Machine learning approaches for high-dimensional genome-wide association studies

Muhammad Ammar Malik

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2022

UNIVERSITY OF BERGEN

Machine learning approaches for high-dimensional genome-wide association studies

Muhammad Ammar Malik

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 24.08.2022

Year: 2022

Title: Machine learning approaches for high-dimensional genome-wide association studies

Name: Muhammad Ammar Malik

Print: Skipnes Kommunikasjon / University of Bergen

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

"Glory be to You! We have no knowledge except what You have taught us. You are truly the All-Knowing, All-Wise." (Qur' ān 2:32)

Dedicated to my parents, Saira Malik and Muhammad Imran;
to my wife, Sumbul Sultana, and my son, Ibraheem;
without their prayers, love and support
this journey would have been impossible.

Scientific environment

The work presented in this thesis was carried at Computational Biology Unit (CBU), Department of Informatics, Faculty of Mathematics and Natural Sciences, Unit of Bergen (UiB).

I was supervised by Prof. Dr. Tom Michoel, and co-supervised by Prof. Dr. Inge Jonassen (Head of Department of Informatics) and Prof. Dr. Alexander Lundervold (professor in the Computer Science group at the Department of Computer science, Electrical engineering and Mathematical sciences, Faculty of Engineering and Science, Western Norway University of Applied Sciences).

During the course of my PhD, I have been member of National Research School in Bioinformatics, Biostatistics and Systems Biology (NORBIS) and Digital Life Norway (DLN) research school.

Acknowledgements

First and foremost, all praises and thanks to Allah Almighty. Nothing happens in this world without His will. After that, I would like to thank my supervisor Tom Michoel. I am truly thankful to him for the way he helped and guided me throughout my PhD journey. He has been extremely kind, and his appreciation of my small efforts was really motivational for me. I am also thankful to my co-supervisors, Inge Jonassen and Alexander Lundervold, for their useful advice.

I would also like to thank the University of Bergen and especially the Department of Informatics for providing me an excellent environment to pursue my PhD. The staff at the department were very helpful all the time.

I would also like to thank my previous supervisors and teachers, especially Dr Ghazanfar Monir and Dr Imran Naseem, back in Pakistan. Both of them are the main reasons for me pursuing my PhD studies. Perhaps without their guidance, I would not have made it this far.

Lastly, but most importantly, my family deserve a special mention here. Their prayers and support helped me reach this far. My wife was really supportive throughout my PhD. Without her love, understanding and support, this journey would have been very difficult. She has been an excellent mother to my beloved son, who used to miss me a lot whenever I was busy. I am also extremely thankful to my parents and brother in Pakistan for their love and prayers. I have met them only twice since 2016, but their constant support has always made me stronger.

Abstract (English)

Genome-wide association studies (GWAS) aim to find statistical associations between genetic variants and traits of interests. The genetic variants that explain a lot of variation in genome-wide gene expression may lead to confounding in expression quantitative trait loci (eQTL) analyses. To account for these confounding factors, in **Article I** we proposed LVREML, a method conceptually analogous to estimating fixed and random effects in linear mixed models (LMM). We showed that the maximum-likelihood latent variables can always be chosen orthogonal to the known factors (such genetic variants). This indicates that the maximum-likelihood variables explain the sample covariances that is not already explained by the genetic variants in the model.

For identifying which traits are effected by the identified genetic variants, we need to reverse the functional relation between genotypes and traits. In this regard, multi-trait approaches are more advantageous than studying the traits individually. The multi-trait approaches benefit from increased power from considering cross-trait covariances and reduced multiple testing burden because a single test is needed to test for associations to a set of traits. In **Article II**, we analyzed various machine learning methods (ridge regression, Naive Bayes/independent univariate correlation, random forests and support vector machines) for reverse regression in multi-trait GWAS, using genotypes, gene expression data and ground-truth transcriptional regulatory networks from the DREAM5 SysGen Challenge and from a cross between two yeast strains to evaluate methods.

In **Article III**, we extended the above approach to human dataset. An important difference between data from **Article II** and **Article III** is that we do not have ground-truth data available for the latter. We used the genotype and brain-imaging features extracted from the MRIs obtained from the ADNI database. The results from both **Article II** and **Article III** showed that the genotype prediction performance varied across genetic variants. This helped in identifying genomic regions that are associated with high number of traits in high-dimensional phenotypic data. We also observed that the feature coefficients of fitted machine learning models correlated with the strength of

association between variants and traits. Our results also showed that non-linear machine learning methods like random forests identified genetic variants distinct from the linear methods. In particular, we observed in **Article III** that random forest was able to identify single-nueclotide-polymorphisms (SNPs) that were distinct from the ones identified by ridge and lasso regression. Further analysis showed that the identified SNPs belonged to genes previously associated with brain-related disorders.

Abstract (Norwegian)

Formålet med Genome-wide association studies (GWAS) er å finne statistiske sammenhenger mellom genetiske varianter og egenskaper av interesser. De genetiske variantene som forklarer mye av variasjonene i genomfattende genekspresjoner kan medføre konfunderende analyser av kvantitative egenskaper ved ekspresjonsplasseringer (eQTL). For å betrakte konfunderende faktorene, presenterte vi LVREML-metoden i **artikkel I**, en metode som er konseptuelt analogt med å estimere faste og tilfeldige effekter i Lineære Blandede modeller (LMM). Vi viste at de latente variablene med "Maximum likelihood" alltid kan velges ortogonalt til de kjente faktorene (som genetiske variasjoner). Dette indikerer at "Maximum likelihood" variablene forklarer utvalgsvariansene som ikke allerede er forklart av de genetiske variantene i modellen.

For å kartlegge hvilke egenskaper som påvirkes av de identifiserte genetiske variantene, må vi reversere den funksjonelle relasjonen mellom genotyper og egenskaper. I denne sammenhengen er en "multi-trait" metode mer fordelaktige enn å studere egenskapene individuelt. "Multi-trait"-metoden drar nytte av økt kapasitet som følge av å vurdere kovarianser på tvers av egenskaper, og redusert multiple tester, fordi det trengs en enkelt test for å teste for sammenhenger til et sett med egenskaper. I **artikkel II** analyserte vi ulike maskinlæringsmetoder (Naive Bayes/independent univariate correlation, random forests og support vector machines) for omvendt regresjon i multi-trekk GWAS, ved bruk av genotyper, genuttrykksdata og "ground-truth" transcriptional regulatory networks fra DREAM5 SysGen Challenge og fra en krysning mellom to gjærstammer for å evaluere metoder.

I artikkel III utvidet vi metoden ovenfor til å behandle menneskelig data. En viktig forskjell mellom data fra **artikkel II** og **artikkel III** er at vi ikke har "Ground-truth" data tilgjengelig for sistnevnte. Vi brukte genotypen og Magnetresonanstomografi (MRI) data hentet fra ADNI-databasen. Resultatene fra både **artikkel II** og **artikkel III** viste at resultat av genotypeprediksjon varierte på tvers av genetiske varianter. Dette hjulpet med å identifisere genomiske regioner som er assosiert med stort

antall egenskaper i høydimensjonale fenotypiske data. Vi observerte også at koeffisientene til maskinlæringsmodeller korrelerte med styrken til assosiasjonene mellom varianter og egenskaper. Resultatene våre viste også at ikke-lineære maskinlæringsmetoder som "random forests" identifiserte genetiske varianter tydeligere enn de lineære metodene. Spesielt observerte vi i **artikkel III** at "random forests" var i stand til å identifisere enkeltnukleotidpolymorfismer (SNP-er) som var forskjellige fra de som ble identifisert "ridge" og"lasso" regresjonsmetodene. Ytterligere analyse viste at de identifiserte SNP-ene tilhørte gener som tidligere var assosiert med hjernerelaterte lidelser.

