participating study are typically not available.

In 2018, we implemented in FINEMAP an approach for estimating how much phenotypic variation can be explained by the causal variants. To demonstrate this, we applied FINEMAP to 110 regions across 51 biomarkers on 5,265 Finnish samples. We compared regional heritability estimation using FINEMAP with both the variance component model BOLT and fixed-effect model HESS in biomarker-associated regions, showing good concordance among all methods. Through simulations with biobank-scale projects, we also illustrated how violations of model assumptions on polygenicity or unspecified genetic architecture induces inaccuracy to the existing heritability estimates that becomes more accentuated as statistical power to identify causal variants increases. Ever increasing GWAS sample sizes, soon reaching millions of samples, provide unprecedented statistical power to decompose heritability estimates from polygenic models into heritability contributions from causal variants.

In conclusion, this doctoral thesis shows that (1) the computational efficiency and accuracy of FINEMAP makes it a promising fine-mapping tool, (2) LD estimates need to be chosen more carefully than previously thought to avoid bias, and (3) large-scale data sets provide new opportunities for fine-mapping to deduce a variant-level picture of regional genetic architecture.

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This doctoral thesis is based on the following original publications and they are referred to in the text by their Roman numerals.

I. **Benner C**, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, and Pirinen M. FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32, 1493-1501 (2016).

II. **Benner C**, Havulinna AS, Järvelin MR, Salomaa V, Ripatti S, and Pirinen M. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* (2017).

III. **Benner C**, Havulinna AS, Salomaa V, Ripatti S, and Pirinen M. Refining fine-mapping: effect sizes and regional heritability. bioRxiv (2018). https://doi.org/10.1101/318618

Author contributions

I. **C.B.** and M.P. designed the study. **C.B.** developed the software tools and conducted the analyses. C.C.A.S. A.S.H., and V.S. provided materials. S.R. and M.P. supervised the research. **C.B.** and M.P. wrote the manuscript. All authors reviewed the manuscript.

II. **C.B.** and M.P. designed the study. **C.B.** developed the software tools and conducted the analyses. A.S.H., M.R.J and V.S. provided materials. S.R. and M.P. supervised the research. **C.B.** and M.P. wrote the manuscript. All authors reviewed the manuscript.

III. **C.B.** and M.P. designed the study. **C.B.** developed the software tools and conducted the analyses. A.S.H. and V.S. provided materials. S.R. and M.P. supervised the research. **C.B.** and M.P. wrote the manuscript. All authors reviewed the manuscript.

ABBREVIATIONS

1000GP	1000 Genomes Project
BVS	Bayesian Variable Selection
DNA	Deoxyribonucleic Acid
DZ	Dizygotic
GRM	Genetic Relatedness Matrix
GWAS	Genome-Wide Association Study / Studies
HRC	Haplotype Reference Consortium
HDL	High Density Lipoprotein
IRLS	Iteratively Reweighted Least Squares
LD	Linkage Disequilibrium
LDL	Low Density Lipoprotein
MLE	Maximum Likelihood Estimate
MCMC	Markov Chain Monte Carlo
MAP	Maximum A Posteriori
MH	Metropolis-Hastings
MAF	Minor Allele Frequency
MZ	Monozygotic
NFBC	Northern Finland Birth Cohort
PON1	Paraoxonase 1
PC	Principal Component
PDF	Probability Density Function
RJMCMC	Reversible Jump Markov Chain Monte Carlo
SSS	Shotgun Stochastic Search
SNP	Single-Nucleotide Polymorphism
SE	Standard Error
UKBB	UK Biobank
WTCCC	Wellcome Trust Case Control Consortium

INTRODUCTION

Genetic factors contribute to complex human phenotypes, together with environmental exposures and lifestyle choices. The explosion of genomic data during the last ten years has remarkably improved our knowledge of the genetic basis of phenotypes and has created comprehensive catalogues of genetic risk factors that affect hundreds of phenotypes[1]; in some disease phenotypes independently of well-known factors[2]. Modern human genetics may truly be on the cusp of uncovering genes that could become drug targets for the development of new medicine, and thereby improve the success rate of drug discovery and development[3].

GWAS have been extremely successful in identifying genomic regions underlying various phenotypes since the landmark publication by the Wellcome Trust Case Control Consortium in 2007[4]. Since then, hundreds of robust associations between genetic markers and phenotypes have been discovered through GWAS for various phenotypes such as lipid traits[5], coronary artery disease[6], Crohn's disease[7], schizophrenia[8], and type 2 diabetes[9]. Robust GWAS regions point to underlying biological mechanisms of the phenotype, but any one associated genomic region harbors thousands of correlated genetic variants, complicating the understanding of these mechanisms. Fine-mapping is a crucial post-GWAS analysis aiming to bridge the gap from GWAS to biology by refining the large set of variants simply associated with phenotypes down to a much smaller set of variants with a direct effect on the phenotypes, hereafter called causal variants, by taking into account the complex correlations between variants[10,11].

Conditional analysis[12] implemented as stepwise greedy forward selection has been a standard approach for pinpointing causal variants. Conditional analysis starts by first selecting the variant with the lowest P-value from the GWAS and then iteratively selects the variant with the lowest conditional P-value until no further variant reaches the genome-wide significance threshold or until a prespecified number of iterations have been conducted. While conditional analysis is a simple approach, the procedure 1) quantifies only the number of causal variants but lacks probabilistic assessments of the causality for individual variants, 2) becomes computationally expensive with increasing

number of variants to condition on, and 3) is unstable when the number of variants to condition on is close to the GWAS sample size.

To overcome the problems of conditional analysis, fine-mapping methods have also been developed using methodology of Bayesian Variable Selection (BVS). These include exhaustive search[13], stochastic search[14-16] or variational approximation[17]. One earlier Bayesian fine-mapping approach circumvented the need for search strategies by assuming only a single causal variant in the region[18], which is often not the case. During the last three years, there has been a surge of Bayesian fine-mapping methods that work directly on the GWAS results to avoid privacy concerns and the logistics of sharing individual-level genotype-phenotype data. Existing Bayesian implementations that use GWAS results and allow for multiple causal variants are CAVIAR[19], PAINTOR[20], CAVIARBF[21], FINEMAP[22], JAM[23] and DAP[24].

This doctoral thesis presents the development of the FINEMAP software. FINEMAP is a program for 1) identifying causal variants, 2) estimating effect sizes of causal variants, and 3) estimating the heritability contribution of causal variants. FINEMAP is computationally efficient by using GWAS results and robust by applying stochastic search. It produces accurate results in a fraction of the processing time of other Bayesian fine-mapping implementations, and this makes it a promising tool for analyzing the growing amounts of data produced in GWAS and in emerging sequencing or biobank projects.

1 REVIEW OF LITERATURE

Modern genetics research applies data analysis to uncover the genetic basis of traits and diseases in humans. Data analysis links genetics to statistics in three areas: (1) description, (2) inference, and (3) prediction. Descriptive statistics are used to characterize data. For example, genetic data can be used to visualize the geographic distribution of known genetic risk factors for rare diseases and to describe patterns of disease occurrence[25]. Attempts at inference use statistical methods to draw inference about some aspect of a population based on data from a subsample of that population. To identify potential drug targets, genomic data and disease status of individuals from a population could be used to infer statistical associations between genetic factors and the disease in that population[26]. The aim of statistical prediction is to quantify the uncertainty about the occurrence of future events. For example, assuming that the genetic risk factors for a disease were known, risk predictions from genomic data could be used to inform medical decision making[27].

A major aim of modern genetics research is to reveal the biology underlying genotype-phenotype associations in order to identify drug targets for the development of new medicine[28]. This doctoral thesis focuses on inferential statistical methods for fine-mapping genotype-phenotype associations. Fine-mapping holds promise for facilitating the discovery of the biological mechanisms behind genotype-phenotype associations. In the following, I will introduce core concepts and terminology in human genetics research and inferential statistical methods. Afterwards, I will outline regression modeling and how this class of methods is applied in GWAS. The review will finish with an illustration of BVS methods for fine-mapping causal variants.

1.1 Concepts and terminology

1.1.1 Human genome

The human genome is located in the nucleus of each cell and contains the information that specifies cellular functioning[29]. The genome is inherited from an individual's parents and each cell carries two copies. The information content of the genome is coded in the chemical compound Deoxyribonucleic Acid (DNA). DNA is a double-stranded molecule, with each strand composed of a linear sequence of nucleotides. A nucleotide consists of a nucleoside and three phosphate groups. A nucleoside is a nitrogenous base connected to a deoxyribose sugar. There are four different nucleotides which are defined by the presence of different nitrogenous bases, which are Adenine (*A*), Cytosine (*C*), Guanine (*G*) and Thymine (*T*). The nucleotides adhere to a specific base pairing in double-stranded DNA: adenine always pairs with thymine whereas cytosine pairs with guanine.

While most individuals in a population carry genomes that have the same nucleotide at a given position, there can be individuals with genomes with a different nucleotide at that position. This nucleotide difference is called a genetic variant. Different nucleotides at the position where the variant occurs are called alleles. There exist biallelic (two allele) and multiallelic (more than two allele) variants, but only biallelic variants are considered in the following work. The incidence of an allele in a population is described by the allele's frequency. In population genetics, the allele that occurs with lower frequency is called the minor allele. Variants with a Minor Allele Frequency (MAF) of at least 5% are typically called Single-Nucleotide Polymorphisms (SNPs), whereas variants between 0.5% and 5% are denoted as low-frequency variants[30]. The genotype of an individual at a variant is the combination of the two alleles of the variant. Let the two alleles of a variant be represented by letters *A* and *B*. The possible genotypes for the variant are then *A*/*A*, *A*/*B* and *B*/*B*. By counting the number of copies of the chosen reference allele *B*, the genotype takes the value 0, 1 or 2.

1.1.2 Linkage disequilibrium

A combination of alleles at nearby variants are often inherited together from the same genome of the parent[31]. Such combinations of alleles are called haplotypes. The coinheritance of alleles causes population-specific patterns of LD between variants which is the statistical dependence of the genotypes for the variants. Let A_1 and B_1 denote the alleles at variant 1 and A_2 and B_2 the alleles at variant 2. A common statistic to quantify LD between the two variants is

$$D' = \frac{|D|}{D_{\max}} \text{ with } D = P_{A_1A_2} - P_{A_1}P_{A_2} \text{ and } D_{\max} = \begin{cases} \min\left(P_{A_1}P_{B_2}, P_{B_1}P_{A_2}\right) & D > 0 \\ \min\left(P_{A_1}P_{A_2}, P_{B_1}P_{B_2}\right) & D < 0 \end{cases},$$

where $P_{A_1}, P_{B_1}, P_{A_2}, P_{B_2}$ are the population frequencies of the alleles and $P_{A_1A_2}$ is the population frequency of the A_1A_2 haplotype. An alternative statistic to measure LD is the square of the Pearson's correlation coefficient r

$$r = \frac{D}{\sqrt{P_{A_1}P_{B_1}P_{A_2}P_{B_2}}},$$

where r is the relevant statistic for fine-mapping methods, as shown in section 1.4.1. Although the Pearson's correlation coefficient is defined for haplotype data, it is commonly estimated from genotype data.

Figure 1 shows LD estimates as squared Pearson's correlations for pairs of variants in the *PON1* locus on chromosome 7. The top of the triangle highlights each variant by a grey square. To obtain the LD estimate for two variants, the columns for the two variants, starting from the grey squares, can be traced until the columns intersect. There is clearly a block of variants in high LD in the middle of the *PON1* locus, indicated by a red line. One way LD between two variants can break down is through exchange of the maternal and paternal genomic segments on which the two variants reside during sexual

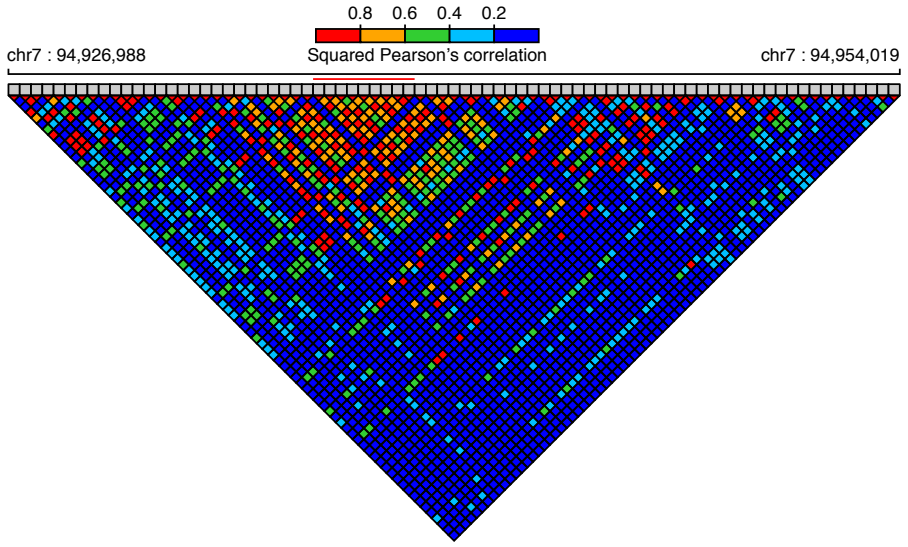


Figure 1. LD of variants within the *PON1* locus. LD was estimated as the squared Pearson's correlation between variants using the genotypes of 27,294 individuals from the FINRISK study. Each variant is represented by a grey square on top of the triangle. For any two variants, the squared Pearson's correlation can be obtained by following the columns for each variant until the columns intersect.

reproduction. The process is called genetic recombination and has higher chance of occurring between two variants if the physical distance between them is large[32].

1.1.3 DNA microarrays

DNA microarray experiments have become a popular approach for the simultaneous investigation of 500,000 to 1,000,000 variants on a genome-wide scale[33]. Figure 2 shows a schematic of a DNA microarray experiment. Single-stranded DNA fragments from the individual being genotyped, commonly referred to as targets, are labeled with fluorescence dye molecules. After labeling, the targets are put in contact with a DNA microarray slide. Fixed to the slides are numerous identical single-stranded nucleic acids called probes. A variant is represented on a slide by including probes which differ in their DNA sequence only with respect to the two alleles of the variant. Since the targets and probes are single-stranded, DNA microarray technology takes advantage of the specific base pairing of single-stranded nucleic acids with their counterpart. Putting the target DNA

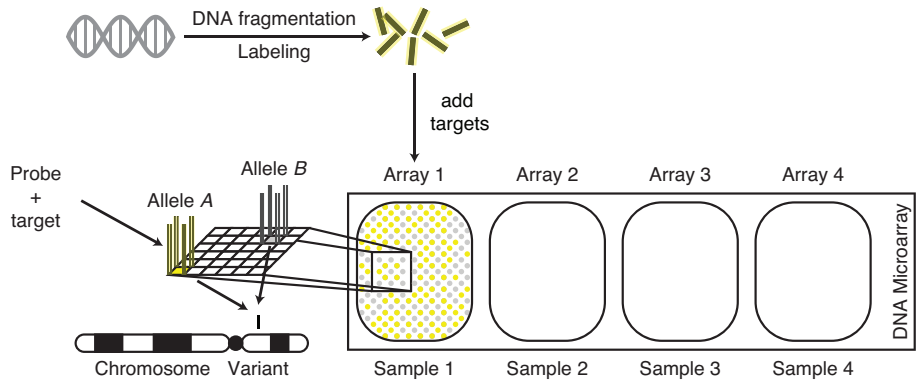


Figure 2. Schematic of a DNA microarray experiment. Single-stranded DNA target fragments are labeled with a fluorescent dye. The labeled targets are put in contact with a DNA microarray slide. Each array contains numerous identical single-stranded nucleic acids called probes fixed to the slide. A probe represents one of the alleles of a variant. The fluorescent dye intensities of the annealed targets and probes can be used to infer the genotypes of an individual at a variant.

in contact with the DNA probes on the microarray slide can thus cause the targets to join with their complementary probes to form double-stranded nucleic acids. Scanning the DNA microarray with a laser causes the fluorescent dye molecules to glow and high-resolution images are obtained. Image analysis algorithms are used to quantify the fluorescent dye intensities of the targets. Inference of the genotype of individuals at a variant is based on clustering the fluorescence dye intensities of the two alleles of a variant.

1.1.4 Genome-wide association studies

The technological innovation of DNA microarrays has enabled GWAS that test millions of variants for a statistical association to a phenotype. A methodological development called imputation has been an additional key analytical advance for GWAS[34]. Imputation uses correlations between nearby variants from a publicly available reference sample, such as the 1000 Genome Project[32] (1000GP) or the Haplotype Reference Consortium[35] (HRC), with dense genotype data. These reference data are used for prediction of the genotypes at the subset of variants not determined directly from the DNA microarray of a large study sample. The use of imputation means that DNA microarrays can genotype

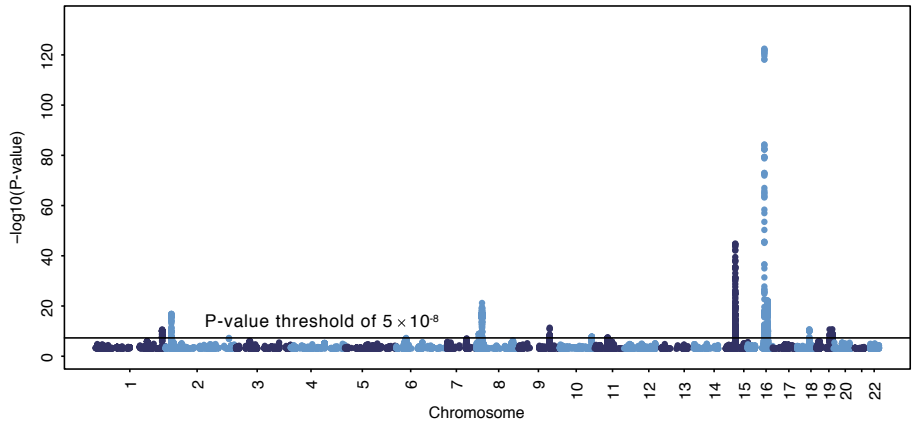


Figure 3. Manhattan plot for a HDL GWAS on 21,320 individuals from the FINRISK study. Negative $\log_{10}$ P-values (y-axis) from independent linear regression of each variant are plotted against the variant's physical position along chromosomes (x-axis).

an individual at about 500,000 to 1,000,000 variants but a much larger number of genome-wide variants are actually examined in the GWAS.

Figure 3 visualizes the results from a GWAS for HDL cholesterol levels in the form of a Manhattan plot. Each point displays the negative $\log_{10}$-transformed P-value from an independent linear regression of the HDL cholesterol levels of 21,320 individuals from the FINRISK study onto genotype data at a variant as a function of the physical position of the variant. Genomic regions with statistically significant associations supported by many correlated neighboring variants show up as peaks in a Manhattan plot. Because millions of independent statistical tests are performed or the prior probability of an association is so small (see Box 2 in [4]), a very small significance threshold of 5×10^{-8} is used to guard against false positive findings[36].

For quantitative phenotypes, the statistical power increases with 1) study sample size, 2) MAF of the variant, and 3) effect size per one reference allele of the variant. For binary disease phenotypes, the statistical power further increases with the proportion of cases with the disease. For illustration, consider Paraoxonase 1 (PON1) serum levels, which have been reported to be inversely associated with systemic oxidative stress and atherosclerosis risk[37]. Figure 4 shows that carrying two copies of the *T* allele at rs1157745 results in a two standard deviation increase in PON1 serum levels compared to carrying no copy of the *T* allele. This result is supported by a genome-wide significant

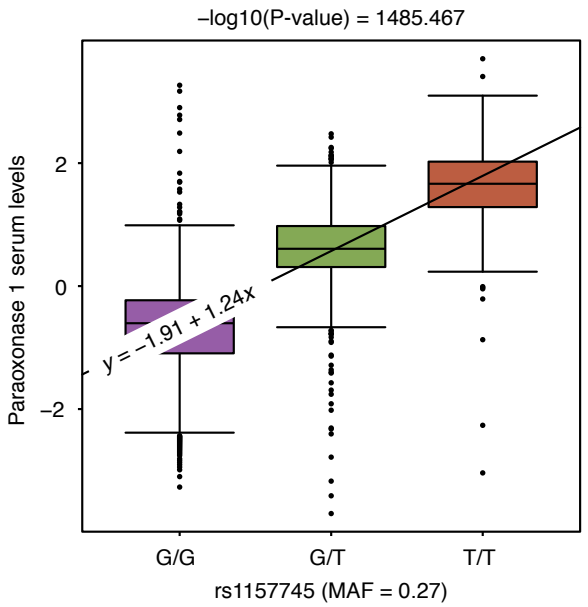


Figure 4. Distribution of PON1 serum levels among the three possible genotypes at rs1157745 among 4,613 individuals from the FINRISK study together with the linear regression line.

P-value from testing whether the slope of the linear regression line is significantly different from zero. With a study sample size of 4,613 individuals from the FINRISK study, the statistical power between PON1 and rs1157745 is practically 100% because of the SNP's large effect size and frequency of the *T* allele.

The first large-scale GWAS was conducted by the Wellcome Trust Case Control Consortium (WTCCC) which was established in 2005[4]. The WTCCC was a collaboration of over 50 research groups in the UK to study the genetic factors underlying bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. The consortium combined 2,000 cases and 3,000 controls for each disease and discovered 24 genomic regions (P-values $< 5 \times 10^{-7}$). To improve the statistical power for variants at lower frequencies or with smaller effect sizes, a common strategy is to increase study sample size by meta-analyzing GWAS results. Recent GWAS consortia combine hundreds of studies throughout the world to investigate, for example, anthropometric traits[38] in up to 339,224 individuals or schizophrenia[8] using 36,989 cases and 113,075 controls.

1.1.5 Heritability

Genetic and environmental factors together contribute to complex human phenotypes. Heritability is a population-level parameter which quantifies the proportion of phenotypic variance that is explained by genetic factors[39]. The concept of heritability can be divided into narrow-sense and broad-sense heritability. Narrow-sense heritability considers only genetic factors acting in an additive fashion, whereas broad-sense heritability considers all genetic factors including those with an additive, dominant or epistatic modes of action. High narrow-sense heritability, which indicates that there is a substantial additive genetic component for the phenotype, can give an idea of how successful a GWAS would be in terms of finding genetic factors for the phenotype.

Traditionally, heritability estimates have been obtained using measures of Monozygotic (MZ) and Dizygotic (DZ) twins. Monozygotic twins are 100% genetically similar, whereas DZ twins share about 50% of their genome with each other. Since the environment is assumed to be similar for MZ and DZ twins, any phenotypic differences between MZ twins cannot be due to either of these factors. Hence, comparison of the phenotypic correlations between pairs of MZ and DZ twins provides an estimate of the magnitude of the narrow-sense heritability of that phenotype, albeit under a strong assumption of similar environments in both types of twins.

With the advent of DNA microarrays, estimation of the heritability contribution from genome-wide significant variants in GWAS has become possible. However, this showed that variants highlighted by GWAS explain only a small proportion of phenotypic variance, much less than that shown by twin studies. This phenomenon has been called the missing heritability problem[40] and has motivated research into methods that can estimate the heritability contribution from all genome-wide variants including those that are not genome-wide significant. Even still, estimates of the heritability contribution from the whole genome are typically much smaller than heritability estimates from twin studies. A possible explanation is that the heritability contribution of rarer variants, with MAF less than 1%, are not adequately captured by DNA microarrays and imputation

methods. On the other hand, heritability estimates from twin studies could be upwardly biased by considering non-additive genetic factors[41] in addition to additive factors.

Among the methods considering the heritability contribution from the whole genome, variance component models (as implemented in the software BOLT[42]) are based on the notion that estimation of the phenotypic variance explained from linear regression using all genotyped variants is equivalent to estimating the genetic variance with a linear mixed model in which the genetic relatedness matrix determines the covariance structure of the random effects. The assumption of these methods is that the phenotype is characterized by a polygenic architecture, under which all variants have tiny effects on the phenotype.

Recently, heritability estimation in genomic regions directly from GWAS results was introduced in the software package HESS[43] under the assumption of arbitrary genetic architecture and a fixed-effect model for causal effect sizes. HESS estimates heritability through a regularized quadratic form built on marginal effect size estimates of all variants and their LD estimates. This is an alternative to computationally expensive variance component models that rely on individual-level genotype-phenotype data.

1.2 Inferential statistics

Inferential statistics[44] is the branch of statistics used to draw conclusions about some aspect of a population based on data from a subsample of that population. In statistical models, data $\boldsymbol{y}$ for n individuals is assumed to be the realization of a random vector $\boldsymbol{Y}$ with probability distribution parameterized by a p-dimensional population parameter $\boldsymbol{\theta}$. The population parameter $\boldsymbol{\theta}$ characterizes the population for which inference is of interest.

Let $\boldsymbol{y} \mapsto f(\boldsymbol{y} \mid \boldsymbol{\theta})$ be the Probability Density Function (PDF) of the distribution of $\boldsymbol{Y}$. Given data $\boldsymbol{y}$ and considered as a function of $\boldsymbol{\theta}$, the PDF becomes the likelihood function $\boldsymbol{\theta} \mapsto f(\boldsymbol{y} \mid \boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y})$. The likelihood function describes for each candidate value $\boldsymbol{\theta}$ the likelihood that these parameter values led to the data $\boldsymbol{y}$. The most likely

candidate values of $\boldsymbol{\theta}$ are called Maximum Likelihood Estimates (MLEs) $\widehat{\boldsymbol{\theta}}$, which are the values that maximize $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y})$.

The variance of $\widehat{\boldsymbol{\theta}}$ is found by noting that the Hessian $\nabla^2_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$ of the log-likelihood function measures the curvature at $\widehat{\boldsymbol{\theta}}$. If the curvature is small at $\widehat{\boldsymbol{\theta}}$, then the log-likelihood function is flat near $\widehat{\boldsymbol{\theta}}$ and there are other likely candidate values; hence the variance estimate for $\widehat{\boldsymbol{\theta}}$ is large. On the other hand, a large curvature at $\widehat{\boldsymbol{\theta}}$ means that the log-likelihood function is sharply peaked at $\widehat{\boldsymbol{\theta}}$ leading to a small variance estimate. The estimate $\widehat{\boldsymbol{\theta}}$ and their Standard Errors (SEs) are computed as

$$\operatorname{diag}\left(\left[-\nabla^2_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right]^{-1}\right)^{\frac{1}{2}}$$

and are used for confidence interval estimation and hypothesis testing of individual values of the population parameter $\boldsymbol{\theta}$.

In the Bayesian framework[45], both $\boldsymbol{Y}$ and $\boldsymbol{\theta}$ are considered random. This means that $\boldsymbol{\theta}$ also follows a probability distribution. The probability distribution of $\boldsymbol{\theta}$ is called the prior distribution and represents the uncertainty about values of the population parameter $\boldsymbol{\theta}$ before data is observed. In Bayesian inference, $\widehat{\boldsymbol{\theta}}$ is not used as a point estimate for $\boldsymbol{\theta}$ but instead the prior distribution is updated with data to a posterior distribution using Bayes's rule. The PDF of the posterior distribution is obtained through Bayes' rule as

$$\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}) f(\boldsymbol{\theta})}{f(\boldsymbol{y})} \propto \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}) f(\boldsymbol{\theta}),$$

where $f(\boldsymbol{\theta})$ is the PDF of the prior distribution of $\boldsymbol{\theta}$ and $f(\boldsymbol{y})$ is the PDF of the marginal distribution of $\boldsymbol{Y}$.

The posterior distribution is used for inference about the population parameter $\boldsymbol{\theta}$. The Bayesian paradigm is different from likelihood inference in that Bayesian inference depends on prior knowledge about the population parameter $\boldsymbol{\theta}$ through the prior

distribution. Such prior knowledge can be particularly useful when there is not much information in the data, for instance due to a small sample size. However, both Bayesian and likelihood frameworks yield similar inference if the sample size is large, resulting in a likelihood function that is strongly concentrated around a particular set of values of the parameter vector and the prior distribution supports that region with a reasonable prior probability.

In this doctoral thesis, Bayesian statistics is used because it provides a probabilistic description of all the parameters of interest, and this can be useful in downstream analysis of the results.

1.2.1 Example: Modeling genotype counts

To illustrate the concepts of Maximum Likelihood and Bayesian inference, let the random variable G denote the number of copies of allele B at a genetic variant in an individual. Since Hardy-Weinberg equilibrium[46] holds at most variants, assume that G follows a binomial distribution with parameters $M = 2$ and Θ, where Θ is the frequency of allele B in the population. The probability mass function of the binomial distribution with parameters $M \in \mathbb{N}$ and $\Theta \in (0,1)$ is

$$\mathrm{Bin}(y \mid M = m, \Theta = \theta) = \binom{m}{y} \theta^y (1-\theta)^{m-y}\,, \qquad y = 0,1,\dots,m\,.$$

Suppose genotype counts $\boldsymbol{g} = (g_1, g_2, \dots, g_n)^T$ of n individuals are independent realizations from random variables $\boldsymbol{G} = (G_1, G_2, \dots, G_n)^T$ following a binomial distribution with unknown allele B frequency Θ. The likelihood function for Θ is

$$\mathcal{L}(\theta \mid \boldsymbol{g}) = \prod_{i=1}^{n} \mathcal{L}(\theta \mid g_i) = \prod_{i=1}^{n} \mathrm{Bin}(g_i \mid 2, \theta) \propto \theta^{n\bar{g}} (1-\theta)^{2n-n\bar{g}}\,,$$

where $\bar{g} = n^{-1}\sum_{i=1}^{n} g_i$. The MLE of Θ can be found by differentiating $\log \mathcal{L}(\theta \mid \boldsymbol{g})$ with respect to θ, setting the results equal to zero, and solving for θ as follows

$$\frac{\mathrm{d}\log \mathcal{L}(\theta \mid \boldsymbol{g})}{\mathrm{d}\,\theta} = \frac{n\bar{g}}{\theta} - \frac{2n - n\bar{g}}{1-\theta} = 0$$
$$\frac{n\bar{g}}{\theta} = \frac{2n - n\bar{g}}{1-\theta}$$
$$\hookrightarrow \hat{\theta} = \frac{\bar{g}}{2}.$$

The negative second derivative of $\log \mathcal{L}(\theta \mid \boldsymbol{g})$ is given by

$$-\frac{\mathrm{d}^2 \log \mathcal{L}(\theta \mid \boldsymbol{g})}{\mathrm{d}\,\theta^2} = \frac{n\bar{g}}{\theta^2} + \frac{2n - n\bar{g}}{(1-\theta)^2}.$$

Plugging in $\hat{\theta}$ yields

$$-\left.\frac{\mathrm{d}^2 \log \mathcal{L}(\theta \mid \boldsymbol{g})}{\mathrm{d}\,\theta^2}\right|_{\theta=\hat{\theta}} = \frac{2n\hat{\theta}}{\hat{\theta}^2} + \frac{2n - 2n\hat{\theta}}{\left(1-\hat{\theta}\right)^2} = \frac{2n}{\hat{\theta}\left(1-\hat{\theta}\right)}.$$

The variance estimate of $\hat{\theta}$ is thus $\widehat{V(\hat{\theta})} = \hat{\theta}\left(1-\hat{\theta}\right)/2n$. A confidence interval estimate with confidence level $(1-\alpha)$ can be approximated by using

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{\theta}\left(1-\hat{\theta}\right)/2n},$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution.

In the Bayesian framework, a beta prior distribution can be specified for the parameter Θ. The PDF for the beta distribution with parameters $\alpha > 0$ and $\beta > 0$ is

$$\text{Beta}(y \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \qquad 0 < y < 1$$

Using Bayes' rule, the posterior density function for Θ is proportional to

$$\begin{aligned} f(\theta \mid \boldsymbol{g}) &\propto \mathcal{L}(\theta \mid \boldsymbol{g}) f(\theta) \\ &= \prod_{i=1}^{n} \mathcal{L}(\theta \mid g_i)\, \text{Beta}(\theta \mid \alpha, \beta) \\ &= \theta^{\alpha + n\bar{g} - 1}(1-\theta)^{\beta + 2n - n\bar{g} - 1} \\ &\propto \text{Beta}(\alpha + n\bar{g}, \beta + 2n - n\bar{g}), \end{aligned}$$

a beta distribution with parameters $\alpha' = \alpha + n\bar{g}$ and $\beta' = \beta + 2n - n\bar{g}$. A point estimate for Θ is the posterior mean of the beta distribution and is given by

$$\text{E}[\theta \mid \boldsymbol{g}] = \frac{\alpha'}{\alpha' + \beta'} = \frac{\alpha + n\bar{g}}{\alpha + \beta + 2n}.$$

A credible interval for Θ can be obtained as an equal tail interval $[F^{-1}(\alpha/2), F^{-1}(1 - \alpha/2)]$ where $F^{-1}(\cdot)$ is the quantile function of the beta posterior distribution and $1 - \alpha$ the coverage probability.

Table 1 shows the genotype counts at rs1157745 among 27,294 individuals from the FINRISK study (see Figure 4). The MLE $\hat{\theta}$ for the population frequency of allele *B* is 0.2737 with 95% confidence interval (0.269997, 0.277477). In the absence of prior knowledge of the population frequency, the parameter of the Beta prior distribution could be set to $\alpha = \beta = 1$ to yield a uniform distribution on the interval $[0,1]$. Using this prior distribution, the posterior mean is 0.2737 with 95% credible interval (0.270012, 0.277493). The absolute difference between $\hat{\theta}$ and $\text{E}[\theta \mid \boldsymbol{g}]$ is smaller than 10^{-5}, showing that the likelihood function dominates the prior distribution with informative data.

Table 1. Genotype counts at rs1157745 among 27,294 individuals from the FINRISK study

G/G	G/T	T/T
14,429	10,787	2,078

1.3 Regression modeling

Millions of variants are tested in GWAS for statistical association to a phenotype. The statistical association tests require a statistical model and data on individuals from the population of interest. The following is an outline of the framework of generalized linear models where two widely used regression models for GWAS, namely linear and logistic regression, can be unified.

1.3.1 Generalized linear models

Generalized linear models[47] were introduced as a generalization of linear regression models. The definition of a generalized linear model is as follows:

(1) A matrix of non-random explanatory variables $\boldsymbol{X}$ of dimension $n \times p$ and regression parameters $\boldsymbol{\beta} = \left[\beta_1, \beta_2, \dots, \beta_p\right]^T$ where the first column of $\boldsymbol{X}$ is $\boldsymbol{x}_{\cdot 1} = \mathbf{1}$.

(2) A random vector $\boldsymbol{Y} = [Y_1, Y_2, \dots, Y_n]^T$ representing the outcome variables. The random variables are independent given the explanatory variables and follow a distribution in the exponential family such as Poisson, binomial or normal.

(3) A link function that combines (1) and (2) as $g(\mu_i) = g(\mathrm{E}[Y_i]) = \eta_i = \boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j$ where the ith row of $\boldsymbol{X}$ is $\boldsymbol{x}_{i\cdot} = \left[x_{i1}, x_{i2}, \dots, x_{ip}\right]^T$.