List of publications

1. Malik MA. and Michoel T. (2022), *Restricted maximum-likelihood method for learning latent variance components in gene expression data with known and unknown confounders*, G3 **12**, 2, 2022; jkab410, https://doi.org/10.1093/g3journal/jkab410

2. Malik MA., Ludl AA., and Michoel T. (2022), *High-dimensional multi-trait GWAS by reverse prediction of genotypes using machine learning methods*, (Accepted at Computational Intelligence Methods for Bioinformatics and Biostatistics CIBB 2021, Extended version submitted at Springer Lecture Notes in Bioinformatics), https://doi.org/10.48550/arXiv.2111.00108

3. Malik MA., Lundervold AS. and Michoel T. (2022), *rfPhen2Gen: A machine learning based association study of brain imaging phenotypes to genotypes* (Submitted at 21st European Conference on Computational Biology ECCB 2022), https://doi.org/10.48550/arXiv.2204.00067

Contents

Chapter 1

Background

1.1 Genetic markers and single-nucleotide polymorphisms

A genetic marker can be described as detectable variation in a deoxyribonucleic acid (DNA) sequence with a known physical location on a chromosome. Genetic markers can be used to identify individuals or populations. Moreover, they can be helpful in studying the relationship between a disease and a gene. Single-nucleotide polymorphisms (SNPs) are one of the well-known genetic markers. SNPs are the most frequently occurring type of genetic variation in the human genome. Each SNP represents a variation in a single DNA building block, called a nucleotide. For example, at a particular genomic location, most individuals might have sequence **GCCTC**, but some individuals instead might have the sequence **GCATT** . So, it is possible to have either nucleotide **C** or **A** at the third position. The third position, in this case, is considered a SNP. Each of two or more variants of a gene at a locus is called an allele. The majority of SNPs in humans are biallelic [1] indicating that only two possibilities of the nucleotide can occur. Therefore, **C** and **A** are the possible alleles for the biallelic SNP in the example cited above. The less common allele is known as a minor allele, and the frequency at which it occurs in a given population is called minor allele frequency (MAF) [2]. SNPs are commonly used as genetic markers to unravel the genetic basis of inherited diseases or traits. Although SNPs are the most common form of variation, various other genetic variations are also present in the human genome, such as structural variants, including copy-number variants, translocations, or inversions of relatively large DNA segments [3]. All individuals inherit two copies of each gene, one from the mother and one from the father. Therefore, for a SNP that has al-

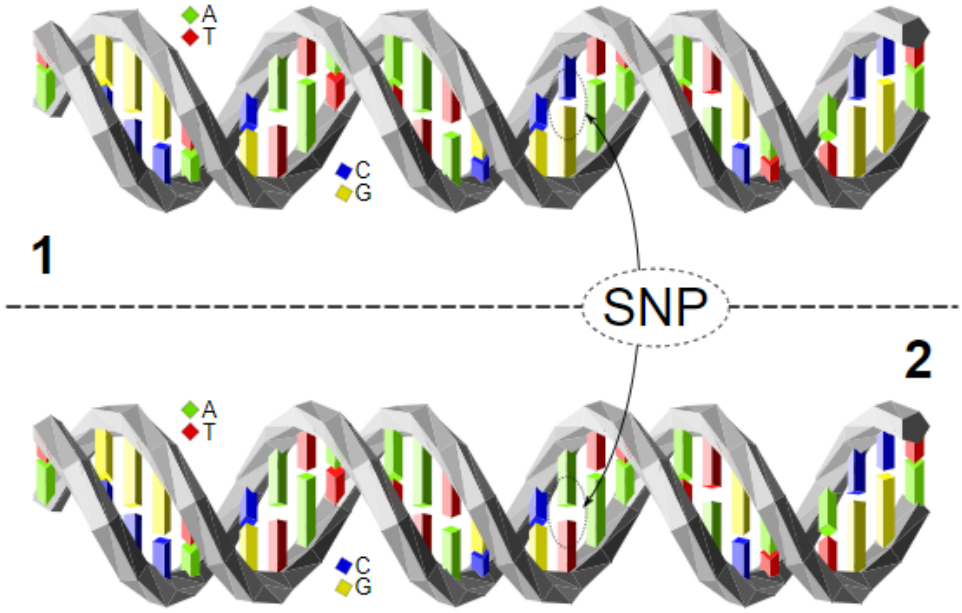


Figure 1.1: Illustration of a SNP. The two DNA molecules are different at a single base-pair location, where the upper DNA molecule has a **C** nucleotide and the lower has an **A**. SNP model by David Eccles (Gringer) [5]

leles C and A, three genotypes are possible; **CC**, **CA**, and **AA**. In the basic form of genetic association analysis, the three genotypes can be used as exposure categories to investigate the association between genes and inherited disease or a trait [4].

1.2 Genome-wide Association Studies (GWAS)

The completion of the human genome sequence, huge improvements in genotyping technology and the initiation of International HapMap Project [6], have set the stage for genome-wide association studies (GWAS). GWAS map the effect of genetic variants on disease risk or severity at single base-pair resolution across the entire human genome [7]. GWAS generally focus on the association between single-nucleotide-polymorphisms (SNPs) and traits, e.g. disease status or a quantitative phenotype such as height, biomarker concentrations or even gene expression [3].

A variation at single-base pair between individuals (SNPs) may cause variation between individuals' traits or phenotypes, for example, a disease risk or physical properties [8]. A GWA study tries to answer if the allele of a genetic variant is found more

often than expected in individuals with the phenotype of interest (e.g. the disease being studied).

Consider a GWA study with a case-control setup. Such a study aims to compare a healthy control group versus a case group affected by a disease. For each of the common known SNPs, it is then investigated if the allele frequency is significantly different between the case and the control group. In such a scenario, a common unit for reporting effect sizes is the *odds ratio*. In the context of GWAS, the odds ratio is the ratio of the percentage of cases among individuals with a specific allele versus individuals who do not have that same allele. The odds ratio is higher than 1 whenever the allele frequency in the cases is higher than the controls. Finding odds ratios that are significantly different from 1 is the objective of the GWA study as it indicates that a SNP is associated with disease [9].

1.2.1 High-dimensional biology

The systematic study of an organism's genome is known as *genomics*. The human genome consists of 3.2 billion bases [10] and an estimated 20000 protein-coding genes. Traditionally, the genes used to be analyzed individually. However, there has been substantial advancement in microarray technology. DNA microarrays can be used to measure the expression levels of a large number of genes simultaneously [11]. Gene expression is usually measured by quantifying levels of the gene product (often a protein) [12]. Another recent technique to quantify RNA and changes in gene expression is RNA-Seq [13]. RNA-Seq utilizes next-generation sequencing (NGS), which refers to any of the several high-throughput approaches for DNA sequencing that uses the concept of massive parallel processing. Hundreds of megabases of nucleotide sequence reads can be generated using NGS parallelization. This is important because it results in a drastic increase in available sequence data and a significant decrease in the cost of sequencing [14].

1.2.2 Multi-trait GWAS

Recent technological advances have made measuring a high number of traits (e.g. gene expression levels.) in an individual feasible and cost-effective. Therefore, in order to investigate the association between genetic variants and multiple traits simultaneously, multi-trait GWAS approaches are necessary. The traditional univariate linear regression or analysis of variance (ANOVA) based approaches test for associ-

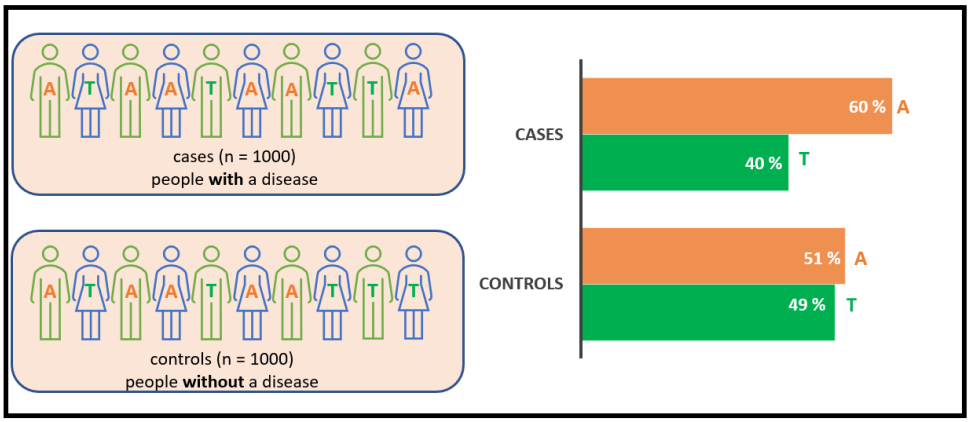


Figure 1.2: Hypothetical results from a case-control GWAS investigating genetic variants associated with a disease. It demonstrates that more people with disease have base "**A**" at this position compared to the control group.

ation between each trait and each genetic variant separately. However, this ignores the cross-trait covariances, and it causes a massive multiple-testing problem. Multiple testing problem occurs when several hypotheses are tested simultaneously, resulting in a large number of false alarms (i.e. most of the significant results are false). Therefore, studies that consider each trait individually suffer from a significant loss of statistical power. Multi-trait GWAS approaches, where multiple correlated traits are studied simultaneously, aim to address both the problems posed by individual trait studies [15, 16, 17].

1.3 Expression quantitative trait loci (eQTL) and GWAS

An expression quantitative trait loci (eQTL) is a genomic locus that explains the genetic variance of a gene expression phenotype. A typical eQTL analysis is based on testing the direct association between genetic variants and gene expression levels. This association analysis can be performed proximally or distally to the gene. These associations are helpful in revealing biochemical processes underlying living systems, discovering the genetic factors causal to specific diseases and determining pathways affected by them. A major advantage of eQTL mapping using the GWAS approach is that it allows the identification of new functional loci without the need of any prior knowledge about specific cis or trans regulatory regions [18]. Regulatory variants in eQTL mapping literature are typically characterized as either cis or trans acting, depending on the physical distance from the gene they regulate and reflecting the pre-

dicted nature of the interaction. Cis-eQTLs commonly refer to genetic variations that act on local genes, and trans-eQTLs are those that act on distant genes and genes residing on different chromosomes. Trans-eQTLs usually have smaller effect sizes than cis-eQTL. However, trans-eQTLs have been known to be relevant for complex traits as compared to stronger cis-eQTL effects [19]. Increasing evidence suggests that single nucleotide polymorphisms (SNPs) associated with complex traits are more likely to be eQTLs than would be expected by chance alone [20]. Expression QTL analyses are useful to identify hotspots (genomic regions affecting multiple transcripts), construct causal networks and select genes and phenotypes for clinical trials [21].

Various approaches for eQTL analysis have been used in the literature. Typically eQTL studies perform separate testing for each transcript-SNP pair [21]. Some of the approaches used to identify the association between expression and genotype include linear regression, analysis of variance (ANOVA) models, generalized linear and mixed models, Bayesian regression [22], and models considering pedigree [23] and latent variables [24]. So far, various methods have been developed for finding the group of SNPs associated with the expression of a single gene [25, 26, 27, 28].

Typically eQTL analyses are known to be computationally intensive as it involves testing for the association of billions of transcript-SNP (single-nucleotide polymorphism) pairs. This issue is further aggravated in modern eQTL datasets having genotype measured over millions of SNPs and gene expression over tens of thousands of transcripts. The separate tests for each transcript-SNP pair in such a dataset would result in over ten billion tests [21]. Furthermore, it has been shown that non-linear methods can be exceptionally slow for large datasets [29, 30, 31]. Despite these limitations, several approaches have been proposed recently for faster eQTL analysis on larger datasets [21, 32, 33].

1.4 Confounding in GWAS

It is a well-known fact that ”correlation does not imply causation” [34]. Confounding can be defined as a spurious association between an exposure and an outcome caused by an independent factor associated with both exposure and outcome [35]. In a GWA study, a trait and a genetic variant may be associated, but the association is not due to a causal link between the trait and the variant. The correlation may be due to some confounding factors (e.g. batch effects, genetic factors in population-based studies, or cell-cycle stage in single-cell studies). To take a classic example, a GWAS for the skill with chopsticks carried out in San Francisco might identify human leukocyte antigen

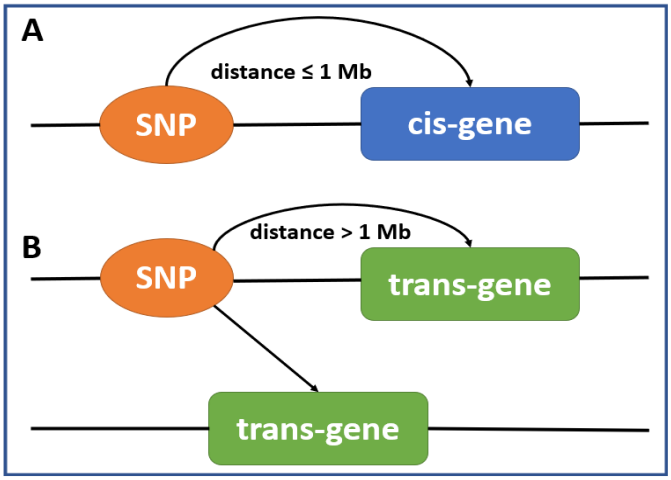


Figure 1.3: Graphical representation of eQTLs. **A.** cis-eQTL, **B.** trans-eQTL.

A1 (HLA-A1) as an allele associated with chopstick skill simply because this allele is more common in people of East Asian origin [36]. Population structure is one of the causes of confounding in GWAS studies. Population structure is referred to as relatedness among individuals [37]. Linear mixed models are widely used to mitigate or correct the effect of these confounding factors [38, 39, 40, 41].

1.4.1 Linear mixed models

Linear mixed models (LMM) are a widely used technique for correcting confounding due to population structure [42, 43, 40, 44, 38]. In a GWAS context, it is assumed that an individual's trait value is a linear function of fixed and random effects, where the random effects are normally distributed with a covariance matrix determined by the genetic similarities between individuals, hence accounting for confounding in the trait data.

A linear mixed-effects models is of the form:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mu + \epsilon \tag{1.1}$$

where $\mathbf{y}$ represents response variable, $\mathbf{X}$ represents explanatory variable, $\mathbf{Z}$ represents design matrix for random effects. β stands for parameter vector for fixed effects, μ stands for vector of random effects and ϵ stands for observation noise.

The random effect vector μ, and the noise vector ϵ, are assumed to have the following

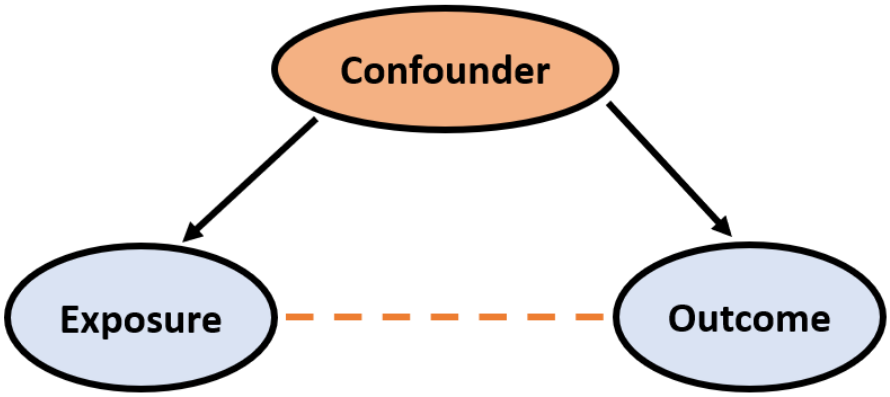


Figure 1.4: A confounder is a spurious factor falsely implying causation between exposure and outcome variables.

independent prior distribution:

$$\mu \sim \mathcal{N}(0, \sigma^2 \mathbf{D}(\theta))$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$$

where $\mathbf{D}$ is a symmetric and positive semidefinite matrix, parameterized by a variance component vector θ, $\mathbf{I}$ represents an identity matrix, and σ^2 is the error variance.