1.3.1.1 Exponential family

The PDF of a random variable Y following a distribution in the exponential family[48] with canonical parameter θ and dispersion parameter ϕ is given by

$$f(y \mid \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}.$$

By equating the expectation of the derivative of $\log \mathcal{L}(\theta, \phi \mid y)$ with respect to the canonical parameter with zero

$$\mathrm{E}\left[\frac{\mathrm{d}\log \mathcal{L}(\theta, \phi \mid y))}{\mathrm{d}\theta}\right] = \frac{1}{a(\phi)}\mathrm{E}\left[y - \frac{\mathrm{d}b(\theta)}{\mathrm{d}\theta}\right] = 0$$

the mean $\mu = \mathrm{E}[Y]$ is given by

$$\mathrm{E}[Y] = \frac{\mathrm{d}b(\theta)}{\mathrm{d}\theta}. \tag{1}$$

The variance of Y is obtained by equating two definitions of the expected Fisher information

$$\begin{aligned} \mathrm{E}\left[\left(\frac{\mathrm{d}\log \mathcal{L}(\theta, \phi \mid y)}{\mathrm{d}\theta}\right)^2\right] &= \mathrm{E}\left[-\frac{\mathrm{d}^2\log \mathcal{L}(\theta, \phi \mid y)}{\mathrm{d}\theta^2}\right] \\ \frac{1}{a(\phi)^2}\mathrm{E}[(Y - \mathrm{E}[Y])^2] &= \frac{1}{a(\phi)}\frac{\mathrm{d}^2 b(\theta)}{\mathrm{d}\theta^2} \\ \hookrightarrow \mathrm{V}[Y] &= a(\phi)\frac{\mathrm{d}^2 b(\theta)}{\mathrm{d}\theta^2}. \end{aligned} \tag{2}$$

1.3.1.2 Maximum likelihood inference

The MLEs of the parameter vector $\boldsymbol{\beta}$ are obtained by Newton-Raphson's method[49]. This method requires a first-order Taylor expansion of the score vector, that is, the gradient ∇ of the joint log-likelihood function $\log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid y_i, \boldsymbol{x}_{i\cdot})$ centered on the current estimate $\boldsymbol{\beta}^*$

$$\nabla_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X}) \approx \nabla_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} + \nabla^2_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) .$$

Note that $\boldsymbol{\theta}$ is a function of the parameter vector $\boldsymbol{\beta}$ through the canonical link function $\boldsymbol{\theta} = g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. Equating the Taylor expansion to $\mathbf{0}$ and solving for $\boldsymbol{\beta}$ yields Newton-Raphson's method

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* + \left(-\nabla^2_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right)^{-1} \nabla_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$$

where $-\nabla^2_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ is the observed information matrix $\mathcal{J}(\boldsymbol{\beta}^*)$ and $\nabla_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ is the score vector. The method of Fisher scoring is obtained by replacing $\mathcal{J}(\boldsymbol{\beta}^*)$ with the expected information matrix

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* + \left(\mathrm{E}\left[-\nabla^2_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right]\right)^{-1} \nabla_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$$

For a GLM with the canonical link function, the score vector is

$$\nabla_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X}) = \begin{bmatrix} \sum_{i=1}^{n} \dfrac{\mathrm{d} \log \mathcal{L}(\theta_i, \phi \mid y_i, , \boldsymbol{x}_{i\cdot})}{d\theta_i} \dfrac{\mathrm{d}\theta_i}{\mathrm{d}\mu_i} \dfrac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} \dfrac{\partial \eta_i}{\partial \beta_1} \\ \vdots \\ \sum_{i=1}^{n} \dfrac{\mathrm{d} \log \mathcal{L}(\theta_i, \phi \mid y_i, , \boldsymbol{x}_{i\cdot})}{d\theta_i} \dfrac{\mathrm{d}\theta_i}{\mathrm{d}\mu_i} \dfrac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} \dfrac{\partial \eta_i}{\partial \beta_p} \end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} \frac{\partial \eta_1}{\partial \beta_1} & \cdots & \frac{\partial \eta_n}{\partial \beta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial \eta_1}{\partial \beta_p} & \cdots & \frac{\partial \eta_n}{\partial \beta_p} \end{bmatrix} \begin{bmatrix} \frac{\mathrm{d} \log \mathcal{L}(\theta_1, \phi \mid y_1, \boldsymbol{x}_{1\cdot})}{d\theta_1} \frac{\mathrm{d}\theta_1}{\mathrm{d}\mu_1} \frac{\mathrm{d}\mu_1}{\mathrm{d}\eta_1} \\ \vdots \\ \frac{\mathrm{d} \log \mathcal{L}(\theta_n, \phi \mid y_n, \boldsymbol{x}_{n\cdot})}{d\theta_n} \frac{\mathrm{d}\theta_n}{\mathrm{d}\mu_n} \frac{\mathrm{d}\mu_n}{\mathrm{d}\eta_n} \end{bmatrix} \\
&= \boldsymbol{X}^T \nabla_{\boldsymbol{\eta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})
\end{aligned}
$$

by the chain rule. The gradient $\nabla_{\boldsymbol{\eta}} log\, \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})$ can be simplified further as

$$
\begin{aligned}
a(\phi) \nabla_{\boldsymbol{\eta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X}) &= \begin{bmatrix} \left(y_1 - \frac{\mathrm{d}b(\theta_1)}{\mathrm{d}\theta_1}\right) \left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\theta_1} \frac{\mathrm{d}\eta_1}{\mathrm{d}\mu_1}\right)^{-1} \\ \vdots \\ \left(y_n - \frac{\mathrm{d}b(\theta_n)}{\mathrm{d}\theta_n}\right) \left(\frac{\mathrm{d}\mu_n}{\mathrm{d}\theta_n} \frac{\mathrm{d}\eta_n}{\mathrm{d}\mu_n}\right)^{-1} \end{bmatrix} \\
&= \begin{bmatrix} (y_1 - \mu_1) \left(\frac{\mathrm{d}^2 b(\theta_1)}{\mathrm{d}\theta_1^2} \frac{\mathrm{d}g(\mu_1)}{\mathrm{d}\mu_1}\right)^{-1} \\ \vdots \\ (y_n - \mu_n) \left(\frac{\mathrm{d}^2 b(\theta_n)}{\mathrm{d}\theta_n^2} \frac{\mathrm{d}g(\mu_n)}{\mathrm{d}\mu_n}\right)^{-1} \end{bmatrix} \\
&= \mathrm{diag}\left(\frac{\mathrm{d}^2 b(\theta_i)}{\mathrm{d}\theta_i^2} \left(\frac{\mathrm{d}g(\mu_i)}{\mathrm{d}\mu_i}\right)^2\right)^{-1} \begin{bmatrix} (y_1 - \mu_1) \frac{\mathrm{d}g(\mu_1)}{\mathrm{d}\mu_1} \\ \vdots \\ (y_n - \mu_n) \frac{\mathrm{d}g(\mu_n)}{\mathrm{d}\mu_n} \end{bmatrix} \\
&= \boldsymbol{W}\boldsymbol{z}\,.
\end{aligned}
$$

where Equation (2) was used in line 2. The expected Fisher information is thus

$$
\mathrm{E}\left[-\nabla_{\boldsymbol{\beta}}^2 \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\right] = \boldsymbol{X}^T E\left[-\nabla_{\boldsymbol{\eta}}^2 \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\right] \boldsymbol{X}
$$

where

$$
\begin{aligned}
& a(\phi)\mathrm{E}\left[-\nabla_{\boldsymbol{\eta}}^2 \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\right] \\
& = a(\phi)\mathrm{diag}\left(\mathrm{E}\left[-\frac{\mathrm{d}^2 \log \mathcal{L}(\theta_i, \phi \mid y_i, \boldsymbol{x}_{i\cdot})}{\mathrm{d}\theta_i^2}\right]\left(\frac{\mathrm{d}\mu_i}{\mathrm{d}\theta_i}\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}\right)^{-2}\right) \\
& = \mathrm{diag}\left(\frac{\mathrm{d}^2 b(\theta_i)}{\mathrm{d}\theta_i^2}\left(\frac{\mathrm{d}^2 b(\theta_i)}{\mathrm{d}\theta_i^2}\frac{\mathrm{d}g(\mu_i)}{\mathrm{d}\mu_i}\right)^{-2}\right) \\
& = \boldsymbol{W}
\end{aligned}
$$

noting that $\frac{\mathrm{d}^2 \log \mathcal{L}(\theta_i, \phi \mid y_i, \boldsymbol{x}_{i\cdot})}{\mathrm{d}\theta_j^2} = 0$ and using Equation (2) in line 2. The MLEs of the parameter vector $\boldsymbol{\beta}$ are thus obtained as

$$
\begin{aligned}
\boldsymbol{\beta} &= \boldsymbol{\beta}^* + \left(\mathrm{E}\left[-\nabla_{\boldsymbol{\beta}}^2 \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right]\right)^{-1} \\
&\quad \times \nabla_{\boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \\
&= \boldsymbol{\beta}^* + (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z} \\
&= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{z} + \boldsymbol{X}\boldsymbol{\beta}^*) \\
&= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\tilde{\boldsymbol{z}}\,,
\end{aligned}
\tag{3}
$$

which corresponds to Iteratively Reweighted Least Squares (IRLS) with Equation (3) as the normal equations.

1.3.1.3 Bayesian inference

Except for linear regression models, a common problem in Bayesian regression modeling is that the posterior distribution for the parameter vector $\boldsymbol{\beta}$ is intractable. In such a situation, a remedy is provided by Markov Chain Monte Carlo (MCMC) methods, which are a popular class of stochastic algorithms[50]. These methods generate a random sample that is approximately from the posterior distribution. The sample can be used for inference about the parameter vector $\boldsymbol{\beta}$. Approximating the posterior distribution with a normal distribution is a deterministic method that provides another solution to the problem.

The analytic expression of the normal approximation facilitates inference about the parameter vector $\boldsymbol{\beta}$.

In the following, MCMC methods such as the Metropolis-Hastings (MH) and Gibbs sampler are introduced. The section ends with a presentation of the normal approximation of the posterior distribution and how the approximation can be utilized within MCMC methods.

Markov Chain Monte Carlo sampling

The MH sampler[51] is a popular MCMC method to generate draws from a posterior distribution. The sampler uses a conditional proposal distribution of $\boldsymbol{\theta}'$ with conditional PDF $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$. The role of the proposal distribution is to suggest a candidate value $\boldsymbol{\theta}'$ given the previous draw $\boldsymbol{\theta}$. Either $\boldsymbol{\theta}'$ or $\boldsymbol{\theta}$ is accepted as the current draw of the sampler. There are different versions of the MH sampler depending on the actual implementation. The blockwise MH sampler is particularly useful when it is not feasible to specify a proposal distribution for the whole parameter vector, but it is possible to find a proposal distribution for each block. In this sampler, the jth candidate state $\boldsymbol{\theta}_j'$ is sampled from the jth proposal distribution with conditional PDF $q_j(\boldsymbol{\theta}_j' \mid \boldsymbol{\theta})$ and accepted with probability

$$\Pr(\boldsymbol{\theta}_j' \mid \boldsymbol{\theta}) = \min\left\{1, \frac{q_j(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_j', \boldsymbol{\theta}_{-j}) f(\boldsymbol{\theta}_j' \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y})}{q_j(\boldsymbol{\theta}_j' \mid \boldsymbol{\theta}_j, \boldsymbol{\theta}_{-j}) f(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y})}\right\},$$

where $\boldsymbol{\theta}_{-j}$ denotes the values of all parameter blocks except $\boldsymbol{\theta}_j$. Any proposal distribution may be chosen as long as it includes the support of the posterior distribution, simulation from it is feasible and the acceptance ratio can be computed.

The jth candidate state $\boldsymbol{\theta}_j'$ is always accepted if the full conditional distribution with PDF $f(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y})$ is chosen as the jth proposal density, because the acceptance probability is then equal to

$$\Pr(\boldsymbol{\theta}_j' \mid \boldsymbol{\theta}) = \min\left\{1, \frac{f(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y}) f(\boldsymbol{\theta}_j' \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y})}{f(\boldsymbol{\theta}_j' \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y}) f(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y})}\right\} = 1 .$$

This version of the MH sampler is called a Gibbs sampler[52]. Formal proofs for the block-wise MH and Gibbs samplers are given in [53].

It is often uncertain how many iterations are needed until a MH or Gibbs sampler provide a good approximation to the posterior distribution. The total number of iterations that are required until sampling from the posterior distribution occurs is called burn-in. A sampler that rapidly explores the parameter space has good mixing properties and requires less burn-in. The actual implementation of a MCMC sampler has great influence on the mixing properties. In practice, convergence is often assessed visually by plotting the generated draws of a parameter against their time index to create a trace plot. The trace plot demonstrates whether a sampler moves quickly away from the starting values and towards the support of the posterior distribution. This is most evident when the generated draws oscillate around a mean value. It should be noted that trace plots are a tool to determine the burn-in, but they can only aid in identifying non-convergence and not prove convergence.

Although MCMC samplers have facilitated Bayesian inference of regression models for GWAS[54], they are computationally expensive for current large data sets. This computational limitation can be resolved by using a normal distribution as an approximation to the posterior distribution for inference.

Normal approximation

Approximation of the posterior distribution by a normal distribution requires a second-order Taylor expansion of the log posterior PDF centered on the Maximum A Posteriori (MAP) estimate $\widetilde{\boldsymbol{\beta}}$, which equals the mode of the posterior PDF,

$$\begin{aligned}\log f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \phi) \\ &\approx \log f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \phi)|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} \\ &-\frac{1}{2}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)^{T} \nabla_{\boldsymbol{\beta}}^{2} \log f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \phi)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right).\end{aligned} \tag{4}$$

Assuming that $-\nabla_{\boldsymbol{\beta}}^{2} \log f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \phi)\big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}$ is positive definite, exponentiating Equation (4) shows that the posterior distribution is approximately

$$N\left(\boldsymbol{\beta} \mid \widetilde{\boldsymbol{\beta}}, \left[-\nabla_{\boldsymbol{\beta}}^{2} \log f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \phi)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}\right]^{-1}\right)$$

around the MAP estimate.

A normal approximation to the posterior distribution can also be combined with an independence MH sampler with proposal density $q(\boldsymbol{\theta}')$. In order to ensure rapid mixing, the proposal distribution is matched to the shape of the posterior distribution near the MAP estimate by using a heavy-tailed multivariate Student's *t*-distribution with low degrees of freedom, location parameter $\boldsymbol{\mu} = \widetilde{\boldsymbol{\beta}}$ and shape matrix $\boldsymbol{\Sigma} = \left[-\nabla_{\boldsymbol{\beta}}^{2} \log f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \phi)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}\right]^{-1}$. In the absence of prior information, the parameters of the proposal distribution are $\boldsymbol{\mu} = \widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma} = \left[\mathcal{J}(\widehat{\boldsymbol{\beta}})\right]^{-1}$.

1.3.1.4 Logistic regression model

Maximum likelihood inference

In logistic regression models, each y_i is thought to be an independent realization of a Bernoulli random variable $Y_i \in (0,1)$ with parameters $\mu_i = \Pr(Y_i = 1)$. The Bernoulli distribution is a member of the exponential family because the probability mass function of the Bernoulli distribution can be written as

$$f(y \mid \mu) = \mu^y (1-\mu)^{1-y} = \exp\left\{ y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu) \right\},$$

where $\theta = \log\left(\frac{\mu}{1-\mu}\right)$, $b(\theta) = \log\left(1+e^{\theta}\right)$, $a(\phi) = 1$ and $c(y,\phi) = 1$. Using Equation (1) and Equation (2), the mean and variance of Y are

$$\begin{aligned} \mathrm{E}[Y] &= \frac{\mathrm{d}\log\left(1+e^{\theta}\right)}{\mathrm{d}\theta} = \frac{e^{\theta}}{1+e^{\theta}} = \mu \\ \mathrm{V}[Y] &= \frac{\mathrm{d}^2\log\left(1+e^{\theta}\right)}{\mathrm{d}\theta^2} = \frac{e^{\theta}}{(1+e^{\theta})^2} = \mu(1-\mu). \end{aligned}$$

In logistic regression, $\mu_i = \mathrm{E}[Y_i] = \Pr(Y_i = 1)$ is thus modeled as a linear combination $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. Since $\eta_i \in (-\infty, \infty)$, a logistic function is applied to each η_i to constrain the value of the linear combination to the unit interval for any parameter values of $\boldsymbol{\beta}$. The canonical link function $\theta_i = g(\mu_i) = \eta_i$ for logistic regression is the logit function

$$\mathrm{logit}(x) = \log\left(\frac{x}{1-x}\right).$$

The IRLS weight matrix is

$$\begin{aligned} \boldsymbol{W} &= \mathrm{diag}\left(\frac{\mathrm{d}^2 b(\theta_i)}{\mathrm{d}\theta_i^2} \left(\frac{\mathrm{d}g(\mu_i)}{\mathrm{d}\mu_i} \right)^2 \right)^{-1} \\ &= \mathrm{diag}\left(\mathrm{V}[Y_i] \left[\frac{\mathrm{d}\log\left(\frac{\mu_i}{1-\mu_i}\right)}{\mathrm{d}\mu_i} \right]^2 \right)^{-1} \\ &= \mathrm{diag}\left(\mu_i(1-\mu_i)\right). \end{aligned}$$

The ith element of the adjusted outcome variable is

Algorithm 1: IRLS estimation for logistic regression

Input: Matrix with explanatory variables $\boldsymbol{X}$ and outcome variable $\boldsymbol{y}$

Output: MLE of $\boldsymbol{\beta}$ and their SEs

Initialize each element of the mean vector $\boldsymbol{\mu}$ and linear combination $\boldsymbol{\eta}$ as $\mu_i = (y_i + 0.5)/2$ and $\eta_i = \log\big(\mu_i(1-\mu_i)\big)$

while estimates of $\boldsymbol{\beta}$ change **do**

- Compute the diagonal weight matrix $\boldsymbol{W}$ with $w_i = \mu_i(1-\mu_i)$
- Calculate the adjusted outcome variable $\tilde{\boldsymbol{z}}$ with $\tilde{z}_i = (y_i - \mu_i)/w_i + \eta_i$
- Estimate the parameter vector as $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\tilde{\boldsymbol{z}}$
- Compute the linear combination $\boldsymbol{\eta}$ with $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ and the mean vector $\boldsymbol{\mu}$ with $\mu_i = 1/(1+e^{-\eta_i})$

end

Obtain the SEs of $\widehat{\boldsymbol{\beta}}$ as the square roots of the diagonal elements of $(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$

$$\tilde{z}_i = (y_i - \mu_i)\frac{\mathrm{d}g(\mu_i)}{\mathrm{d}\mu_i} + \boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta} = \frac{y_i - \mu_i}{\mu_i(1-\mu_i)} + \eta_i\,.$$

An algorithm for the IRLS method is shown in Algorithm 1.

Parameter interpretation

The logistic regression model relates the log-odds, which is the canonical parameter of the Bernoulli distribution,

$$\operatorname{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \log\left(\frac{\Pr(Y_i = 1)}{1 - \Pr\,(Y_i = 1)}\right)$$

to the linear combination $\eta_i = \boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}$. The effect of explanatory variables on the log-odds is linear on logarithmic scale but multiplicative on the original scale. The logistic regression equation is

$$\operatorname{logit}(\mu_i) = \beta_1 + \sum_{j=2}^{p} x_{ij}\beta_j\,,$$

which indicates that β_1 represents the log-odds when all explanatory variables are equal to zero. Equivalently, e^{β_1} denotes the odds ratio when all explanatory variables are equal to zero.