In this model, the parameters that need to be estimated are the fixed effect coefficient β, and the variance components θ and σ^2. Two popular approaches to estimate these parameters are maximum likelihood and restricted maximum likelihood (REML) [45].

1.5 Approaches for multi-trait GWAS

Generally, in multi-trait GWAS methods, a single genetic variant (SNP) is considered at a time. Let us represent it by a random variable Y. Also consider p traits represented by random variables $X_1, X_2, \ldots, X_p$ taking real values. We define a "forward" multi-trait association model probabilistically through a conditional distribution $p(X_1, \ldots, X_p \mid Y)$, which corresponds to the natural direction where variation in Y causes variation in the X_i. Using Bayes' formula, we can write the same model in the reverse causal direction using Y as the dependent variable:

$$P(Y \mid X_1, \ldots, X_p) = P(X_1, \ldots, X_p \mid Y) \frac{P(Y)}{P(X_1, \ldots, X_p)} \tag{1.2}$$

where $P(Y)$ and $P(X_1, \ldots, X_p)$ are prior distributions. Conversely, a forward model $P(X_1, \ldots, X_p \mid Y)$ can be obtained from a reverse model $P(Y \mid X_1, \ldots, X_p)$ using the same formula.

We have data in the form of independent random samples from the joint distribution $P(Y, X_1, \ldots, X_p)$ in n individuals, represented by a genotype vector $\mathbf{y} \in \mathbb{R}^n$ and trait vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p \in \mathbb{R}^n$, which we gather in a matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$. The log-likelihood of observing the data is the log-probability density

$$\begin{aligned} \mathcal{L} = \log p(\mathbf{y}, \mathbf{X}) &= \log \prod_{j=1}^{n} p(y_j, x_{j1}, \ldots, x_{jp}) \\ &= \sum_{j=1}^{n} \log p(y_j, x_{j1}, \ldots, x_{jp}), \end{aligned}$$

which can be expressed in terms of the forward or reverse conditional probabilities depending on the type of model being fit. We now review how existing as well as newly proposed, and low-dimensional as well as high-dimensional multi-trait GWAS methods fit within this framework.

1.5.1 Univariate tests

The simplest method for multi-trait GWAS in the high-dimensional setting consists of testing each trait for association with the genetic variant independently. In this case we fit, by maximum-likelihood, a model $p(x_i \mid y)$ for each trait X_i independently using a linear model

$$p(x_i \mid y) = \mathcal{N}(\mu_y, \sigma^2)$$

a normal distribution with mean μ_y dependent on the genotype value y. This corresponds to the multi-trait model

$$p(x_1, \ldots, x_p \mid y) = \prod_{i=1}^{p} p(x_i \mid y)$$

Using Bayes' rule eq. (1.2), we obtain

$$P(y \mid x_1, \ldots, x_p) = p(x_1, \ldots, x_p \mid y) \frac{P(y)}{p(x_1, \ldots, x_p)}$$
$$\propto P(y) \prod_{i=1}^{p} p(x_i \mid y)$$

where $P(y)$ is the prior probability (background frequency) of observing genotype class y. This is the formula for a *Naive Bayes classifier* of the genotype y given features x_i. Naive Bayes classifiers are family of "probabilistic classifiers" based on Bayes' theorem with the "naive" assumption that every pair of features given the value of the class variable is conditionally independent [46]. In the univariate approach, statistical tests are carried out to determine whether a genotype-dependent model $p(x_i \mid y)$ is more likely or not than a model where the trait is independent of the genotype. This is equivalent to doing a feature selection to determine which traits to include in the naive Bayes classifier.

1.5.2 Canonical correlation analysis

MV-PLINK [16] is a multivariate method based on Canonical Correlation Analysis (CCA). Given two sets of random variables $(X_1, X_2, \ldots, X_p)$ and $(Y_1, Y_2, \ldots, Y_q)$, CCA finds linear coefficients $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ that maximize the correlation

$$\rho(\mathbf{a}, \mathbf{b}) = \operatorname{corr}\left(\sum_{i=1}^{p} a_i X_i, \sum_{j=1}^{q} b_j Y_j\right)$$

It can be shown [1] that if $q = 1$, then the maximizing coefficients $\hat{\mathbf{a}}$ are given by $\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where $\mathbf{X}$ and $\mathbf{y}$ are the data sampled from the joint distribution $P(Y, X_1, X_2, \ldots, X_p)$. These are the same coefficients that would be obtained from a *linear regression* model where Y is modelled as a linear function of the predictors $(X_1, X_2, \ldots, X_p)$, or from the maximum-likelihood solution of a reverse probabilistic model

$$p(y \mid x_1, \ldots, x_p) = \mathcal{N}\left(\sum_{i=1}^{p} a_i x_i, \sigma^2\right). \tag{1.3}$$

[1] see supplementary information for **Article II** section S1 in Appendices

1.5.3 Reverse logistic regression

MultiPhen [47] is a method that is described directly in terms of a model to predict genotypes from multiple traits, using proportional odds *logistic regression*, that is, instead of fitting the genotype class probabilities $P(y = m \mid x_1, \ldots, x_p)$, for $m = 0, 1, 2$ (for biallelic data), the method fits

$$P(y \leq m \mid x_1, \ldots, x_p) = \frac{1}{1 + e^{-\alpha_m - \sum_{i=1}^{p} \beta_i x_i}}$$

Then a likelihood ratio test is used to determine if this model fits the data better than a model where $\beta_1 = \cdots = \beta_p = 0$, thus carrying out a single test for each genetic variant, testing whether the variant is associated with *any* of the traits using the logistic regression model.

1.5.4 L2-Regularized reverse regression

Expressing CCA for multi-trait GWAS as a linear regression of the variant genotype on the trait values eq. (1.3) immediately leads to a generalization to the high-dimensional setting in the form of regularizing the regression coefficients, that is, augmenting eq. (1.3) with a prior distribution $p(a_i) = \mathcal{N}(0, \alpha)$, $i = 1, \ldots, p$.

Finding the maximum-likelihood values of the regression coefficients is equivalent to L_2-regularized or ridge regression. This is the approach followed by [48], who combined it with a likelihood ratio test to determine whether the fitted model is more likely than a model where the genotype is independent of the traits ($a_i = 0$ for all i) and obtain a single association p-value for each variant.

1.5.5 Reverse genotype prediction using machine learning methods

From the above, we conclude that existing multi-trait GWAS methods can be described as reverse genotype prediction methods. From this perspective, L_2-regularized linear regression is but one of many established machine learning methods that could be used to predict an outcome variable Y from a high number of predictors or features X_i, $i = 1, \ldots, p$.

	Supervised learning	Unsupervised learning	Reinforcement learning
Definition	Machine learns using labeled input and output data.	Machine learns without labeled data.	A computer program interacts with its environment by performing actions & learning from errors or rewards.
Goal	Learn a general rule to map the input to output.	Discover hidden patterns in data.	The program navigates its problem space and tries to maximize the rewards based on the feedback.
Problem type	Regression & classification	Association & clustering	Reward-based.

Table 1.1: Comparison of three main machine learning approaches

1.6 Machine learning

To easily understand the concept of machine learning (ML), conisder the following formalism suggest by Tom Mitchell in his book *Machine Learning* [49]:

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Therefore, machine learning is the study of *computer programs/algorithms* that can improve automatically through experience by using the data. Machine learning algorithms build a model based on some available data, known as *training data*, to make predictions on unseen data, known as *testing data*.

Machine learning approaches can be divided into three main categories, depending on the nature of the "signal" or "feedback" available to the learning system, namely *supervised learning*, *unsupervised learning*, and *reinforcement learning* (Table 1.1).

Since, the scope of this thesis is limited to *supervised learning*, in the following subsection we give brief introduction of *supervised learning* algorithms used in our study.