For continuous variables of interest, $\beta_2, \ldots, \beta_p$ represent the change in log-odds for an increase of the value c of the respective explanatory variable x_{ij^*} by one unit while other explanatory variables are held constant because

$$\begin{aligned}
&\operatorname{logit}\left(\mu_i\left(x_{ij^*} = c+1\right)\right) - \operatorname{logit}\left(\mu_i\left(x_{ij^*} = c\right)\right) \\
&\qquad = \beta_1 + \beta_{j^*}(c+1) + \sum_{\{2,\ldots,p\}\setminus j^*} x_{ij}\beta_j - \beta_1 - \beta_{j^*}c - \sum_{\{2,\ldots,p\}\setminus j^*} x_{ij}\beta_j \\
&\qquad = \beta_{j^*} .
\end{aligned}$$

On the original scale

$$\mu_i\left(x_{ij^*} = c+1\right) = e^{\beta_{j^*}}\mu_i\left(x_{ij^*} = c\right)$$

so that $e^{\beta_{j^*}}$ denotes the change in odds corresponding to an increase of the value of the explanatory variable x_{ij^*} by one unit. The sign of β_{j^*} indicates whether an increase of x_{ij^*} by one unit is associated with an increase or decrease in odds.

For a categorical explanatory variable with K levels, dummy variables $d_1, d_2 \ldots, d_K$ are defined with $d_{ij} = 1$ if the value of the ith sample of the categorical explanatory variable is at the jth level, otherwise $d_{ij} = 0$. The first dummy variable d_1 is not needed if the remaining $K-1$ dummy variables are coded relative to the first level. This corresponds to the set-to-zero constraint for the regression parameter α_1 of the first dummy variable. The constraint is applied to make the regression parameters identifiable because adding a constant c to the intercept β_0 and subtracting it from α_1

$$\text{logit}(\mu_i) = (\beta_1 + c) + (\alpha_1 - c)d_{i1} + \sum_{j=2}^{K} d_{ij}\alpha_j + \sum_{j=2}^{p} x_{ij}\beta_j$$

results in a model with the same value for the $\text{logit}(\mu_i)$ for any c. The logistic regression equation with the set-to-zero constraint under the common slope formulation with one categorical explanatory variable is

$$\text{logit}(\mu_i) = \begin{cases} \beta_1 + \sum_{j=2}^{p} x_{ij}\beta_j & \text{if } d_{i2} = \cdots = d_{iK} = 0 \\ \beta_1 + \alpha_2 + \sum_{j=2}^{p} x_{ij}\beta_j & \text{if } d_{i2} = 1 \\ \vdots & \vdots \\ \beta_1 + \alpha_K + \sum_{j=2}^{p} x_{ij}\beta_j & \text{if } d_{iK} = 1 \end{cases} .$$

which takes the form of $K - 1$ regression lines with different intercepts but common slopes. The assumption of common slopes implies that a continuous explanatory variable has the same effect on the outcome variable regardless of the level of the categorical explanatory variable. As a result of setting $\alpha_1 = 0$, the intercept β_1 represents the log-odds for the first level of the categorical explanatory variable if all continuous explanatory variables are zero. The parameter α_ℓ is the change in log-odds for the ℓth level compared to the first level of the categorical explanatory variable while continuous explanatory variables are held constant because

$$\begin{aligned} &\text{logit}\big(\mu_i(d_{i\ell} = 1)\big) - \text{logit}\big(\mu_i(d_{i2} = \cdots = d_{i\ell} = \cdots = d_{iK} = 0)\big) \\ &\quad = \beta_1 + \alpha_\ell + \sum_{j=2}^{p} x_{ij}\beta_j - \beta_1 - \sum_{j=2}^{p} x_{ij}\beta_j = \alpha_\ell \,. \end{aligned}$$

In the presence of a categorical explanatory variable, the parameter β_j represents the change in log-odds for an increase of the jth variable of interest while the categorical explanatory variable is fixed at some level ℓ and all other continuous variables of interest are held constant because

$$\begin{aligned}\operatorname{logit}\left(\mu_i\left(x_{ij^*} = c+1, d_{i\ell} = 1\right)\right) - \operatorname{logit}\left(\mu_i\left(x_{ij^*} = c, d_{i\ell} = 1\right)\right) \\ = \beta_1 + \alpha_\ell + \beta_{j^*}(c+1) + \sum_{\{2,\dots,p\}\setminus j^*} x_{ij}\beta_j - \beta_1 - \alpha_\ell - \beta_{j^*}c \\ - \sum_{\{2,\dots,p\}\setminus j^*} x_{ij}\beta_j = \beta_{j^*}\,.\end{aligned}$$

The effect of multiple categorical explanatory variables on the log-odds is additive under the common slope formulation. The change in log-odds for some levels ℓ_1 and ℓ_2 compared to the baseline of two categorical explanatory variables while holding continuous explanatory variables constant is

$$\begin{aligned}\operatorname{logit}\left(\mu_i\left(d^1_{i\ell_1} = 1, d^2_{i\ell_2} = 1\right)\right) \\ - \operatorname{logit}\Big(\mu_i\big(d^1_{i2} = \cdots = d^1_{i\ell_1} = \cdots = d^1_{iK} = 0, d^2_{i2} = \cdots = d^2_{i\ell_1} = \cdots \\ = d^2_{iK} = 0\big)\Big) = \beta_1 + \alpha_{\ell_1} + \alpha_{\ell_2} + \sum_{j=2}^{p} x_{ij}\beta_j - \beta_1 - \sum_{j=2}^{p} x_{ij}\beta_j \\ = \alpha_{\ell_1} + \alpha_{\ell_2}\,,\end{aligned}$$

where $d^1_{\ell_1}$ and $d^2_{\ell_2}$ denote the ℓ_1th and ℓ_2th dummy variable of the first and second categorical explanatory variable and α_{ℓ_1} and α_{ℓ_2} the respective regression parameters.

Example

Note that the following is a continuation of the example given in Section 1.1.4 but rs62470411 is used in place of rs1157745. Figure 5 shows the distribution of PON1

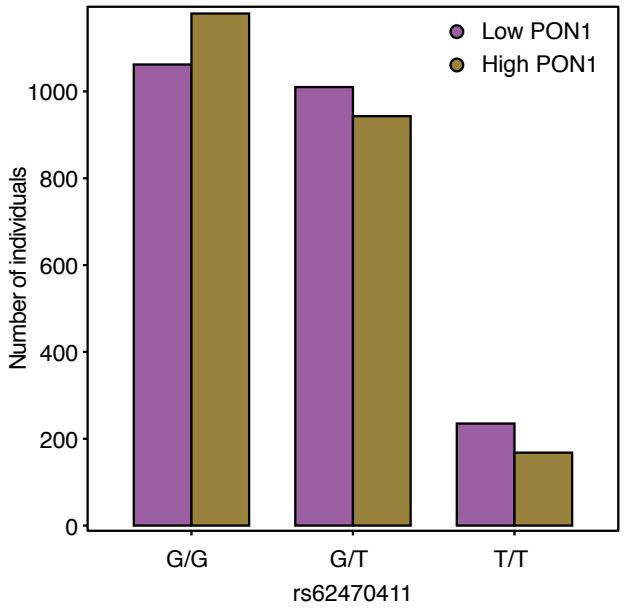


Figure 5. Distribution of low/high PON1 serum levels among three possible genotypes at rs62470411 among 4,613 individuals from the FINRISK study. Quantitative measurements of PON1 serum levels were dichotomized according to the median PON1 serum level among the individuals.

serum levels of 4,613 individuals from the FINRISK study among three possible genotypes at rs62470411. Assume that there is a hypothetical test that can only classify individuals into low/high PON1 serum levels and that outcomes from such a test are available by dichotomizing according to the median PON1 serum level among the 4,613 individuals.

Logistic regression is applied to study the relationship between rs62470411 and dichotomized PON1 serum levels. The outcome variable y_i is one if the ith individual has PON1 serum levels above the median in the data and zero otherwise. The value of the variable of interest x_i is the count of the number of copies of the *T* allele. Table 2 shows the results from IRLS estimation. The change in odds of having higher PON1 serum levels is $e^{\beta 1} = 0.811$ for each copy of the *T* allele. Compared to individuals with two copies of the G allele, the odds of having higher serum levels is 19% lower in individuals with one copy of the T allele and 34% lower in individuals with two copies of the *T* allele. This result is statistically significant at the 10^{-5} level according to a Wald statistical test.

Table 2. IRLS estimation for logistic regression for rs62470411 and dichotomized PON1 serum levels among 4,613 individuals from the FINRISK study.

Iteration	β_1	SE of β_1	P-value† for β_1
1	-0.1395	0.0527	0.0081917
2	-0.1539	0.0639	0.0160650
3	-0.1873	0.0498	0.0001700
4	-0.2095	0.0461	0.0000055
5	-0.2101	0.0459	0.0000048

† Wald test for H_0: $\beta_1 = 0$

1.3.1.5 Linear regression model

In linear regression models, each y_i is thought to be an independent realization of a normal random variable $Y_i \in (-\infty, \infty)$ with parameters $\mu_i \in \mathbb{R}$ and $\sigma^2 > 0$. The normal distribution is a member of the exponential family because the probability density function of the normal distribution can be written as

$$f(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} = \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\},$$

where $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$ and $c(y, \phi) = -y^2/2\sigma^2 - \log(2\pi\sigma^2)/2$. The mean and variance are readily available as $\mathrm{E}[Y] = \mu$ and $\mathrm{V}[Y] = \sigma^2$. The expectation $\mathrm{E}[Y_i] = \mu_i$ is thus modeled as a linear combination $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. The linear combination η_i can take any value in $(-\infty, \infty)$ and therefore the link function $g(\cdot)$ is the identity link.

The IRLS weight matrix is

$$\boldsymbol{W} = \mathrm{diag}\left(\frac{\mathrm{d}^2 b(\theta_i)}{\mathrm{d}\theta_i^2}\left(\frac{\mathrm{d}g(\mu_i)}{\mathrm{d}\mu_i}\right)^2\right)^{-1} = \mathrm{diag}\left(\mathrm{V}[Y_i]\left[\frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_i}\right]^2\right)^{-1} = \sigma^2\boldsymbol{I}_n\,.$$

The ith element of the adjusted outcome variable is

$$\tilde{z}_i = (y_i - \mu_i)\frac{\mathrm{d}g(\mu_i)}{\mathrm{d}\mu_i} + \boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta} = y_i\,.$$

The normal equations are thus given by

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\tilde{\boldsymbol{z}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{5}$$

and can be solved noniteratively to obtain $\widehat{\boldsymbol{\beta}}$. Equation (5) is solved by computing the QR decomposition[55] of $\boldsymbol{X}$. This decomposition yields $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q}$ is an orthogonal matrix of dimension $n \times p$ and $\boldsymbol{R}$ is an upper triangular matrix of dimension $p \times p$. Substituting this decomposition into Equation (5) results in

$$\begin{aligned}
(\boldsymbol{Q}\boldsymbol{R})^T(\boldsymbol{Q}\boldsymbol{R})\boldsymbol{\beta} &= (\boldsymbol{Q}\boldsymbol{R})^T\boldsymbol{y} \\
\boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R}\boldsymbol{\beta} &= \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{y} \\
\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{\beta} &= \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{y} \\
(\boldsymbol{R}\boldsymbol{R}^{-1})^T\boldsymbol{R}\boldsymbol{\beta} &= \boldsymbol{Q}^T\boldsymbol{y} \\
\hookrightarrow \widehat{\boldsymbol{\beta}} &= \boldsymbol{R}^{-1}\boldsymbol{z}\,,
\end{aligned}$$

which is an upper triangular system of equations that can be solved by back substitution.

Parameter interpretation

The linear regression model relates the expectation $\mu_i = \mathrm{E}[Y_i]$ to the linear combination $\eta_i = \boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}$. The effect of explanatory variables on the expectation of Y_i is linear. The linear regression equation is

$$\mu_i = \mathrm{E}[Y_i] = \beta_1 + \sum_{j=2}^{p} x_{ij}\beta_j\,,$$

which indicates that β_1 represents the expectation of Y_i when all explanatory variables are equal to zero.

For continuous explanatory variables, $\beta_2, \dots, \beta_p$ represent the change in the expectation of Y_i for an increase of the value c of the respective explanatory variable x_{ij^*} by one unit while other explanatory variables are held constant because

$$\begin{aligned}\mu_i(x_{ij^*} = c+1) - \mu_i(x_{ij^*} = c) \\ &= \beta_1 + \beta_{j^*}(c+1) + \sum_{\{2,\dots,p\}\setminus j^*} x_{ij}\beta_j - \beta_1 - \beta_{j^*}c - \sum_{\{2,\dots,p\}\setminus j^*} x_{ij}\beta_j \\ &= \beta_{j^*}\,.\end{aligned}$$

The sign of β_{j^*} indicates whether an increase of x_{ij^*} by one unit is associated with an increase or decrease in the expectation of Y_i.

For a categorial explanatory variable with K levels, the linear regression equation with the set-to-zero constraint under the common slope formulation with one categorical explanatory variable is

$$\mu_i = \begin{cases} \beta_1 + \sum_{j=2}^{p} x_{ij}\beta_j & \text{if } d_{i2} = \cdots = d_{iK} = 0 \\ \beta_1 + \alpha_2 + \sum_{j=2}^{p} x_{ij}\beta_j & \text{if } d_{i2} = 1 \\ \quad\vdots & \quad\vdots \\ \beta_1 + \alpha_K + \sum_{j=2}^{p} x_{ij}\beta_j & \text{if } d_{iK} = 1 \end{cases}.$$

which takes the form of $K-1$ regression lines with different intercepts but common slopes. The intercept β_1 represents the expectation of Y_i for the first level of the categorical explanatory variable if all continuous explanatory variables are zero. The parameter α_ℓ is the change in the expectation of Y_i for the ℓth level compared to the first level of

the categorical explanatory variable while the continuous explanatory variables are held constant because

$$\mu_i(d_{i\ell} = 1) - \mu_i(d_{i2} = \cdots = d_{i\ell} = \cdots = d_{iK} = 0)$$
$$= \beta_1 + \alpha_\ell + \sum_{j=2}^{p} x_{ij}\beta_j - \beta_1 - \sum_{j=2}^{p} x_{ij}\beta_j = \alpha_\ell \,.$$

In the presence of an explanatory variable, the parameter β_j represents the change in the expectation of Y_i for an increase of the jth explanatory variable while the categorical explanatory variable is fixed at some level ℓ and all other continuous explanatory variables are held constant because

$$\mu_i\big(x_{ij^*} = c + 1, d_{i\ell} = 1\big) - \mu_i\big(x_{ij^*} = c, d_{i\ell} = 1\big)$$
$$= \beta_1 + \alpha_\ell + \beta_{j^*}(c + 1) + \sum_{\{2,\ldots,p\}\setminus j^*} x_{ij}\beta_j - \beta_1 - \alpha_\ell - \beta_{j^*}c$$
$$- \sum_{\{2,\ldots,p\}\setminus j^*} x_{ij}\beta_j = \beta_{j^*} \,.$$

The effect of multiple categorical explanatory variables on the expectation of Y_i is additive under the common slope formulation. The change in the expectation of Y_i for some levels ℓ_1 and ℓ_2 compared to the baseline of two categorical explanatory variables while holding continuous explanatory variables constant is

$$\mu_i\big(d^1_{i\ell_1} = 1, d^2_{i\ell_2} = 1\big)$$
$$- \mu_i\big(d^1_{i2} = \cdots = d^1_{i\ell_1} = \cdots = d^1_{iK} = 0, d^2_{i2} = \cdots = d^2_{i\ell_1} = \cdots = d^2_{iK}$$
$$= 0\big) = \beta_1 + \alpha_{\ell_1} + \alpha_{\ell_2} + \sum_{j=2}^{p} x_{ij}\beta_j - \beta_1 - \sum_{j=2}^{p} x_{ij}\beta_j = \alpha_{\ell_1} + \alpha_{\ell_2} \,.$$

where $d^1_{\ell_1}$ and $d^2_{\ell_2}$ denote the ℓ_1th and ℓ_2th dummy variable of the first and second categorical explanatory variable and α_{ℓ_1} and α_{ℓ_2} the respective regression parameters.

Example

The following is a continuation of the example in Section 1.3.1.4. Sometimes linear regression is applied in the case/control setting to study the effect of a variant on a disease phenotype, even if the outcome variable is not continuous. Using linear regression instead of logistic regression offers computational benefits and works well[56] if 1) the effect of the variant on the disease in terms of odds ratio is smaller than 1.3, 2) the proportion of cases for the disease is between 0.3 and 0.7, and 3) the MAF is greater than 0.05.

The P-value from a Wald statistical test that rs62470411 has no effect on dichotomized PON1 serum levels is practically the same in linear and logistic regression (Table 3). However, the interpretation of the effect size from linear regression is difficult because the estimate is not on the log-odds scale. Turning the effect size estimate from linear regression into an estimate on the log-odds scale can be done by using the following approximation[56]

$$\hat{\theta}_{\text{log-odds}} \approx \left[\frac{(1-2\phi)\left(1-2\hat{P}\right)}{2} + \frac{\phi(1-\phi)}{\hat{\theta}_{\text{linear}}} - \hat{\theta}_{\text{linear}}\frac{0.084 + 0.9\phi(1-\phi)\hat{P}\left(1-\hat{P}\right)}{\phi(1-\phi)}\right]^{-1},$$

where ϕ is the proportion of cases for the disease and $\hat{P}$ is the frequency of the reference allele in the data. The approximation works well for rs62470411, because the case proportion is about 0.5, the MAF is 0.3 and the effect size estimate on log-odds scale is small.

Table 3. IRLS estimation for linear regression for rs62470411 and dichotomized PON1 serum levels among 4,613 individuals from the FINRISK study

	β_1	SE of β_1	P-value† for β_1
Linear regression	-0.05233	0.01140	0.00000457
Logistic regression	-0.21007	0.04595	0.00000484
Approximation	-0.21009	0.04562	0.00000412

† Wald test for H_0: $\beta_1 = 0$

1.4 Statistical variable selection

High-throughput technologies such as DNA microarrays enable genotyping of hundreds of thousands of variants. In GWAS, the genotype-phenotype data on thousands of individuals is used to test each variant independently for a statistical association with a phenotype. While GWAS is informative about which genomic regions are associated with the phenotype, it typically fails to reveal the underlying biology or suggest potential drug targets due to the large number of candidate variants in strong LD.