1.6.1 Linear Models

Given a vector of inputs $X^T = (X_1, X_2, ..., X_p)$, a linear model predicts the output Y using:

$$\hat{Y} = \hat{\beta_0} + \sum_{j=1}^{p} X_j \hat{\beta_j} \tag{1.4}$$

The term $\hat{\beta_0}$ represents intercept and is also known as the *bias* in machine learning. If constant variable 1 is added in input X, then we can include the $\hat{\beta_0}$ term in the vector of coefficients $\hat{\beta}$. If we denote $\mathbf{X}$ by $N \times (p+1)$ matrix with each row an input vector (with a 1 in first position), and similarly let $\mathbf{y}$ be the N-vector of outputs in the training set then, the linear model can be written in vector form as inner product:

$$\hat{\mathbf{y}} = \mathbf{X}^T \hat{\beta} \tag{1.5}$$

If we are modeling K outputs then $\hat{Y}$ can be K-vector. In that case β would be $p \times K$ matrix of coefficients. In the $(p+1)$-dimensional input-output space, $(X, \hat{Y})$ represents a hyperplane. Viewed as p-dimensional input space, $f(X) = X^T\beta$ is linear, and the gradient $f'(X) = \beta$ is a vector in linear in input space pointing in the steepest uphill direction.

The most popular method to fit the linear model to set of training data is *least squares*, in which we pick the coefficients $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ to minimize the residual sum of squares (RSS):

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \end{aligned} \tag{1.6}$$

The eq. (1.6) can be written in matrix notation as:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \tag{1.7}$$

Differentiating eq. (1.7) w.r.t β and setting it to 0, we get the estimates for β as:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{1.8}$$

Using the estimated $\hat{\beta}$ the output $\hat{\mathbf{y}}$ at training inputs $\mathbf{X}$ and training outputs $\mathbf{y}$ can be given as:

$$\hat{\mathbf{y}} = \mathbf{X}^T\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{1.9}$$

Ridge regression

Ridge regression (also known as **L2 regularization**) shrinks the regression coefficients (β) by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{1.10}$$

The amount of shrinkage is controlled by a complexity parameter $\lambda \geq 0$. Larger value of λ implies greater amount of shrinkage. The coefficients are shrunk toward zero (and each other). In a standard linear regression with p variables, the degree-of-freedom of the fit is p, i.e. free parameters. The intuition behind this is that even-though all p coefficients in a ridge model will be non-zero, they are fit in a restricted manner controlled via λ. Therefore the degree of freedom is p when $\lambda = 0$ (no regularization, and it approaches ∞ when $\lambda \to \infty$. The ridge solutions are not equivariant under scaling of the input, so the input needs to be standardized before solving eq. (1.10). By centering the input (i.e. subtract the mean from each input), the intercept term β_0 can be left out. Now the input matrix $\mathbf{X}$ has p (rather than $p+1$) columns. The eq. (1.10) can be written in matrix form as:

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T((\mathbf{y} - \mathbf{X}\beta) + \lambda(\beta^T)(\beta) \tag{1.11}$$

Solving eq. 1.11 for β we get:

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{1.12}$$

where **I** is a $p \times p$ identity matrix.

1.6.2 Support vector machines (SVM)

Given a training set of N points of the form $(x_1, y_1), ..., (x_N, y_N)$, where y_i indicates the class to which the point x_i belongs (i.e.$y_i \in \{-1, 1\}$) and each x_i is p-dimensional real vector. The aim of a linear SVM is to find the *maximum-margin hyperplane* dividing the group of points of x_i based on the class of y_i. The *maximum-margin* can be defined as the maximum distance between the hyperplane and the nearest point x_i from either classes. The hyperplane is defined by:

$$\{x : f(x) = x^T\beta + \beta_0\} \tag{1.13}$$

where β is a unit vector (i.e. $||\beta|| = 1$). To find the hyperplane with *maximum-margin* between the training points belonging to each class, we need to solve the following optimization problem:

$$\max_{\beta,\beta_0,||\beta||=1} M \tag{1.14}$$
$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq 1$$

where M is distance from the decision boundary to the hyperplane for either class.

It is shown in [46] that above optimization problem can be rewritten more conveniently as:

$$\min_{\beta,\beta_0} ||\beta|| \tag{1.15}$$
$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq 1$$

The detailed discussion of SVM is beyond the scope of this thesis therefore further details about computing support vector classifier can be found in [46].

1.6.3 Random Forests

Random forest is a class of supervised machine learning algorithms widely used for classification and regression problems. Random forest utilize the modified version of

bagging technique for building many uncorrelated trees and averaging them. Bagging refers to ensemble machine learning technique, where a set of weak learners is combined for creating a strong learner that achieves better performance than a single learner. Bagging reduces the variance (error resulting from sensitivity to small fluctuations in the the training set) by averaging many noisy but nearly unbiased models (i.e. models with lesser erroneous assumptions in the learning algorithms). Since decision trees can capture complex structures in the data, it makes them perfect candidate for bagging. The averaging of trees helps mitigate the noisy nature of trees. The idea in random forests is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables [46]. The random forest regression algorithm can be defined as in Table 1.2.

Table 1.2: **Algorithm for Random forest regression**

1. For $b = 1$ to B:

   (a) Draw a bootstrap sample $\mathbf{Z}\star$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data,
   by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_min is reached.
   i. Select m variables at random from the p variables.
   ii. Pick the best variable/split-point among the m.
   iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

Regression: $\hat{f}_B^{\mathrm{rf}}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Random forests are preferable choice for variety of problems because any transformation of a single variable is implicitly captured by a tree, therefore the random forests do not need scaling of the variables. Moreover, as already mentioned above the random forest uses bagging technique to prevent overfitting (i.e. prevents models to perform extremely well on training data but poorly on unseen test data). Random forests have also been known to deal with missing data [50]. Lastly, feature selection in random forest is relatively straight forward.

Feature importance

Prediction of a response variable from a set of predictor variables is an important task in many scientific fields. However, in many cases, the aim is not only to make accurate predictions but also to identify which predictor variables (explanatory features) are most important in making these predictions. This helps us to understand the underlying process. In case of random forests, the importance of a variable X_j to predict Y is computed by adding up the weighted impurity decrease for all nodes t where X_j is used, averaged over all trees ψ_m (for $m = 1, ..., M$) in the forest as follows:

$$Imp(X_j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t \epsilon \psi_m} 1(j_t = j)[p(t)\Delta(s_t, t)] \tag{1.16}$$

where $p(t)$ is the proportion $\frac{N_t}{N}$ of samples reaching t and where j_t represents the identifier of the variable used for splitting node t. $\Delta(s_t, t)$ represents the impurity decrease. When impurity function being used is Gini index, then this measure is known as *Gini importance*. In case of classification Gini index is defined as follows:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{1.17}$$

Where C represents number of classes and p_i denothes the probability that the sample belongs to ith class [51].

1.7 Neuroimaging genetics

Neuroimaging genetics, also known as imaging genomics or imaging genetics, is a useful tool to investigate the associations between genetic variants and variation in brain structure among individuals [52]. The discovery of biomarkers jointly from imaging and genetic data helps us to better understand the underlying pathological processes of neuropsychiatric and neurodegenerative diseases [53, 54]. Moreover, neuroimaging may help us discover the genetic pathways through which genes affect the above-mentioned diseases by identifying associations between causal genes and variations in brain regions [55, 56]. And lastly, imaging genetics studies have been shown to have increased statistical power when compared with conventional case-control studies and therefore have decreased sample size requirement [57].

Recently a large number of neuroimaging studies have been conducted to explore the

association between neurodegenerative disease and brain structure [58, 59, 52, 60, 61]. Some of these studies have focused on understanding the genetic causes of these diseases (for example, Alzheimer's disease), whereas the other genome-wide association studies (GWAS) focus on identifying the genetic variations that influence brain structure and function. A common issue with most imaging genetics studies is the reduction in either imaging or genetic data (or sometimes both). For example, whole-brain studies have mostly focused on a small number of genetic variants [62, 63, 64, 65], whereas whole-genome studies have focused on a limited number of quantitative imaging traits [66, 67]. This restriction in either genotype or phenotype data can greatly hinder our ability to identify important associations.