The top panel in Figure 6 shows the negative $\log_{10}$ P-values from a PON1 GWAS on 4,613 individuals from the FINRISK study for variants in a 29 kilobase region around the *PON1* gene. Variants with similar magnitude of correlation with the variant with the smallest P-value in the region reach almost identical statistical significance. The bottom panel shows the absolute value of pairwise sample Pearson correlations between the variants in the region. Clearly, many variants between positions 94,952,431 and 95,054,011 are in high LD with each other and reach similar statistical significance. Since these variants are not in strong LD with the variant at 94,941,038 that has the smallest P-value, there could be statistical evidence for multiple causal variants. Fine-mapping is the post-GWAS statistical analysis used to identify causal variants in genomic regions in order to prioritize variants for follow-up analysis.

There is also a rich literature on BVS[62] such as Gibbs Variable Selection[63], Stochastic Search Variable Selection[64] and the sampler of Kuo and Mallick[65]. In the Bayesian framework, the aim is to obtain a probabilistic description about the importance of explanatory variables rather than choosing a single set; this is also an important reason to use the framework in fine-mapping of GWAS regions.

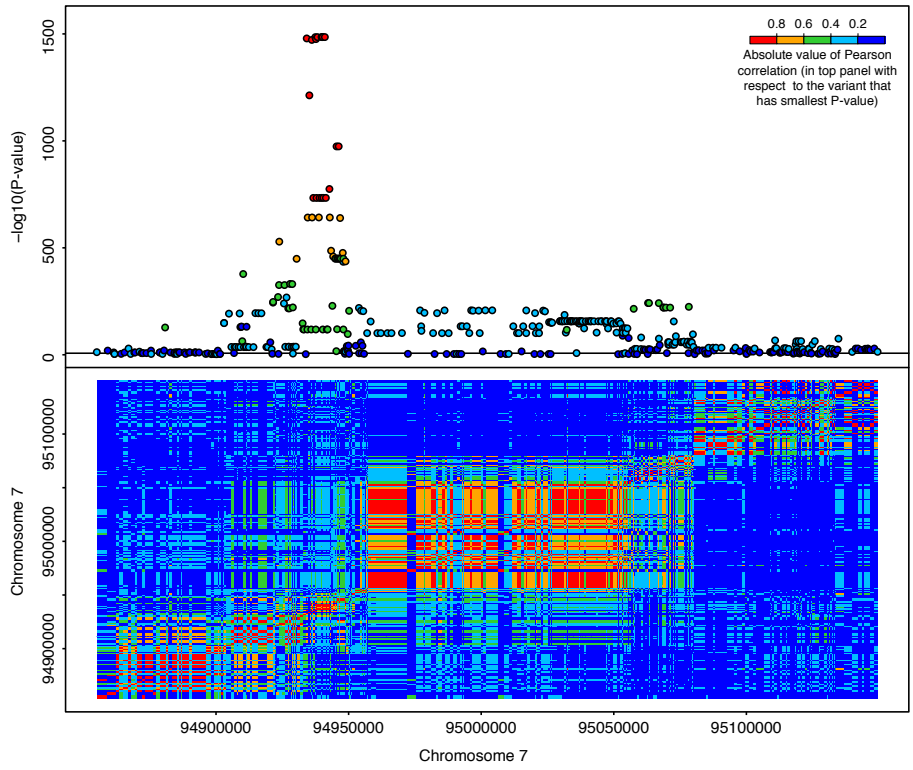


Figure 6. Regional plot for PON1 serum level GWAS on 4,613 individuals from the FINRISK study. Top panel shows negative $\log_{10}$ P-values (y-axis) from independent linear regression of each variant in a 0.29 mega base region around the *PON1* gene against the variant's genomic position along chromosome 7 (x-axis). Variants are colored with respect to their absolute value of Pearson correlation with the variant that has smallest P-value in the region. Bottom panel shows the absolute value of Pearson's correlations among variants in the region.

The advantage of stochastic search in BVS compared to exhaustive search is the computational savings that results from fitting only some of the most probable models. Interestingly, BVS implemented in recent fine-mapping methods CAVIAR, PAINTOR and CAVIARBF rely on computationally expensive exhaustive search.

1.4.1 Bayesian fine-mapping methods

The following is an outline of the statistical model implemented in fine-mapping methods CAVIAR, CAVIARBF, PAINTOR and FINEMAP. Although CAVIAR,

CAVIARBF and PAINTOR are useful fine-mapping methods, they rely on computationally expensive exhaustive search that restricts their usefulness in practice. This computational limitation motivated the development of FINEMAP, which uses more efficient stochastic search. Some additional fine-mapping methods that have been published after FINEMAP are described in the Discussion.

1.4.1.1 CAVIAR (2014) and CAVIARBF (2015)

GWAS for a quantitative trait relies on individual linear regression for each variant. Fine-mapping methods extend the statistical model to multiple linear regression. CAVIAR[19,66] and CAVIARBF[67] consider the following multiple linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\lambda}_c + \boldsymbol{\epsilon}\,,$$

where $\boldsymbol{y}$ is a standardized vector of length n with values of a quantitative trait, $\boldsymbol{X}$ is a genotype matrix at m variants with standardized columns of length n, $\boldsymbol{\lambda}_c$ is a effect size vector of length m that is indexed by a binary indicator vector $\boldsymbol{c}$ such that $\lambda_{c_\ell} \neq 0$ if the ℓth element c_ℓ equals one and $\lambda_{c_\ell} = 0$ if $c_\ell = 0$, and $\boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon} \mid 0, \sigma^2 \boldsymbol{I}_n)$, where $\hat{\sigma}^2 \approx \mathrm{V}(\boldsymbol{y}) = 1$ in the following steps by assuming that variants have small effect sizes typical for GWAS.

The MLE for $\boldsymbol{\lambda}_c$ can be computed from the sample correlation matrix $\boldsymbol{\Sigma} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ and z-scores $s_\ell = \hat{\beta}_\ell \times \mathrm{V}\left[\hat{\beta}_\ell\right]^{-1/2} = n^{-1/2}\boldsymbol{x}_{\cdot\ell}^T\boldsymbol{y}$ from individual linear regression of each variant

$$\hat{\boldsymbol{\lambda}}_c = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = n^{-1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\,,$$

where $\boldsymbol{S} = (s_1, s_2, \ldots, s_m)^T$. Asymptotically, the likelihood function $\mathcal{L}(\boldsymbol{\lambda}_c \mid \boldsymbol{y}, \boldsymbol{X})$ is proportional to the PDF $N\left(\hat{\boldsymbol{\lambda}}_c \mid \boldsymbol{\lambda}_c, \mathrm{V}\left[\hat{\boldsymbol{\lambda}}_c\right]\right)$ of a normal distribution where $\mathrm{V}\left[\hat{\boldsymbol{\lambda}}_c\right] =$

$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = n^{-1}\boldsymbol{\Sigma}^{-1}$. Using the affine transformation $\boldsymbol{S} = \sqrt{n}\boldsymbol{\Sigma}\hat{\boldsymbol{\lambda}}_c$, the likelihood function can also be expressed as $\mathcal{L}(\boldsymbol{\lambda}_c \mid \boldsymbol{y}, \boldsymbol{X}) \propto N\left(\boldsymbol{S} \mid \sqrt{n}\,\boldsymbol{\Sigma}\boldsymbol{\lambda}_c, \boldsymbol{\Sigma}\right)$.

CAVIAR specifies a prior distribution for $\boldsymbol{\lambda}_c$ with PDF $N(\boldsymbol{\lambda}_c \mid \boldsymbol{0}, \boldsymbol{\Sigma}_c)$ where $\boldsymbol{\Sigma}_c = \mathrm{diag}\left(\sigma^2_{c_1}, \sigma^2_{c_2}, \ldots, \sigma^2_{c_m}\right)$ with $\sigma^2_{c_\ell} = \sigma^2_g$ if the ℓth variant is causal and $\sigma^2_{c_\ell} \approx 10^{-4}$, otherwise. The effect size vector $\boldsymbol{\lambda}_c$ is not of interest in identifying the causal status of variants and therefore is integrated out. The marginal likelihood for the causal status vector $\boldsymbol{c}$ is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{c} \mid \boldsymbol{y}, \boldsymbol{X}) &= \int \mathcal{L}(\boldsymbol{\lambda}_c \mid \boldsymbol{y}, \boldsymbol{X}) f(\boldsymbol{\lambda}_c \mid \boldsymbol{c})\, \mathrm{d}\boldsymbol{\lambda}_c \\ &= \int N\left(\boldsymbol{S} \mid \sqrt{n}\boldsymbol{\Sigma}\boldsymbol{\lambda}_c, \boldsymbol{\Sigma}\right) N(\boldsymbol{\lambda}_c \mid \boldsymbol{0}, \boldsymbol{\Sigma}_c)\, \mathrm{d}\boldsymbol{\lambda}_c \\ &= N(\boldsymbol{S} \mid \boldsymbol{0}, \boldsymbol{\Sigma} + n\boldsymbol{\Sigma}\boldsymbol{\Sigma}_c\boldsymbol{\Sigma})\,.\end{aligned}$$

Computation of unnormalized posterior probabilities $\Pr^*(\boldsymbol{c} \mid \boldsymbol{y}, \boldsymbol{X}) \propto \mathcal{L}(\boldsymbol{c} \mid \boldsymbol{y}, \boldsymbol{X}) \Pr(\boldsymbol{c})$ requires prior probabilities $\Pr(\boldsymbol{c})$ for each causal status vector. CAVIAR assumes that each variant has the same prior probability of being causal leading to the following prior probability

$$\Pr(\boldsymbol{c}) = \prod_{\ell=1}^{m} \left(\frac{1}{m}\right)^{c_\ell} \left(\frac{m-1}{m}\right)^{1-c_\ell} = \left(\frac{1}{m}\right)^{k} \left(\frac{m-1}{m}\right)^{m-k},$$

when $k = \sum_{\ell=1}^{m} c_\ell$.

One of the novel concepts introduced by CAVIAR is the computation of the likelihood function $\mathcal{L}(\boldsymbol{\lambda}_c \mid \boldsymbol{y}, \boldsymbol{X}) \propto N\left(\boldsymbol{S} \mid \sqrt{n}\boldsymbol{\Sigma}\boldsymbol{\lambda}_c, \boldsymbol{\Sigma}\right)$ on the basis of summary-level data $(\boldsymbol{S}, \boldsymbol{\Sigma})$ instead of individual-level data $(\boldsymbol{y}, \boldsymbol{X})$. However, each likelihood evaluation requires $O(m^3)$ operations. One of the novelties in CAVIARBF is that the Bayes Factor $\mathrm{BF}(\boldsymbol{c} : \boldsymbol{c}_0) = \mathcal{L}(\boldsymbol{c} \mid \boldsymbol{y}, \boldsymbol{X}) / \mathcal{L}(\boldsymbol{c}_0 \mid \boldsymbol{y}, \boldsymbol{X})$ can be computed with $O(k^3)$ operations by using only data on variants that are assumed to be causal according to $\boldsymbol{c}$. This insight was obtained by shrinking ϵ in the PDF of the prior distribution $N(\boldsymbol{\lambda}_c \mid \boldsymbol{0}, \boldsymbol{\Sigma}_c)$ towards zero.

The mathematical derivation of this result was not shown by CAVIARBF, but is given in Article I and as a different version in Section 3.2.5.

1.4.1.2 PAINTOR (2014)

PAINTOR is similar to CAVIAR and CAVIARBF, but also allows for joint fine-mapping of multiple genomic regions and incorporation of annotations for the variants through the prior distribution of the causal status vector to improve fine-mapping accuracy.

For brevity, assume that there is only a single genomic region. PAINTOR treats the causal status vector as missing data and obtains the incomplete data likelihood function by summing the complete data likelihood function over all possible causal status vectors

$$\mathcal{L}(\boldsymbol{\lambda}_c, \boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{A}) = \sum_{\boldsymbol{c}} f(\boldsymbol{y}, \boldsymbol{c} \mid \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c, \boldsymbol{\gamma}) = \sum_{\boldsymbol{c}} f(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c, \boldsymbol{c}, \boldsymbol{\gamma}) f(\boldsymbol{c} \mid \boldsymbol{A}, \boldsymbol{\gamma}),$$

where $f(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c, \boldsymbol{c}, \boldsymbol{\gamma}) = N\left(\boldsymbol{S} \mid \sqrt{n}\boldsymbol{\Sigma}\boldsymbol{\lambda}_c, \boldsymbol{\Sigma}\right)$ and $\boldsymbol{\gamma}$ are the annotation effect sizes. The prior distribution for $\boldsymbol{c}$ is specified through a standard logistic regression model as

$$f(\boldsymbol{c} \mid \boldsymbol{A}, \boldsymbol{\gamma}) = \prod_{\ell} \left[\frac{1}{1 + \exp\left(-\boldsymbol{a}_{\ell\cdot}^T \boldsymbol{\gamma}\right)}\right]^{c_\ell} \left[\frac{1}{1 + \exp\left(\boldsymbol{a}_{\ell\cdot}^T \boldsymbol{\gamma}\right)}\right]^{1-c_\ell},$$

where the ℓth row of $\boldsymbol{A}$ is the binary annotation vector $\boldsymbol{a}_{\ell\cdot} = [a_{\ell 1}, a_{\ell 2}, \ldots, a_{\ell d}]^T$ with elements being equal to one if the ℓth variant has the annotation and zero, otherwise.

Computation of the posterior probability $\Pr(c_\ell = 1 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c, \boldsymbol{\gamma})$ for each variant requires an Expectation Maximization (EM) algorithm[68]. This involves iteratively computing $\Pr(c_\ell = 1 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c', \boldsymbol{\gamma}')$ given current estimates $(\boldsymbol{\lambda}_c', \boldsymbol{\gamma}')$ followed by maximization of an objective function $Q(\boldsymbol{\lambda}_c, \boldsymbol{\gamma} \mid \boldsymbol{\lambda}_c', \boldsymbol{\gamma}')$ to obtain new parameter estimates for $(\boldsymbol{\lambda}_c', \boldsymbol{\gamma}')$ given newly computed posterior probabilities for the causal status of

the variants. The posterior probability $\Pr(c_\ell = 1 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c', \boldsymbol{\gamma}')$ for each variant is computed by exhaustive enumeration of all possible causal status vectors and their posterior probability

$$\Pr(\boldsymbol{c} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c', \boldsymbol{\gamma}') = \frac{f(\boldsymbol{y}, \boldsymbol{c} \mid \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c', \boldsymbol{\gamma}')}{\sum_{\boldsymbol{c}} f(\boldsymbol{y}, \boldsymbol{c} \mid \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c', \boldsymbol{\gamma}')}.$$

New parameter estimates are subsequently computed by maximizing the following objective function

$$\begin{aligned} Q(\boldsymbol{\lambda}_c, \boldsymbol{\gamma} \mid \boldsymbol{\lambda}_c', \boldsymbol{\gamma}') &= \sum_{\boldsymbol{c}} \Pr(\boldsymbol{c} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c', \boldsymbol{\gamma}') \times \log f(\boldsymbol{y}, \boldsymbol{c} \mid \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\lambda}_c', \boldsymbol{\gamma}') \\ &= \sum_{\boldsymbol{c}} Q(\boldsymbol{\lambda}_c \mid \boldsymbol{\lambda}_c') + Q(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}'), \end{aligned}$$

where $\boldsymbol{\lambda}_c'$ is fixed to the z-scores in PAINTOR during maximization.

2 AIMS OF THE STUDY

(1) Existing fine-mapping methods rely on computationally expensive exhaustive search that restricts their use to only a few hundred variants. Since GWAS regions can span several mega bases and contain thousands of variants, improving the computational efficiency of fine-mapping methods is paramount to facilitate extraction of valuable information from GWAS regions which could otherwise remain undetected due to computational limitations.

- Aim 1 is to scale up fine-mapping methods to genomic regions with thousands of variants while maintaining the accuracy of gold standard exhaustive search.

(2) Fine-mapping methods that work directly on GWAS results also require LD estimates as input. All existing fine-mapping methods can use LD estimates from publicly available reference genotype panels such as the 1000 Genomes Project or the Haplotype Reference Consortium. The hope has been that such LD estimates perform well, but the impact of these estimates has not been comprehensively studied.

- Aim 2 is to investigate how LD estimates from reference genotype panels perform in fine-mapping analysis in comparison with LD estimates from the original individual-level GWAS genotype data.

(3) The output from Bayesian fine-mapping methods are typically a list of possible configurations of variants and their posterior probabilities, as well the posterior probability of causality for each variant. These probabilities contain all the information needed for follow-up downstream analysis. Examples of useful downstream analyses are estimation of effect sizes of variants and phenotypic variance explained by fine-mapped variants.

- Aim 3 is to explore whether large-scale GWAS sample sizes in biobank studies can provide opportunities to routinely estimate the heritability for GWAS regions with a fine-mapping model compared to using a computationally expensive variance component model.

3 MATERIALS AND METHODS

3.1 Cohorts

In article I, we used data from the FINRISK study and the WTCCC2. In article II, we used data from the FINRISK study, the 1966 Northern Finland Birth Cohort (NFBC1966) and the UK Biobank (UKBB). In article III, we used data from the FINRISK study and UKBB.

FINRISK is a representative, cross-sectional survey of the Finnish working-age population. Since 1972, a random sample of 6,000–8,000 individuals has been collected every 5 years for the study of risk factors of chronic diseases. The study protocols of the FINRISK surveys used in this work (1992, 1997, 2002, 2007, and 2012) were approved by the ethics committee of the National Public Health Institute until 1997 and by the ethics committee of Helsinki and Uusimaa Hospital District after that. All participants of FINRISK have provided written informed consent.

NFBC1966 is a longitudinal study of individuals from the provinces of Oulu and Lapland in northern Finland and was approved by the ethics committee of the Northern Ostrobothnia Hospital District Federation of Municipalities. The cohort was originally collected for the study of risk factors for birth-related complications and includes 12,068 mothers and 12,231 children. All participants of NFBC1966 have provided written informed consent.

WTCCC2 is a collaboration of research groups in the UK to study the genetic factors underlying common diseases such as Parkinson's disease or schizophrenia. Since 2005, the consortium has advanced genetics research through novel biological discoveries and by developing widely used methodologies for GWAS.

UKBB is a longitudinal study of individuals from 40 to 69 years of age in the United Kingdom and was approved by the North West Multi-center Research Ethics Committee. From 2006 to 2010, a sample of 500,000 individuals was collected for the investigation of genetic and environmental factors involved in disease development. All participants of UKBB have provided written informed consent.

3.2 Methods

3.2.1 Genotype data quality control and imputation

For FINRISK and NFBC1966, genotype data was processed by standard sample and variant quality control procedures, pre-phased with Eagle v2.3[69] and genotype imputation was carried out with the IMPUTE2[70] software and a population-specific reference panel consisting of 2,690 high-coverage whole-genome sequenced Finnish individuals. Documentation about the imputation of the UKBB genotype data is given at http://www.ukbiobank.co.uk/wp-content/uploads/2014/04/imputation_documentation_May2015.pdf

3.2.2 Phenotype transformations

In article I, we used measurements of HDL cholesterol levels on 19,115 individuals included in FINRISK 1992-2007. In Article II, we used measurements of Low Density Lipoprotein (LDL) cholesterol levels on 15,626 individuals included in FINRISK 1992-2007. In article III, we used measurements of lipid levels on 21,320 unrelated individuals included in FINRISK 1992-2012. We further used serum levels of 51 biomarkers representing inflammation, metabolism, vascular function, oxidative stress, coagulation, renal function, angiogenesis, and myocardial necrosis from 5,267 unrelated individuals included in the FINRISK 1997. The details about lipid and biomarker measurements are given in [5] and [71].

We excluded individuals on lipid-lowering medication from the analyses in articles I, II and III. Phenotypic values $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ for n individuals were adjusted by the covariates sex, age, age^2 and the first few Principal Components (PCs) of population structure[72] using the following linear model

$$\boldsymbol{y} = \alpha + \boldsymbol{X\beta} + \boldsymbol{\epsilon},$$

where α is the regression intercept, $\boldsymbol{X}$ is a matrix of dimension $n \times p$ with values of p covariate values, $\boldsymbol{\beta}$ are the effect size estimates of the covariates and $\boldsymbol{\epsilon}$ is normal noise with mean 0 and variance σ_ϵ^2. The covariate adjustments were performed with the statistical software R[73] using the lm() function. Regression residuals were normalized in R using a rank-based transformation

$$y_i = \Phi^{-1}\left(\frac{r_i - 0.5}{n}\right)$$

to the standard normal distribution[74], where r_i is the rank of the ith regression residual and $\Phi^{-1}(\cdot)$ is the inverse cumulative function of the normal distribution.

PC analysis was performed from the estimated Genetic Relatedness Matrix (GRM) $\boldsymbol{R}$. The elements of $\boldsymbol{R}$ are computed as

$$\boldsymbol{R}_{ij} = \frac{1}{m}\sum_{\ell=1}^{m}\frac{\left(g_{\ell i} - 2\hat{P}_\ell\right)\left(g_{\ell j} - 2\hat{P}_\ell\right)}{2\hat{P}_\ell(1-\hat{P}_\ell)},$$

where $g_{\ell i}$ and $g_{\ell j}$ are respectively the number of copies of the minor allele for the ith and jth individual at the ℓth variant and $\hat{P}_\ell$ is the sample frequency of the minor allele at the ℓth variant. We estimated the GRM using the software package EIGENSOFT[72,75] and a set of LD pruned variants outside long-range LD regions[76] with MAF greater than 5% and imputation quality score above 0.95. In article III, we performed PC analyses separately for FINRISK 1992-2012 and FINRISK 1997.

3.2.3 Phenotype simulations

We simulated datasets in all projects to validate our statistical methods. Our simulations were based on real genotype data on individuals from the Finnish population or the UKBB. In article I, we used FINRISK genotype data for variants in the *PCSK9* locus on chromosome 1. In article II, we performed genome-wide simulations over 100 GWAS regions (outside the *HLA* region) based on GWAS meta-analyses for lipid traits[5], coronary artery disease[6], Crohn's disease[7], schizophrenia[8], and type 2 diabetes[9] in FINRISK and NFBC1966 data. We further performed simulations with UKBB genotype data for variants in the *ABO* locus on chromosome 9. In article III, we utilized UKBB data for 98 of the 100 GWAS regions from project II.

Datasets were generated by using the following linear model

$$\boldsymbol{y} = \sum_{c \in C} \beta_c \boldsymbol{g}_c + \boldsymbol{\epsilon}\,,$$

where C is the set of causal variants, $\boldsymbol{g}_c$ is the vector of standardized genotypes at the cth causal variant, β_c is the standardized effect size of the cth variant and $\boldsymbol{\epsilon}$ is normal noise with mean 0 and variance σ_ϵ^2. In article I and II, we set the residual variance

$$\sigma_\epsilon^2 = 1 - \sum_{c \in C} \beta_c^2\,.$$

We specified the effect sizes of the causal variants so that the statistical power was approximately 0.5 at the genome-wide significance threshold of 5×10^{-8} with 18,834 individuals in article I and 5,363 in article II. In article III, we set the residual variance $\sigma_\epsilon^2 = 1 - h^2$, where h^2 is the regional heritability, and specified the joint effect sizes of the causal variants such that in the true causal configuration the variants accounted for different proportions of the regional heritability.

3.2.4 Single-variant association testing

In lipid and biomarker analyses, we performed single-variant association testing with a linear regression model implemented in the software SNPTEST[77]. SNPTEST requires genotype probability data for a set of variants and a list of individuals. For a biallelic variant, the ith individual has three genotype probabilities

$$\begin{aligned}\Pr_{A/A} &= \Pr(i\text{th individual has genotype } A/A)\\ \Pr_{A/B} &= \Pr(i\text{th individual has genotype } A/B)\\ \Pr_{B/B} &= \Pr(i\text{th individual has genotype } B/B)\end{aligned}$$

Using the option `–method expected`, SNPTEST performs association testing from expected allele dosages. For the ith individuals, the expected allele dosages of allele B is

$$\mathrm{E}[g_i] = 0 \times \Pr_{A/A} + 1 \times \Pr_{A/B} + 2 \times \Pr_{B/B}\,.$$

The expected allele dosages are used in an individual linear regression for each variant

$$\boldsymbol{y} = \alpha + \beta_\ell \boldsymbol{g}_\ell + \boldsymbol{\epsilon}\,, \tag{6}$$

where $\boldsymbol{y}$ is the vector of phenotypic values for n individuals, α is the regression intercept, $\boldsymbol{g}_\ell$ is a vector of expected allele dosages at the ℓth variant for n individuals, β_ℓ is the effect size of the variant, and $\boldsymbol{\epsilon}$ is normal distributed residual error with mean 0 and variance σ_ϵ^2. SNPTEST outputs the MLE and its SE for the effect size parameter β_ℓ as well as a P-value for testing the hypothesis that $\beta_\ell = 0$ against $\beta_\ell \neq 0$.

We performed single-variant association testing in each simulated dataset by applying linear regression as implemented in the lm() function in R. This function outputs the MLE and its SE for the effect size parameter β_ℓ.

3.2.5 Fine-mapping causal variants

The following is an alternative outline of the statistical fine-mapping model given in article I. The outline provides a mathematical derivation of the result in CAVIARBF of computing the Bayes factor for assessing the evidence in favor of a causal configuration by using only causal variants. The statistical model is built on a Bayesian linear regression model

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\lambda} &\sim N(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\lambda}, \hat{\sigma}^2 \boldsymbol{I}_n) \\ \boldsymbol{\lambda} \mid s_\lambda^2 &\sim N(\boldsymbol{\lambda} \mid \boldsymbol{0}, s_\lambda^2 \hat{\sigma}^2 \boldsymbol{\Delta}_\gamma), \end{aligned}$$

where $\boldsymbol{y}$ is a standardized vector of phenotypic values for n individuals, $\boldsymbol{X}$ is a matrix of dimension $n \times m$ with standardized genotypes of m variants in the region, $\boldsymbol{\lambda}$ are the standardized effect sizes, $\boldsymbol{\Delta}_\gamma = \mathrm{diag}(\delta_{\gamma_1}, \delta_{\gamma_2}, \ldots, \delta_{\gamma_m})$ with $\delta_{\gamma_\ell} = 1$ if $\gamma_\ell = 1$ and $\delta_{\gamma_\ell} \approx 10^{-4}$ if $\gamma_\ell = 0$, and $\hat{\sigma}^2 \approx \mathrm{V}(\boldsymbol{y}) = 1$ is the residual variance that we assume to be close to one because of small effect sizes typical in GWAS. The indicator vector $\boldsymbol{\gamma}$ represents different configurations of causal variants by setting elements equal to one if a variant is causal and zero, otherwise. Integrating out the effect size parameters yields an analytic form for the marginal likelihood for $\boldsymbol{\gamma}$

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\gamma}, s_\lambda^2) &= \int N(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\lambda}, \boldsymbol{I}_n)\, N(\boldsymbol{\lambda} \mid \boldsymbol{0}, s_\lambda^2 \boldsymbol{\Delta}_\gamma)\, \mathrm{d}\boldsymbol{\lambda} \\ &= N(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{I}_n + s_\lambda^2 \boldsymbol{X}\boldsymbol{\Delta}_\gamma \boldsymbol{X}^T) \\ &\propto \det(\boldsymbol{I}_m + s_\lambda^2 \boldsymbol{\Delta}_\gamma \boldsymbol{X}^T \boldsymbol{X})^{-1/2} \\ &\times \exp\left\{ \frac{1}{2} \boldsymbol{y}^T \boldsymbol{X} (s_\lambda^{-2} \boldsymbol{\Delta}_\gamma + \boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \right\}, \end{aligned} \tag{7}$$

where I used, in line 2, the Matrix determinant lemma

$$\det(\boldsymbol{A}_{n\times n} + \boldsymbol{U}_{n\times k}\boldsymbol{W}_{k\times k}\boldsymbol{V}_{k\times n}) = \det(\boldsymbol{A}_{n\times n})\det(\boldsymbol{W}_{k\times k})\det\left(\boldsymbol{W}_{k\times k}^{-1} + \boldsymbol{V}_{k\times n}\boldsymbol{A}_{n\times n}^{-1}\boldsymbol{U}_{n\times k}\right)$$

and Woodbury matrix identity

$$(\boldsymbol{A}_{n\times n} + \boldsymbol{U}_{n\times k}\boldsymbol{W}_{k\times k}\boldsymbol{V}_{k\times n})^{-1} = \boldsymbol{A}_{n\times n}^{-1} - \boldsymbol{A}_{n\times n}^{-1}\boldsymbol{U}_{n\times k}\left(\boldsymbol{W}_{k\times k}^{-1} + \boldsymbol{V}_{k\times n}\boldsymbol{A}_{n\times n}^{-1}\boldsymbol{U}_{n\times k}\right)^{-1}\boldsymbol{V}_{k\times n}\boldsymbol{A}_{n\times n}^{-1}$$

to reduce computational complexity. This step is different from the marginal likelihood derivation in section 2.3.1 in article I where the following affine transformation was used

$$\hat{\boldsymbol{z}} = n^{1/2}\widehat{\boldsymbol{R}}\hat{\boldsymbol{\lambda}}$$

to obtain an expression for the marginal likelihood for $\boldsymbol{\gamma}$. Instead, I replace $(\boldsymbol{y}, \boldsymbol{X})$ with $\left(\hat{\boldsymbol{z}}, \widehat{\boldsymbol{R}}\right)$ in Equation (7) by substitution where

$$\begin{aligned}\boldsymbol{x}_{\cdot\ell}^{T}\boldsymbol{y} &= \sqrt{n}\hat{z}_{\ell}\\ \boldsymbol{X}^{T}\boldsymbol{y} &= \sqrt{n}\hat{\boldsymbol{z}}\\ \boldsymbol{X}^{T}\boldsymbol{X} &= n\widehat{\boldsymbol{R}}\,.\end{aligned}$$

This yields the following expression for the marginal likelihood for $\boldsymbol{\gamma}$

$$\begin{aligned} f\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\gamma}, s_{\lambda}^{2}\right) &\propto \det\left(\boldsymbol{I}_m + \boldsymbol{\Sigma}_{\gamma}\widehat{\boldsymbol{R}}\right)^{-1/2}\exp\left\{\frac{1}{2}\hat{\boldsymbol{z}}^{T}\left(\boldsymbol{\Sigma}_{\gamma}^{-1} + \widehat{\boldsymbol{R}}\right)^{-1}\hat{\boldsymbol{z}}\right\}\\ &= \det\left(\boldsymbol{I}_m + \boldsymbol{\Sigma}_{\gamma}\widehat{\boldsymbol{R}}\right)^{-1/2}\\ &\times \exp\left\{\frac{1}{2}\hat{\boldsymbol{z}}^{T}\left(\widehat{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1}\left(\boldsymbol{\Sigma}_{\gamma} + \widehat{\boldsymbol{R}}^{-1}\right)^{-1}\widehat{\boldsymbol{R}}^{-1}\right)\hat{\boldsymbol{z}}\right\} \qquad (8)\\ &= \det\left(\boldsymbol{I}_m + \boldsymbol{\Sigma}_{\gamma}\widehat{\boldsymbol{R}}\right)^{-1/2}\\ &\times \exp\left\{\frac{1}{2}\hat{\boldsymbol{z}}^{T}\widehat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{z}} - \frac{1}{2}\boldsymbol{z}^{T}\left(\widehat{\boldsymbol{R}}\boldsymbol{\Sigma}_{\gamma}\widehat{\boldsymbol{R}} + \widehat{\boldsymbol{R}}\right)^{-1}\hat{\boldsymbol{z}}\right\}, \end{aligned}$$

where I defined $\boldsymbol{\Sigma}_\gamma = ns_\lambda^2\boldsymbol{\Delta}_\gamma$ and applied in line 2 the Woodbury matrix identity to the inverse matrix in the exponential function. I now partition the GWAS summary data $\left(\hat{\boldsymbol{z}}, \widehat{\boldsymbol{R}}\right)$ into components of k causal variants C and $m-k$ non-causal variants N to obtain an expression for the marginal likelihood such that its evaluation requires only information about the causal variants. That is, I write the z-scores as

$$\hat{\boldsymbol{z}} = \begin{bmatrix} \hat{\boldsymbol{z}}_C \\ \hat{\boldsymbol{z}}_N \end{bmatrix}$$

and permute the rows and columns of the correlation matrix $\widehat{\boldsymbol{R}}$ and $\boldsymbol{\Sigma}_\gamma$ such that

$$\begin{aligned} \widehat{\boldsymbol{R}} &= \begin{bmatrix} \widehat{\boldsymbol{R}}_{CC} & \widehat{\boldsymbol{R}}_{CN} \\ \widehat{\boldsymbol{R}}_{NC} & \widehat{\boldsymbol{R}}_{NN} \end{bmatrix} \\ \boldsymbol{\Sigma}_\gamma &= \begin{bmatrix} \boldsymbol{\Sigma}_{CC} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{NN} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{CC} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\ \widehat{\boldsymbol{R}}\boldsymbol{\Sigma}_\gamma\widehat{\boldsymbol{R}} + \widehat{\boldsymbol{R}} &= \begin{bmatrix} \widehat{\boldsymbol{R}}_{CC} + \widehat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\widehat{\boldsymbol{R}}_{CC} & \widehat{\boldsymbol{R}}_{CN} + \widehat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\widehat{\boldsymbol{R}}_{CN} \\ \widehat{\boldsymbol{R}}_{NC} + \widehat{\boldsymbol{R}}_{NC}\boldsymbol{\Sigma}_{CC}\widehat{\boldsymbol{R}}_{CC} & \widehat{\boldsymbol{R}}_{NN} + \widehat{\boldsymbol{R}}_{NC}\boldsymbol{\Sigma}_{CC}\widehat{\boldsymbol{R}}_{CN} \end{bmatrix}, \end{aligned}$$

where I have set $\boldsymbol{\Sigma}_{NN} = \mathbf{0}$ in order to derive the Bayes factor for assessing the evidence in favor of a causal configuration by using only causal variants. Using the general result for inversion of a 2×2 block matrix

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \\ -(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \end{bmatrix}.$$