1.8 Alzheimer's disease

Alzheimer's disease (AD) is a neurodegenerative disorder and one of the most common forms of dementia [68]. It has an adverse effect on an individual's thinking, behavior and memory [69]. Often, people aging 65 or above are more prone to AD, and it is the fifth leading cause of death among such individuals in the United States [70]. AD is characterized by pathological features such as synaptic loss, neuroinflammation, neuronal cell death, cortical atrophy, and the presence of neurofibrillary tangles and senile plaques. Unfortunately, no standard methods for the diagnosis of AD exist and moreover, no cure or disease-modifying therapy has been developed so far, which is an indication of AD being a major public global health problem. AD starts with a synaptic loss followed by neuronal death and the formation of neurofibrillary tangles and senile plaques at the later stages of the disease. AD is developed with multiple interacting causes over many years. The pathological and clinical complexity and heterogeneity of the disease is the most challenging aspect of AD diagnostic research. Moreover, a clear diagnosis of AD is often difficult due to the simultaneous presence of other older-age neurodegenerative diseases such as vascular dementia (VaD), Lewy body disease (LBD), Parkinson's disease (PD), frontotemporal dementia (FTD), amyotrophic lateral sclerosis (ALS), and tauopathy [71].

1.8.1 Neuroimaging Biomarkers

Neuroimaging biomarkers are among the most widely researched Alzheimer's disease (AD) biomarkers. Various neuroimaging modalities have been applied for the detection of AD biomarkers, such as single-photon emission computed tomography

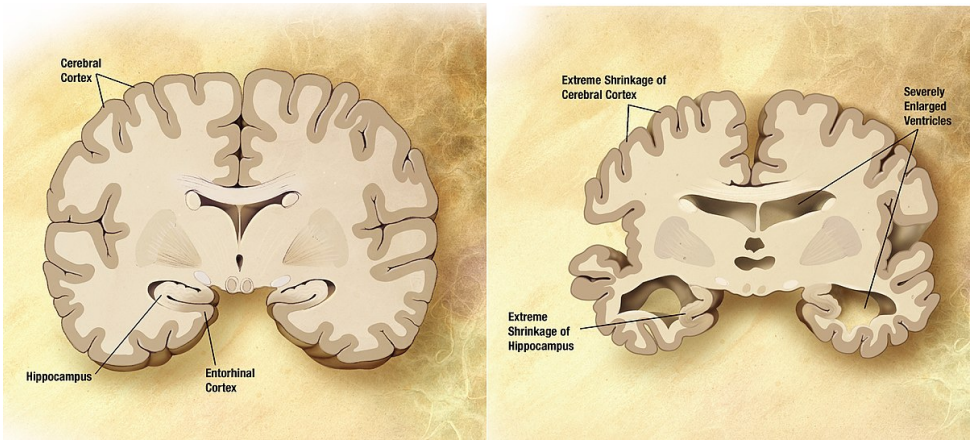


Figure 1.5: Diagram showing changes of the brain caused by Alzheimer's disease [72]

(SPECT), positron emission tomography (PET), computed tomography (CT), magnetic resonance imaging (MRI), and magnetic resonance spectroscopy (MRS). Several neuropathological abnormalities characteristic of AD can be identified using modern neuroimaging methods, including atrophy in specific brain regions or the whole brain (shrinkage), brain Aβ accumulation (amyloid plaques), hyperphosphorylated tau (p-tau) deposition, neuronal damage (loss of neurons), abnormal cerebral blood flow, reduced levels of brain metabolites (indicating reduced activity of the brain), abnormal neural activity, and regional inflammation of the brain. In recent years, neuroimaging has been used for the detection of abnormal neuronal network connectivity, which is believed as the cause of neurological dysfunction in various disorders. Neuroimaging may help in the detection of biomarkers associated with preclinical AD. This may help in identifying high-risk individuals who might benefit from early therapeutic intervention before the widespread neuronal loss [71].

1.9 Performance evaluation

1.9.1 Reverse genotype prediction

For genotype prediction using machine learning models, the phenotype values (gene expression or brain measurements in our case) were treated as explanatory variables, whereas the genotype value of a variant was treated as a response variable. The prediction performance was measured by computing the root mean squared error (RMSE) between the predicted and the actual genotype value of variants.

1.9.2 Trans-eQTL prediction

In cases where ground-truth data were available (Article **II**), trans-eQTL target prediction was done using weights assigned to the features by the machine learning methods: feature importance in case of random forest regression (RFR) and coefficients for support vector regression (SVR) and ridge regression (RR). We computed the area under the receiver operating characteristic (AUROC) curve to measure prediction performance by comparing the weights against the true targets in the ground truth for each variant.

In cases where ground-truth was not available (Article **III**), we simply display the feature weights of the trained machine learning method. We hypothesise that the genetic variants with good prediction performance, the feature weights in the fitted models measure the strength of biological association between a variant and a trait.

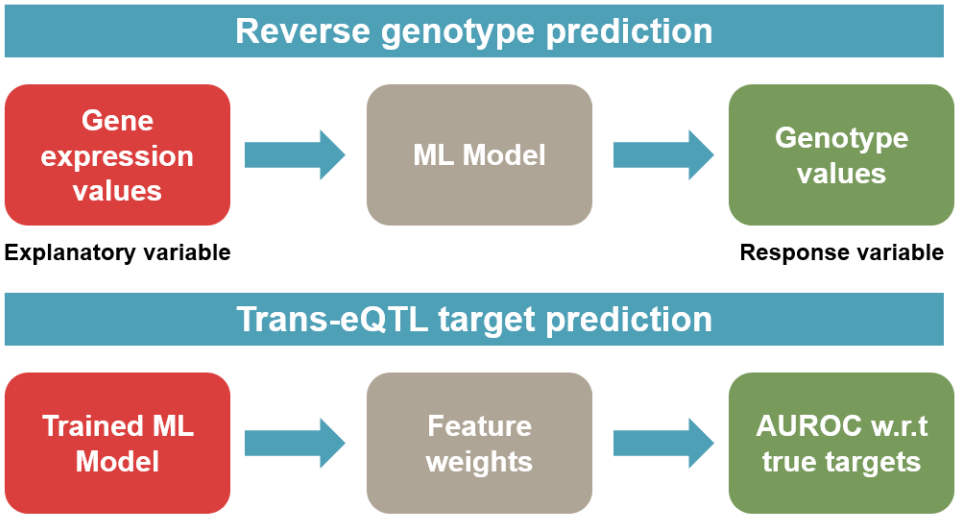


Figure 1.6: Figure summarizing the approach proposed in this thesis for *reverse genotype* and *trans-eQTL* prediction

Chapter 2

Aim of the study

Random effects models are popular statistical models for detecting and correcting spurious sample correlations due to hidden confounders in genome-wide gene expression data. In **Article I**, using the proposed random-effects models, we found out that some genetic variants explain a high proportion of variation. In order to identify the actual traits affected by such variants, we need to consider multi-trait GWAS approaches.

When multiple correlated traits are studied simultaneously, joint, multi-trait approaches can be more advantageous than studying the traits individually due to increased power from taking into account cross-trait covariances and reduced multiple-testing burden by performing a single test for the association to a set of traits [15, 16, 17, 73].

The most commonly used multi-trait GWAS approaches are based on a multivariate analysis of variance (MANOVA) or canonical correlation analysis (CCA) [16]. However, these are applicable only to studies where the number of traits is relatively small, especially in comparison to the number of samples. When analyzing the effects of genetic variants on molecular traits (gene or protein expression levels, metabolite concentrations) or imaging features, we have to deal with a large number, often an order of magnitude or more greater than the sample size, of correlated traits simultaneously. For such studies, the standard procedure is still to conduct univariate linear regression or ANOVA tests for each genetic variant against each trait separately. While efficient algorithms exist to undertake this task [21, 74, 75], the massive multiple-testing problem results in a significant loss of statistical power.

An alternative approach to multi-trait GWAS has been to reverse the functional relation between genotypes and traits, and fit a multivariate regression model that pre-

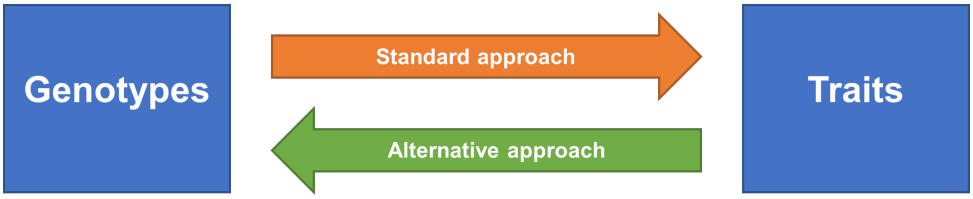


Figure 2.1: The *standard* multi-trait association approach involves regressing traits on genotypes. Whereas the *alternative* approach involves reversing this functional relation.

dicts genotypes from multiple traits simultaneously, instead of the usual approach to regress traits on genotypes. The first study to do this explicitly used logistic regression and showed a significant increase in power compared to univariate methods, without being dependent on assuming normally distributed genotypes like MANOVA or CCA [47]. Although the method as presented in [47] is still only valid when the number of traits is small, extending multivariate regression methods to high-dimensional settings is straightforward. Thus a recent study used L2-regularized linear regression of single nucleotide polymorphisms (SNPs) on gene expression traits to identify trans-acting expression quantitative trait loci (trans-eQTLs), and showed that this approach aggregates evidence from many small trans-effects while being unaffected by strong expression correlations [48]. In a very different application domain, regularized regression of SNP genotypes on longitudinal image phenotypes was used to identify time-dependent genetic associations with imaging phenotypes [59].

Despite these advances, several limitations and open questions remain unanswered in high-dimensional GWAS. Firstly, linear models search for the linear combination of traits that is most strongly associated to the genetic variant, but there is no *a priori* biological reason why only linear combinations should be considered. Secondly, while L2-regularization allows to deal with high-dimensional traits, it does not address the problem of variable selection. For instance, in the case of gene expression, we expect that trans-eQTLs are potentially associated with *many*, but not *all* genes. Indeed, in [48] a secondary set of univariate tests is carried out to select genes associated to trans-eQTLs identified by the initial multi-variate regression. Thirdly, a systematic biological validation and comparison of the available methods is lacking.

Moreover, the approach to extend the GWAS for mapping the effect of genetic variants on molecular phenotypes are often not feasible. This is because molecular measurements are invasive, requiring tissue biopsies, and therefore cannot be obtained for most tissue types during the lifetime of an individual. This is a particular hindrance to the study of complex brain-related disorders. To gain insight in the genetic factors causing these disorders and the pathways through which they might act,

multi-modal imaging of brain structure and function has emerged as a rich source of information. However, analyzing image data using conventional genetic association study approaches requires either to predefine a set of image features that are thought to be of relevance to the disorder being studied, or to test all possible genetic variants against all possible image regions in 3-dimensional grid (volume elements, or "voxels"). Both have obvious drawbacks. Predefining features limits the analysis to features that are already known to be important, and prevents the discovery of genetic associations with unexpected or unknown features. Scanning associations voxel-wise leads to a massive multiple-testing problem, such that only the most extreme associations will survive multiple-testing correction. Furthermore, voxels are defined by a regularly-spaced image grid, and hence voxel-wise analyses are inherently inefficient, and potentially insensitive, to detect associations with unknown features manifested on significantly larger or smaller scales than the predefined grid scale.

Therefore in order to address the above mentioned issues, we considered wider range of machine learning methods including random forest regression, support vector regression, univariate and L2-regularized linear regression, for reverse genotype prediction in multi-trait GWAS in **Article II**. Encouraged by our results we proceeded with brain imaging data where no ground truth data was available (**Article III**) to identify significant SNPs as well as significant imaging phenotypes related to brain function.

2.1 Main hypothesis

In this thesis, we study the effects of genetic variants on high-dimensional phenotype data such as gene expression and imaging data using machine learning approaches.

Our overall hypothesis is that genetic variants whose genotypes can be predicted with higher accuracy are more likely to affect some or all of the traits under consideration than variants whose genotypes cannot be predicted well and that feature weights in the fitted models measure the strength of biological association between a variant and a trait.

Chapter 3

Summary of the articles

3.1 Article I

Restricted maximum-likelihood method for learning latent variance components in gene expression data with known and unknown confounders

Random effects models are popular statistical models for detecting and correcting spurious sample correlations due to hidden confounders in genome-wide gene expression data. In standard eQTL analyses, the genetic variants that affect a lot of genes result in confounding. In the presence of some genetic variants as known confounding factors, simultaneous estimation of contributions from known and latent variance components in random effect models is challenging. The current solutions rely on numerical gradient-based optimizers for maximizing the likelihood function. This is unsatisfactory because the resulting solution is poorly characterized, the relation between the known and latent factors is obscured, and the efficiency of the method may be suboptimal.

In this study, we proved analytically; that by including additional parameters in the proposed model to account for nonzero covariances among the effects of known covariates and latent factors, the latent factors can always be chosen orthogonal to the known confounding factors. In other words, the maximum-likelihood latent variables explain sample covariances not already explained by known factors. This helps in inferring latent factors that are used to correct for correlation structure in the data. Moreover, the latent factors are also used as new data-derived "endophenotypes", that is, determinants of gene expression whose own genetic associations are biologically informative.

The proposed method, called latent variable restricted maximum-likelihood (LVREML), relies on analytic restricted maximum-likelihood (REML) solution. LVREML estimates the latent variables by maximizing the likelihood on the restricted subspace orthogonal to the known factors, and we show that this reduces to probabilistic PCA on that subspace. It then estimates the variance–covariance parameters by maximizing the remaining terms in the likelihood function given the latent variables, using a newly derived analytic solution for this problem. When compared with gradient-based optimizers, our method attains greater or equal likelihood values and can be computed using standard matrix operations. The proposed method results in latent factors that do not overlap with any known factors, and has a runtime reduced by several orders of magnitude. In summary, the proposed LVREML method facilitates the application of random effects modeling strategies for learning latent variance components to much larger gene expression datasets than possible with current methods.

3.2 Article II

High-dimensional multi-trait GWAS by reverse prediction of genotypes

Multi-trait genome-wide association studies (GWAS) use multi-variate statistical methods to identify associations between genetic variants and multiple correlated traits simultaneously, and have higher statistical power than independent univariate analysis of traits. The conventional multi-trait GWAS approaches rely on regressing multiple traits on genotypes simultaneously.

However, an emerging alternative approach to multi-trait GWAS is to reverse the functional relation between genotypes and traits. This reverse regression is a promising approach, especially in high-dimensional settings where the number of traits exceeds the number of samples. Present studies focus only on linear models, which essentially search for the linear combination of traits that are most strongly associated to the genetic variant. But there is no *a priori* biological reason to consider only linear combinations. Moreover, while L2-regularization allows dealing with high-dimensional traits, it does not address the problem of variable selection. Lastly, a systematic biological validation and comparison of the available methods is lacking.

In this paper, we extended this approach and analyzed different machine learning methods (ridge regression, random forests and support vector machines)for reverse regression in multi-trait GWAS, using genotypes, gene expression data and ground-truth transcriptional regulatory networks from the DREAM5 SysGen Challenge and

from a cross between two yeast strains [76] to evaluate methods.