I rewrite the quadratic forms in the exponential function of Equation (8) as

$$\begin{aligned} \hat{\boldsymbol{z}}^T\widehat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{z}} = \hat{\boldsymbol{z}}_C^T\widehat{\boldsymbol{R}}_{CC}^{-1}\hat{\boldsymbol{z}}_C \\ + \hat{\boldsymbol{z}}_C^T\widehat{\boldsymbol{R}}_{CC}^{-1}\widehat{\boldsymbol{R}}_{CN}\left(\widehat{\boldsymbol{R}}_{NN} - \widehat{\boldsymbol{R}}_{NC}\widehat{\boldsymbol{R}}_{CC}^{-1}\widehat{\boldsymbol{R}}_{CN}\right)^{-1}\widehat{\boldsymbol{R}}_{NC}\widehat{\boldsymbol{R}}_{CC}^{-1}\hat{\boldsymbol{z}}_C - 2\hat{\boldsymbol{z}}_C^T\widehat{\boldsymbol{R}}_{CC}^{-1}\widehat{\boldsymbol{R}}_{CN}\left(\widehat{\boldsymbol{R}}_{NN}\right. \\ \left. - \widehat{\boldsymbol{R}}_{NC}\widehat{\boldsymbol{R}}_{CC}^{-1}\widehat{\boldsymbol{R}}_{CN}\right)^{-1}\hat{\boldsymbol{z}}_N + \hat{\boldsymbol{z}}_N^T\left(\widehat{\boldsymbol{R}}_{NN} - \widehat{\boldsymbol{R}}_{NC}\widehat{\boldsymbol{R}}_{CC}^{-1}\widehat{\boldsymbol{R}}_{CN}\right)^{-1}\hat{\boldsymbol{z}}_N \end{aligned}$$

and

$$
\begin{aligned}
&\hat{\boldsymbol{z}}^T\left(\hat{\boldsymbol{R}}\boldsymbol{\Sigma}_\gamma\hat{\boldsymbol{R}}+\hat{\boldsymbol{R}}\right)^{-1}\hat{\boldsymbol{z}} \\
&\quad=\hat{\boldsymbol{z}}_C^T\Big[\left(\hat{\boldsymbol{R}}_{CC}+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)^{-1} \\
&\quad+\left(\hat{\boldsymbol{R}}_{CC}+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)^{-1}\left(\hat{\boldsymbol{R}}_{CN}\right. \\
&\quad\left.+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CN}\right)\Big[\hat{\boldsymbol{R}}_{NN}+\hat{\boldsymbol{R}}_{NC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CN} \\
&\quad-\left(\hat{\boldsymbol{R}}_{NC}+\hat{\boldsymbol{R}}_{NC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)\left(\hat{\boldsymbol{R}}_{CC}+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)^{-1}\left(\hat{\boldsymbol{R}}_{CN}\right. \\
&\quad\left.+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CN}\right)\Big]^{-1}\left(\hat{\boldsymbol{R}}_{NC}+\hat{\boldsymbol{R}}_{NC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)\left(\hat{\boldsymbol{R}}_{CC}+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)^{-1}\Big]\hat{\boldsymbol{z}}_C \\
&\quad-2\hat{\boldsymbol{z}}_C^T\left(\hat{\boldsymbol{R}}_{CC}+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)^{-1}\left(\hat{\boldsymbol{R}}_{CN}\right. \\
&\quad\left.+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CN}\right)\Big[\hat{\boldsymbol{R}}_{NN}+\hat{\boldsymbol{R}}_{NC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CN} \\
&\quad-\left(\hat{\boldsymbol{R}}_{NC}+\hat{\boldsymbol{R}}_{NC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)\left(\hat{\boldsymbol{R}}_{CC}+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)^{-1}\left(\hat{\boldsymbol{R}}_{CN}\right. \\
&\quad\left.+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CN}\right)\Big]^{-1}\hat{\boldsymbol{z}}_N \\
&\quad+\hat{\boldsymbol{z}}_N^T\Big[\hat{\boldsymbol{R}}_{NN}+\hat{\boldsymbol{R}}_{NC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CN} \\
&\quad-\left(\hat{\boldsymbol{R}}_{NC}+\hat{\boldsymbol{R}}_{NC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)\left(\hat{\boldsymbol{R}}_{CC}+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)^{-1}\left(\hat{\boldsymbol{R}}_{CN}\right. \\
&\quad\left.+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CN}\right)\Big]^{-1}\hat{\boldsymbol{z}}_N \\
&\quad=\hat{\boldsymbol{z}}_C^T\left(\hat{\boldsymbol{R}}_{CC}+\hat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right)^{-1}\hat{\boldsymbol{z}}_C \\
&\quad+\hat{\boldsymbol{z}}_C^T\hat{\boldsymbol{R}}_{CC}^{-1}\hat{\boldsymbol{R}}_{CN}\left(\hat{\boldsymbol{R}}_{NN}-\hat{\boldsymbol{R}}_{NC}\hat{\boldsymbol{R}}_{CC}^{-1}\hat{\boldsymbol{R}}_{CN}\right)^{-1}\hat{\boldsymbol{R}}_{NC}\hat{\boldsymbol{R}}_{CC}^{-1}\hat{\boldsymbol{z}}_C \\
&\quad-2\hat{\boldsymbol{z}}_C^T\hat{\boldsymbol{R}}_{CC}^{-1}\hat{\boldsymbol{R}}_{CN}\left(\hat{\boldsymbol{R}}_{NN}-\hat{\boldsymbol{R}}_{NC}\hat{\boldsymbol{R}}_{CC}^{-1}\hat{\boldsymbol{R}}_{CN}\right)^{-1}\hat{\boldsymbol{z}}_N \\
&\quad+\hat{\boldsymbol{z}}_N^T\left(\hat{\boldsymbol{R}}_{NN}-\hat{\boldsymbol{R}}_{NC}\hat{\boldsymbol{R}}_{CC}^{-1}\hat{\boldsymbol{R}}_{CN}\right)^{-1}\hat{\boldsymbol{z}}_N\,.
\end{aligned}
$$

I also rewrite the determinant in Equation (8) as

$$
\det\left(\boldsymbol{I}_m+\boldsymbol{\Sigma}_\gamma\hat{\boldsymbol{R}}\right)=\det\left(\begin{bmatrix}\boldsymbol{I}_k+\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC} & \hat{\boldsymbol{R}}_{CN}\\ \boldsymbol{0} & \boldsymbol{I}_{m-k}\end{bmatrix}\right)=\det\left(\boldsymbol{I}_k+\boldsymbol{\Sigma}_{CC}\hat{\boldsymbol{R}}_{CC}\right).
$$

This means that the exponential function in the marginal likelihood in Equation (8) can be evaluated by using only information about the causal variants

$$
\begin{aligned}
f\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\gamma}, s_\lambda^2\right) &\propto \det\left(\boldsymbol{I}_m + \boldsymbol{\Sigma}_\gamma \widehat{\boldsymbol{R}}\right)^{-1/2} \exp\left\{\frac{1}{2}\hat{\boldsymbol{z}}^T \widehat{\boldsymbol{R}}^{-1} \hat{\boldsymbol{z}} - \frac{1}{2}\boldsymbol{z}^T \left(\widehat{\boldsymbol{R}}\boldsymbol{\Sigma}_\gamma \widehat{\boldsymbol{R}} + \widehat{\boldsymbol{R}}\right)^{-1} \hat{\boldsymbol{z}}\right\} \\
&= \det\left(\boldsymbol{I}_k + \boldsymbol{\Sigma}_{CC} \widehat{\boldsymbol{R}}_{CC}\right)^{-1/2} \\
&\times \exp\left\{\frac{1}{2}\hat{\boldsymbol{z}}_C^T \widehat{\boldsymbol{R}}_{CC}^{-1} \hat{\boldsymbol{z}}_C + \frac{1}{2}\hat{\boldsymbol{z}}_C^T \left(\widehat{\boldsymbol{R}}_{CC} + \widehat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\widehat{\boldsymbol{R}}_{CC}\right)^{-1} \hat{\boldsymbol{z}}_C\right\}.
\end{aligned}
$$

I can thus evaluate the Bayes factor for assessing the evidence in favor of a causal configuration $\boldsymbol{\gamma}$ against the null configuration $\boldsymbol{\gamma}_0$ by using a ratio of normal densities evaluated at the z-scores of the causal variants

$$
\begin{aligned}
\mathrm{BF}(\gamma : \gamma_0) &= \frac{f\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\gamma}, s_\lambda^2\right)}{f\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\gamma}_0, s_\lambda^2\right)} \\
&= \det\left(\widehat{\boldsymbol{R}}_{CC}\right)^{-1/2} / \det\left(\widehat{\boldsymbol{R}}_{CC}\right)^{-1/2} \det\left(\boldsymbol{I}_k + \boldsymbol{\Sigma}_{CC}\widehat{\boldsymbol{R}}_{CC}\right)^{-1/2} \\
&\times \exp\left\{\frac{1}{2}\boldsymbol{z}_C^T \widehat{\boldsymbol{R}}_{CC}^{-1} \boldsymbol{z}_C + \frac{1}{2}\hat{\boldsymbol{z}}_C^T \left(\widehat{\boldsymbol{R}}_{CC} + \widehat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\widehat{\boldsymbol{R}}_{CC}\right)^{-1} \hat{\boldsymbol{z}}_C\right\} \\
&= \frac{N\left(\hat{\boldsymbol{z}}_C \mid \boldsymbol{0}, \widehat{\boldsymbol{R}}_{CC} + \widehat{\boldsymbol{R}}_{CC}\boldsymbol{\Sigma}_{CC}\widehat{\boldsymbol{R}}_{CC}\right)}{N\left(\hat{\boldsymbol{z}}_C \mid \boldsymbol{0}, \widehat{\boldsymbol{R}}_{CC}\right)}.
\end{aligned}
$$

4 RESULTS

4.1 Efficient and accurate GWAS summary statistics-based fine-mapping using stochastic search

Privacy concerns and the logistics of sharing individual-level data have led to the development of statistical fine-mapping methods that work directly on GWAS summary data[12,19-23]. Working with GWAS summary data instead of individual-level genotype-phenotype data is computationally attractive, because the GWAS summary data size is much smaller. Computational improvements in statistical fine-mapping methods have also been achieved by optimizing likelihood evaluations through linear algebra factorizations[21]. However, fine-mapping causal variants is a combinatorial problem and fine-mapping methods previously[19-21] relied on computationally expensive exhaustive search.

Through the intuition that the majority of possible causal configurations have small posterior probability and can thus be neglected, we approached this hard combinatorial problem by implementing an ultra-fast shotgun stochastic search algorithm in the FINEMAP software[22]. We showed that stochastic search in FINEMAP is computationally efficient and yields accurate results by comparing it to the state-of-the-art exhaustive search implemented in CAVIARBF. In simulations, we generated datasets with 150 variants of which five were causal. The causal variants were selected such that there was at least one non-causal variant with absolute correlation above 0.5 with a causal variant. This selection of causal variant was chosen to verify whether a stochastic search can deal with highly correlated proxies.

We showed that stochastic search is as accurate as exhaustive search by comparing the posterior probability of causality for each variant from CAVIARBF and FINEMAP. The largest absolute difference between the posterior probabilities was 0.11 with the median being smaller than 6×10^{-4}. We also noticed that FINEMAP achieves even higher accuracy when the maximum number of allowed causal variants in CAVIARBF must be restricted due to combinatorial explosion.

Furthermore, we demonstrated that stochastic search performs fine-mapping in a fraction of the processing time of exhaustive search. We performed fine-mapping in simulated datasets with varying numbers of variants. The processing time for the stochastic search implemented in FINEMAP remained approximately constant as a function of the number of variants in the region and increased only marginally with increasing maximum allowed number of causal variants. The computational efficiency of FINEMAP is due to the stochastic search, which starts to concentrate quickly on regions of the model space where high posterior probability models can be found. In contrast, we observed that increasing the number of variants in CAVIARBF results in a combinatorial explosion that significantly extends the processing time of the implemented exhaustive search. To investigate how computationally inefficient exhaustive search is, we also performed fine-mapping in datasets with a few hundred variants and varied the maximum allowed numbers of causal variants. We observed that the difference in processing times between stochastic and exhaustive search become even more accentuated.

Computationally expensive exhaustive search restricts fine-mapping to a few hundred variants. The efficiency of FINEMAP scales up fine-mapping to regions with thousands of variants, thus enabling the extraction of valuable information which could otherwise remain undetected due to the limitations of previous methods. For example, fine-mapping of the association between PON1 serum levels and a genomic region of 7.5 mega bases with 16,189 variants around the *PON1* gene was completed in about 14 minutes using FINEMAP v1.3, while allowing for maximally ten causal variants to enable identification of the large number of known PON1 variants[37].

4.2 Importance of choosing the correct LD information in GWAS summary statistics-based fine-mapping

Fine-mapping methods that work directly on GWAS summary statistics require as input estimates of pairwise correlations between the variants. Previously, it has been thought that correlation estimates could be obtained from publicly available reference genotype data such as the 1000GP[32] or the HRC[35]. However, the impact of this strategy had not

been studied previously. Motivated by feedback from FINEMAP users on the performance of the 1000GP data for obtaining correlation estimates, we studied the consequences of misspecification of correlations that could happen when publicly available reference genomes are used.

As a motivating example, we fine-mapped the association of the *APOE* locus to LDL cholesterol level. Using correlations between variants from the same genotype data on the same samples used in GWAS yielded the highest evidence for the two well-known missense variants, rs7412 and rs429358, as well as a third variant rs35136575, which had been identified in an earlier targeted study of the *APOE* locus[78]. Performing fine-mapping using the same GWAS summary statistics with correlation estimates obtained from the 1000GP data resulted in different results. There was still high evidence for rs7412 and rs35136575 to be causal, but rs429358 showed no evidence of being causal. Additionally, several variants which showed no evidence of being causal with correlation estimates computed from the original GWAS genotype data suddenly showed decisive evidence of being causal.

The *APOE* example motivated us to conduct comprehensive simulations to examine the performance of reference genotype panels. We used genotype data from 5,363 samples from the NFBC1966 study for phenotype simulations and generation of GWAS summary statistics. We used genotype data from the FINRISK study to create reference genotype panels with 100, 1,000 and 5,000 samples. We computed correlations between variants from the same NFCB1966 genotype data used to generate the GWAS summary statistics, as well as generating correlations from the FINRISK reference panels. Fine-mapping using NFBC1966 GWAS summary statistics was performed by using either NFBC1966 or FINRISK correlation estimates. Comparison of the two sets of fine-mapping results showed that correlation estimates obtained from reference panels with 100 samples perform significantly worse than using correlation estimates from the original NFBC1966 GWAS data. However, reference panels with at least 1,000 samples achieved very comparable performance compared to using the original NFBC1966 GWAS genotype data to obtain correlation estimates.

We also investigated the performance of reference panels when the GWAS sample size is 50,000 and noticed that the size of the reference panel needs to scale with the GWAS sample size. In particular, a reference panel size of 1,000 samples that performed well with a GWAS sample size of about 5,000 samples was no longer adequate with a GWAS sample size of 50,000. A reference panel size of 10,000 was needed to achieve the same performance as using the original GWAS genotype data to estimate correlations. We further confirmed this empirical result that the size of the reference panel needs to scale proportionally with the GWAS sample size by studying theoretically the asymptotic behavior of the posterior probabilities for a pair of variants.

This finding has important consequences for the application of all fine-mapping methods using GWAS summary statistics from GWAS consortia in which correlation estimates from each participating study are typically not available. For that reason, we provided the LDstore[79] software to help compute and share correlation estimates in this setting.

4.3 Heritability estimation from fine-mapped variants and large effect size regions

The output from FINEMAP is a list of causal configurations and their posterior probabilities, as well as the posterior probability of causality for each individual variant. These probabilities contain all the information needed for downstream analyses in fine-mapped regions.

We implemented a method for the estimation of effect sizes and regional heritability in FINEMAP to shed light on the regional genetic architecture of complex phenotypes. To verify our implementation of regional heritability decomposition into causal variants, we compared FINEMAP with the variance components model implemented in BOLT[42] and the fixed-effect model implemented in HESS[43] which are based on fundamentally different model assumptions. In simulated datasets with few causal variants, we observed that FINEMAP and BOLT are nearly unbiased and provide accurate uncertainty quantification of the regional heritability. In contrast, HESS was downward biased

and showed low precision in the estimates. This is likely due to the regularization method in HESS removing too much signal from the data, because deactivating the regularization solved the problem for small levels of regional heritability.

In our biomarker analysis with 5,265 Finnish samples, we investigated whether there is a residual polygenic signal in the biomarker-associated regions by comparing the heritability estimates from FINEMAP with BOLT and HESS. We indeed noticed that heritability estimates from FINEMAP were 20% lower than those from BOLT. A possible explanation is that FINEMAP underestimates and BOLT overestimates the regional heritability when the sample size is small, but there could also be a true polygenic signal in these regions. HESS showed downward bias in our biomarker analysis which could not be solved by deactivating the regularization method in HESS. Scaling the analysis for lipid traits from 5,265 Finns to 21,320 Finns shows good agreement between FINEMAP and BOLT, whereas HESS estimates were consistently lower.

FINEMAP was developed for genomic regions where variants have small effect sizes as are typical in GWAS. This allowed us to fix residual variance parameters in the model to one. However, this approximation does not work with large-effect size regions such as in our biomarker analysis. Fine-mapping biomarker-associated regions with a wide range of heritability levels motivated us to develop the statistical model further to handle large-effect size regions. In the era of biobank-scale data, extending FINEMAP in v1.3 to handle large-effect size regions and estimation of regional heritability and effect sizes allows for the investigation of regional genetic architecture at the variant level.

5 DISCUSSION

GWAS discoveries have created robust catalogues of statistical associations between thousands of genomic regions and hundreds of phenotypes. Fine-mapping causal variants in these regions has the potential to help translate GWAS discoveries into drug targets for the development of new medicines. Several fine-mapping methods[19-21,23,24] have recently been developed to facilitate this step. This doctoral research presented the development of the FINEMAP[22] software for fine-mapping causal variants.

5.1 Efficiency and accuracy of fine-mapping

In article I, we implemented a stochastic search algorithm in the FINEMAP software to resolve computational shortcomings in existing fine-mapping methods that hindered their use in practice. We compared the stochastic search in FINEMAP with the gold standard exhaustive search implemented in CAVIARBF[67]. We showed through comprehensive simulations that stochastic search is able to maintain the accuracy of exhaustive search while requiring only a fraction of its processing time. For example, FINEMAP opened up completely new opportunities by fine-mapping the HDL cholesterol level association of the *LIPC* locus with 20,000 variants in less than 90 seconds while CAVIARBF was estimated to run over 300 years for the same analysis.

Current datasets, for instance from the UKBB, contain genetic data on 500,000 individuals at over 90 million variants. According to the authors from the widely used software package BOLT[42], one GWAS on UKBB genotype-phenotype data would require several days to a week using high performance parallel computing and more than 100 gigabytes of computer memory. Utilizing GWAS results with FINEMAP makes the computation independent from the GWAS sample sizes; this allows FINEMAP to cope with ever-increasing sample sizes in future GWAS. In contrast to standard GWAS software, FINEMAP completes fine-mapping of GWAS regions within minutes to hours on an off-the-shelf desktop computer.

FINEMAP uses a stochastic search algorithm that generates a list of possible causal configurations and their posterior probabilities. We highlighted in article I that modeling the whole *LIPC* locus with FINEMAP enabled the identification of a 3-SNP configuration with 190-fold higher likelihood than the top configuration from conditional analysis. The result from FINEMAP suggested that a missense variant and a promoter polymorphism are likely to be causal whereas the lead variant in GWAS had less evidence of being causal than a regulatory variant correlated with it. This example shows that stochastic search can identify more plausible causal configurations than standard conditional analysis.

Conditional analysis may, however, be an attractive approach if only a single causal configuration is required. A single causal configuration may suffice, for instance, for phenotype predictions or quantification of the number of causal variants. Conditional analysis is, however, lacking probabilistic assessments of the causality for individual variants, making it less suitable in downstream analyses such as colocalization of GWAS signals or when overlaying fine-mapping results with functional variant annotations. Since stochastic search has not yet been comprehensively compared with conditional analysis, it is unclear how often the methods give different solutions. It can be expected that differences between both approaches become more accentuated in regions with multiple causal variants and in regions where causal variants are tagged by many variants[23].

JAM[23] is a fine-mapping method with a similar statistical model to FINEMAP. It was published after the publication of article I and uses a particular version of the MH sampler called Reversible Jump MCMC[80] (RJMCMC) as the search strategy. The RJMCMC sampler uses delete/change/add moves to explore the high-dimensional model space of causal configurations, making RJMCMC comparable to the Shotgun Stochastic Search (SSS) algorithm implemented in FINEMAP in terms of fine-mapping accuracy. Although SSS is related to MCMC sampling, it improves upon sequential MCMC by exploring the whole set of neighboring causal configurations at each step; this enables SSS to discover causal configurations with high posterior probability much more rapidly than MCMC[81].

DAP[24] is another fine-mapping method with a similar statistical model to FINEMAP. It was also published after the publication of article I. DAP uses preselected variants[82] with the strongest marginal association and a Bayesian version of conditional analysis[83] to perform variable selection in GWAS and large effect size regions. The authors simulated datasets where the heritability contribution from causal variants is about 0.26 to compare DAP to FINEMAP version 1.1, which is not suitable for such large-effect regions. The performance of DAP's preselection step has not been validated in regions where causal variants have much smaller effects. This would be important to show because GWAS regions typically harbor variants with significantly smaller heritability contribution. In GWAS regions, plausible causal configurations could be missed by DAP because it preselects variants on the basis of their marginal association statistics, but a significant association of causal variants is only highlighted in their joint model (see Parkinson's disease example in section 5.3 of article I). A comprehensive review and discussion of all fine-mapping methods under comparable settings is currently lacking and would be useful for the genetics community.

5.2 Importance of choosing the correct LD information

Fine-mapping methods that work on GWAS results require estimates of LD between variants. Ideally, LD estimates are obtained from the same genotype data on the same individuals that were used in the GWAS. However, LD estimates can also be obtained from publicly available reference genotype panels such as the 1000GP and the HRC. Alarming signs of incompatibility between GWAS results and LD estimates from reference panels[84], and similar feedback from FINEMAP users, motivated us to investigate the impact of using LD estimates from reference panels on fine-mapping accuracy in article II. We reported the important practical results that a reference panel can cause bias if its size is insufficient compared to the GWAS sample size. This finding has important consequences for the application of all fine-mapping methods using GWAS results from GWAS consortia in which accurate LD estimates from each participating study are typically not available.

GWAS consortia have thus far approached fine-mapping of causal variants by either 1) meta-analyzing results from conditional analysis requiring multiple rounds of time-consuming coordination between participating groups[85], 2) by performing simplistic fine-mapping under the assumption of a single causal variant in the region to avoid LD estimates altogether[86,87], or 3) by using a publicly available reference panel or genotype data from the largest participating GWAS to obtain LD estimates[88], which can be inaccurate in the light of article II.

To explain why misspecification of LD estimates can cause bias, consider the scenario of one causal and one non-causal variant. Let the two variants be highly correlated in the original GWAS data but assume that the LD estimate from the reference panel underestimates their correlation. By using the LD estimate from the reference panel, the fine-mapping method will regard both variants as nearly independent and incorrectly label both as causal. On the other hand, let the correlation between the two variants be moderate, but assume that the LD estimate from the reference panel overestimates their correlation. By using the LD estimate from the reference panel, the fine-mapping model will label both variants as causal in order to make the GWAS results for the two variants consistent with their LD estimate.

Using conditional analysis implemented in the GCTA software and LD estimates from a reference panel, fine-mapping efforts by the GIANT consortium found on average 4.6 causal variants per height-associated region with a standard deviation of 6 causal variants per region. Interestingly, 19 causal variants were found in a region around the *IGF1* locus. In light of the results from article II, it would be worthwhile to study this region with LD estimates from the original GWAS data to investigate whether external LD estimates may have caused false positives. It should be noted that it is currently difficult to obtain LD estimates from publicly available reference panels that accurately match GWAS results, because the individual ancestry proportions used in imputation of each GWAS cohort are unknown. To provide GWAS consortia the possibility to use the latest generation of fine-mapping methods, we developed, in article II, the LDstore software for sharing LD estimates computed from the original individual-level GWAS genotype data.

5.3 Heritability estimation from fine-mapped variants and large effect size regions

Ever-increasing sample sizes through large-scale sequencing efforts or biobank projects allow for unprecedented causal inference. In article III, we utilized the estimation of effect sizes at causal variants to estimate the heritability contribution from causal variants in GWAS regions. We compared FINEMAP in 110 regions across 51 biomarkers on 5,265 Finns with approaches implemented in the software BOLT and HESS that make fundamentally different modeling assumptions. Our results showed good concordance among all methods in regions with negligible contribution to the genome-wide heritability, whereas BOLT and HESS yielded respectively larger and smaller estimates than FINEMAP in regions with moderate to high heritability levels. Scaling the analysis for lipid traits from 5,265 Finns to 21,320 Finns showed good agreement between FINEMAP and BOLT also for moderate to high levels of regional heritability, whereas HESS estimates are consistently lower at these levels. Comparison on biobank-scale data suggested that an upward bias of BOLT and a downward bias of FINEMAP could together explain the observed difference between the methods.

One of the ultimate goals in modern genetics research is to reveal the biological mechanisms behind phenotype-associated GWAS regions. Comparing regional heritability estimates originating from different model assumptions would therefore provide important information about the genetic architecture of each region. In addition to regional heritability estimation, effect size estimates at causal variants could be used for genetic risk score evaluations[89] or in Mendelian randomization[90]. Folding in probabilistic descriptions about causal variants could be useful in these downstream analyses and, where possible, yield a detailed variant-level picture of genetic architecture.

5.4 Impact

Fine-mapping of GWAS regions is one of the great challenges in modern genetics research of complex phenotypes. Innovative, practical and well supported solutions are

needed in the post-GWAS era to solve such great challenges. Since its publication in 2016, the FINEMAP software has already been adopted by various GWAS to investigate, for example, lipid levels[85], cytokines and growth factors[91], bone mineral density[92], pulmonary function[93], C-reactive protein[94], and psychiatric disorders[95-97]. The availability of genotype data and exceptionally rich phenotypic information on all 500,000 UKBB participants has also led to the application of FINEMAP and LDstore in biobank studies of several phenotypes such as bone mineral density[98], physical activity and sleep[99] or blood cell traits[100]. This doctoral thesis project has helped set standards of analysis within this field and, for the UKBB, it has made fine-mapping as simple and seamless to do as running a GWAS using standard GWAS software. For instance, FINEMAP version 1.3 supporting on-the-fly LD estimation from UKBB genotype data opens up the possibility to leverage publicly available UKBB GWAS results (available at https://sites.google.com/broadinstitute.org/ukbbgwasresults) for rapid and unparalleled fine-mapping of causal variants across thousands of phenotypes.

In order to contribute to the GWAS field in impactful ways, mere technical innovations and focus on implementation of scientific software is not sufficient. Establishing collaborative links with research groups and actively supporting robust and user-friendly software is key to making statistical methods most accessible to the international research community.

6 CONCLUSIONS AND FUTURE ASPECTS

The last ten years of highly successful GWAS produced a wealth of robust statistical associations between genomic regions and various human traits and diseases. The next major challenge in genetics research is to understand the biological processes behind these statistical associations. Better understanding of the biology underlying the associated genomic regions is needed to identify genes that could become drug targets for the development of new therapeutic interventions. Some evidence has already emerged that genetics research can improve decision making at early stages of the drug development process[3].

Fine-mapping is an important post-GWAS analysis that holds the promise of advancing the understanding of complex traits and diseases by breaking down GWAS associations into causal variants. This doctoral thesis presented the novel contributions to fine-mapping methodology that resulted in the FINEMAP software. The role of FINEMAP as an innovative, practical solution to the fine-mapping problem has been demonstrated by comparison to the gold standard state-of-the-art. Confusion about using reference genotype panels in fine-mapping has been resolved by demonstrating why and when misspecification of LD estimates can create problems. The benefits of harnessing increased statistical power provided by biobank-scale data for fine-mapping causal variants has been demonstrated by comparing heritability estimates using a fine-mapping model to heritability estimates using a polygenic model.

While this doctoral thesis has led to important improvements in fine-mapping methodology, there exists several methodological directions into which fine-mapping can be developed further. For instance, it would be useful to extend FINEMAP to multi-phenotype and cross-population fine-mapping analysis to increase statistical power. This will be crucial for large-scale meta-analysis of GWAS in phenotype rich biobank projects across populations with varying ancestries. Another way to develop FINEMAP further is to fold in functional variant annotations in order to improve fine-mapping resolution in high LD regions. Such a feature would also enable extraction of additional, valuable biological information from GWAS regions.

In conclusion, this doctoral thesis presented and demonstrated cutting-edge innovations in statistical fine-mapping methodology. The fact that many research groups have already adopted FINEMAP in their projects shows that the community finds it a useful tool that could potentially bring human genetics research a step closer towards discovering new genes and genetic variants that would become drug targets for the development of new medicines.

7 REFERENCES

1. Visscher, P.M. *et al*. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).
2. Khera, A.V. *et al*. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *N Engl J Med* **375**, 2349-2358 (2016).
3. Nelson, M.R. *et al*. The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856-60 (2015).
4. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
5. Surakka, I. *et al*. The impact of low-frequency and rare variants on lipid levels. *Nat Genet* **47**, 589-97 (2015).
6. Nikpay, M. *et al*. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47**, 1121-30 (2015).
7. Jostins, L. *et al*. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
8. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).
9. Morris, A.P. *et al*. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981-90 (2012).
10. Schaid, D.J., Chen, W. & Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* (2018).
11. Spain, S.L. & Barrett, J.C. Strategies for fine-mapping complex traits. *Hum Mol Genet* **24**, R111-9 (2015).
12. Yang, J. *et al*. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).
13. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**, e114 (2007).
14. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780-1815 (2011).
15. Bottolo, L. *et al*. GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genet* **9**, e1003657 (2013).
16. Bottolo, L. & Richardson, S. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal.* **5**, 583-618 (2010).
17. Carbonetto, P. & Stephens, M. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Anal.* **7**, 73-108 (2012).

18. Wellcome Trust Case Control Consortium *et al*. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294-301 (2012).
19. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497-508 (2014).
20. Kichaev, G. *et al*. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**, e1004722 (2014).
21. Chen, W. *et al*. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**, 719-36 (2015).
22. Benner, C. *et al*. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-501 (2016).
23. Newcombe, P.J., Conti, D.V. & Richardson, S. JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genet Epidemiol* **40**, 188-201 (2016).
24. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *Am J Hum Genet* **98**, 1114-1129 (2016).
25. Marcus, J.H. & Novembre, J. Visualizing the geography of genetic variants. *Bioinformatics* **33**, 594-595 (2017).
26. Chen, B. & Butte, A.J. Leveraging big data to transform target selection and drug discovery. *Clin Pharmacol Ther* **99**, 285-97 (2016).
27. Abraham, G. & Inouye, M. Genomic risk prediction of complex human disease and its clinical application. *Curr Opin Genet Dev* **33**, 10-6 (2015).
28. Flannick, J. *et al*. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet* **46**, 357-63 (2014).
29. Alberts, B. *et al*. Molecular Biology of the Cell, Sixth Edition. *Molecular Biology of the Cell, Sixth Edition*, 1-1342 (2015).
30. 1000 Genomes Project Consortium *et al*. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
31. Collins, A. *Linkage Disequilibrium and Association Mapping: Analysis and Applications*, (Humana Press, 2007).
32. 1000 Genomes Project Consortium *et al*. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
33. LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* **37**, 4181-93 (2009).
34. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
35. McCarthy, S. *et al*. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
36. Sham, P.C. & Purcell, S.M. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* **15**, 335-46 (2014).

37. Tang, W.H. *et al*. Clinical and genetic association of serum paraoxonase and arylesterase activities with cardiovascular risk. *Arterioscler Thromb Vasc Biol* **32**, 2803-12 (2012).
38. Locke, A.E. *et al*. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
39. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* **9**, 255-66 (2008).
40. Manolio, T.A. *et al*. Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
41. Vinkhuyzen, A.A. *et al*. Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. *Transl Psychiatry* **2**, e102 (2012).
42. Loh, P.R. *et al*. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-90 (2015).
43. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am J Hum Genet* **99**, 139-53 (2016).
44. Casella, G. & Berger, R.L. *Statistical Inference*, (Duxbury Thomson Learning, 2002).
45. Press, S.J. *Subjective and objective Bayesian statistics: principles, models, and applications*, (Wiley-Interscience, 2003).
46. Stern, C. The Hardy-Weinberg Law. *Science* **97**, 137-8 (1943).
47. McCullagh, P. *Generalized linear models*, (Chapman and Hall, London ;, 1983).
48. Pawitan, Y. *In all likelihood : statistical modelling and inference using likelihood*, xiii, 528 pages (Oxford University Press, Oxford, 2013).
49. Sauer, T. *Numerical Analysis*, (Pearson Addison Wesley, 2006).
50. Tierney, L. Markov Chains for Exploring Posterior Distributions. *Ann. Statist.* **22**, 1701-1728 (1994).
51. Hastings, W.K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**, 97-109 (1970).
52. Geman, S. & Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *Ieee Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741 (1984).
53. Geweke, J. Contemporary Bayesian Econometrics and Statistics. *Contemporary Bayesian Econometrics and Statistics*, 1-308 (2005).
54. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821-4 (2012).
55. Gentle, J.E. *Matrix Algebra: Theory, Computations, and Applications in Statistics*, (Springer New York, 2007).
56. Pirinen, M., Donnelly, P. & Spencer, C.C.A. Efficient Computation with a Linear Mixed Model on Large-Scale Data Sets with Applications to Genetic Studies. *Annals of Applied Statistics* **7**, 369-390 (2013).
57. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716-723 (1974).

58. Schwarz, G. Estimating the Dimension of a Model. *Ann. Statist.* **6**, 461-464 (1978).
59. Furnival, G.M. & Wilson, R.W. Regressions by Leaps and Bounds. *Technometrics* **16**, 499-511 (1974).
60. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288 (1996).
61. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67**, 301-320 (2005).
62. O'Hara, R.B. & Sillanpaa, M.J. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal.* **4**, 85-117 (2009).
63. Dellaportas, P., Forster, J.J. & Ntzoufras, I. *Bayesian variable selection using the Gibbs sampler*, (Taylor & Francis, 2000).
64. George, E.I. & McCulloch, R.E. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association* **88**, 881-889 (1993).
65. Kuo, L. & Mallick, B. Variable Selection for Regression Models. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)* **60**, 65-81 (1998).
66. Lozano, J.A., Hormozdiari, F., Joo, J.W., Han, B. & Eskin, E. The Multivariate Normal Distribution Framework for Analyzing Association Studies. *bioRxiv* (2017).
67. Xu, Z. *et al.* DISSCO: direct imputation of summary statistics allowing covariates. *Bioinformatics* **31**, 2434-42 (2015).
68. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1-38 (1977).
69. Loh, P.R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448 (2016).
70. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
71. Blankenberg, S. *et al.* Contribution of 30 biomarkers to 10-year cardiovascular risk estimation in 2 population cohorts: the MONICA, risk, genetics, archiving, and monograph (MORGAM) biomarker project. *Circulation* **121**, 2388-97 (2010).
72. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
73. R Core Team. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, 2018).
74. Bliss, C.I. *Statistics in Biology*, (McGraw-Hill, New York, 1969).
75. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
76. Price, A.L. *et al.* Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* **83**, 132-5; author reply 135-9 (2008).

77. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
78. Klos, K. *et al*. APOE/C1/C4/C2 hepatic control region polymorphism influences plasma apoE and LDL cholesterol levels. *Hum Mol Genet* **17**, 2039-46 (2008).
79. Benner, C. *et al*. Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *Am J Hum Genet* **101**, 539-551 (2017).
80. Green, P.J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732 (1995).
81. Hans, C., Dobra, A. & West, M. Shotgun Stochastic Search for "Large p" Regression. *Journal of the American Statistical Association* **102**, 507-516 (2007).
82. Jianqing, F. & Jinchi, L. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849-911 (2008).
83. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet* **9**, e1003486 (2013).
84. Chen, W., McDonnell, S.K., Thibodeau, S.N., Tillmans, L.S. & Schaid, D.J. Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics. *Genetics* **204**, 933-958 (2016).
85. Iotchkova, V. *et al*. Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. *Nat Genet* **48**, 1303-1312 (2016).
86. Locke, A.E. *et al*. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
87. Gormley, P. *et al*. Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nat Genet* **48**, 856-66 (2016).
88. Wood, A.R. *et al*. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-86 (2014).
89. Vilhjalmsson, B.J. *et al*. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97**, 576-92 (2015).
90. Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B.B. & Hopewell, J.C. Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. *Genetic Epidemiology* **41**, 714-725 (2017).
91. Ahola-Olli, A.V. *et al*. Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am J Hum Genet* **100**, 40-50 (2017).
92. Greenbaum, J. & Deng, H.W. A Statistical Approach to Fine Mapping for the Identification of Potential Causal Variants Related to Bone Mineral Density. *J Bone Miner Res* **32**, 1651-1658 (2017).
93. Wyss, A.B. *et al*. Multiethnic Meta-analysis Identifies New Loci for Pulmonary Function. *bioRxiv* (2017).

94. Kocarnik, J.M. *et al*. Discovery, fine-mapping, and conditional analyses of genetic variants associated with C-reactive protein in multiethnic populations using the Metabochip in the Population Architecture using Genomics and Epidemiology (PAGE) study. *Hum Mol Genet* (2018).
95. Schork, A.J. *et al*. A genome-wide association study for shared risk across major psychiatric disorders in a nation-wide birth cohort implicates fetal neurodevelopment as a key mediator. *bioRxiv* (2017).
96. Bipolar, D., Schizophrenia Working Group of the Psychiatric Genomics Consortium. Electronic address, d.r.v.e., Bipolar, D. & Schizophrenia Working Group of the Psychiatric Genomics, C. Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705-1715 e16 (2018).
97. Pardinas, A.F. *et al*. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* **50**, 381-389 (2018).
98. Kemp, J.P. *et al*. Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat Genet* **49**, 1468-1475 (2017).
99. Doherty, A. *et al*. GWAS identifies 10 loci for objectively-measured physical activity and sleep with causal roles in cardiometabolic disease. *bioRxiv* (2018).
100. Lareau, C.A. *et al*. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *bioRxiv* (2018).