Our results show that genotype prediction performance, in terms of root mean squared error (RMSE), allowed us to distinguish between genomic regions with high and low transcriptional activity. Moreover, model feature coefficients correlated with the strength of association between variants and individual traits; and were predictive of true trans-eQTL target genes, with complementary findings across methods.

3.3 Article III

rfPhen2Gen: A machine learning based association study of brain imaging phenotypes to genotypes

Imaging genetic studies aim to find associations between genetic variants and imaging quantitative traits. Traditional genome-wide association studies (GWAS) are based on univariate statistical tests, but when multiple traits are analyzed together, they suffer from a multiple-testing problem and from not taking into account correlations among the traits.

As discussed in **Article II**, an alternative approach to multi-trait GWAS is to reverse the functional relation between genotypes and traits by fitting a multi-variate regression model to predict genotypes from multiple traits simultaneously. And the current reverse genotype prediction approaches are mostly based on linear models.

Moreover, to the best of our knowledge, a genome-wide analysis of machine learning methods for reverse genotype prediction in human GWAS has not yet been conducted. In this paper, we evaluated random forest regression (RFR) as a method to predict SNPs from imaging QTs and identify biologically relevant associations. We learned machine learning models to predict all 518485 SNPs across the whole genome (selected after the quality control procedure) from the 56 brain imaging quantitative traits using data from the ADNI database.

Our results showed that genotype regression error is a better indicator of permutation p-value significance than genotype classification accuracy. Moreover, SNPs within the known Alzheimer's disease (AD) risk gene APOE had the lowest RMSE for lasso and random forest, but not ridge regression. Moreover, when compared across the whole genome, random forests produced a distinct list of selected SNPs, based on RMSE prediction performance, than the linear methods (ridge and lasso regression), which were highly similar to each other. This indicates that using non-linear multi-

variate GWAS methods may help in identification of genetic variants distinct from those selected by conventional linear methods. Feature selection in random forests identified well-known brain regions associated with AD, like the hippocampus and amygdala, as important predictors of the most significant SNPs. Lastly, extending the analysis to the top 1,000 SNPs predicted by random forests, we observed clustering of image features, showing that groups of variants, not colocated on the genome, tend to associate with similar brain regions or features.

Chapter 4

Data and Software

4.1 Yeast data

The yeast data used in **Article I** and **Article II** was obtained from [76]. The expression data contains expression values for 5,720 genes in 1,012 segregants. The genotype data consists of binary genotype values for 42,052 genetic markers in the same 1,012 segregants. Following [76], we removed batch and optical density effects from the expression data using categorical regression. This was achieved using *statsmodels* python package. The expression data was then normalized to have zero mean and unit standard deviation.

In **Article II**, to match variants to genes, we considered the list of genome-wide significant eQTLs provided by [76] whose confidence interval (of variable size) overlapped with an interval covering a gene plus 1000 bp upstream and 200 bp downstream of the gene position. This resulted in a list of 2884 genes and for each of these genes we defined its matching variant as the most strongly associated variant from the list.

We obtained networks of known transciptional regulatory interactions in yeast (S. cerevisiae) from the YEASTRACT ((Yeast Search for Transcriptional Regulators And Consensus Tracking) [77]. Regulation matrices were obtained from `http://www.yeasttract.com/formregmatrix.php`. We retrieved the ground-truth matrix containing all reported interactions of the type DNA binding and expression evidence. Self regulation was removed from the ground-truths. The Ensembl database (release 83, December 2015) [78] was used to map gene names to their identifiers. After overlaying the ground-truth with the set of genes with matching cis-eQTL, a ground-truth network of 80 transcription factors (TFs) with matching cis-eQTL and 3394 target

genes was obtained.

The expression dataset was then filtered to contain only the genes present in the ground truth network, and ground-truth trans-eQTL sets for the 80 TF-associated cis-eQTL genetic variants were defined as direct targets of the corresponding TFs in the ground-truth network.

4.2 Simulated data

In **Article II**, we also used simulated data obtained from DREAM5 Systems Genetic Challenge A ((`https://www.synapse.org/#!Synapse:syn2820440/wiki/`), generated by the SysGenSIM software [79]. The DREAM data consists of simulated genotype and transcriptome data of synthetic gene regulatory networks. The dataset consists of 15 sub-datasets, where 5 different networks are provided and for each network 100, 300 and 999 samples are simulated. Every sub-dataset contains 1000 genes. We used the networks with 999 samples only.

Each genetic variant in DREAM data, is associated to a unique causal gene that mediates its effect. Therefore, we defined ground-truth trans-eQTL targets for each variant as the causal gene's direct targets in the ground-truth network.

In the DREAM data 25% of the variants acted in cis, meaning they affected expression of their causal gene directly. The remaining 75% of the variants acted in trans. Since the identities of the cis and trans eQTLs are unknown, we computed the P-values of genotype-gene expression associations between matching variantgene pairs using Pearson correlation and selected all genes with P-values less than 1/750 to identify cisacting eQTLs.

4.3 Brain-imaging data

Brain-imaging and genotype data for **Article III** used in this study was obtained from the ADNI (Alzheimer's Disease Neuroimaging Initiative) database (`adni.loni.ucla.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and

early Alzheimer's disease (AD). The baseline T1-weighted MRI images from the four phases of the ADNI study, the Illumina SNP genotyping data, demographic information, APOE genotype, and baseline diagnosis was downloaded from the ADNI database. The demographic information of the samples used in the study can be found in Table 1. Details about the standardized imaging protocols used in ADNI can be found in `https://adni.loni.usc.edu/methods/documents/mri-protocols/`.

Table 4.1: Demographic information of the samples used in the study. CN: Controls, MCI: Mild Cognitive Impairment, AD: Alzheimer's Disease

	CN	**MCI**	**AD**
No. of subjects	211	359	178
Gender(M/F)	112/99	234/125	94/84
Baseline age(years:mean $\pm$ SD)	75.7$\pm$4.9	74.7$\pm$7.3	75.4$\pm$7.3
Education(years:mean $\pm$ SD)	16.0$\pm$2.8	15.7$\pm$3.0	14.6$\pm$3.2
Race (Caucasian/Non-Caucasian)	191/20	325/34	161/17

4.3.1 MRI data and imaging phenotype extraction

We extracted subcortical segmentation and cortical parcellation from the T1-weighted images using FreeSurfer v6.0 [80] to obtain imaging phenotypes. Following [52] we defined 56 volumetric and cortical thickness values mentioned in(Table 4.2).

4.3.2 SNP genotypes

The SNP data from ADNI database were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA, USA). The genotype data consists of 620,901 SNPs. The SNP data was screened using the following quality control (QC) steps: (1) call rate check per subject ($\geq$ 90%) and per SNP marker ($\geq$ 90%), (2) gender check (3) marker removal according to the minor allele frequency (MAF) $\geq$ 5% and (4) Hardy-Weingberg equilibrium (HWE) test of $p \leq 10^{-6}$. The remaining missing genotype values were imputed as the modal values. After the QC procedure, 749 subjects and 518,484 SNPs remained in the data. The APOE gene is one of the important causal genes for AD, but the previously identified APOE SNPs (rs429358/rs7412) were not available on the Illumina array. Therefore, the APOE genotype was coded from the `ADNIMERGE.csv` file prepared by the ADNI study by using the number of APOE-ε4 risk alleles.

4.4 Software

The code for the work in this thesis was written in Python []. Some of the python libraries we utilized in our work include, NumPy [81], Matplotlib [82], Pandas [83], seaborn [84], statsmodels [85] and scikit-learn [86]. The T1-weighted MRI images were processed using FreeSurfer [80]. Whereas, the genotype data from ADNI was

Table 4.2: List of FreeSurfer phenotypes defined as volume or cortical thickness of various region of interests (ROI)[a]

Phenotype description (Phenotype ID)	
Volume of amygdala (AmygVol)	Volume of cerebral cortex (CerebCtx)
Volume of cerebral white matter (CerebWM)	Volume of hippocampus (HippVol)
Volume of inferior lateral ventricle (InfLatVent)	Volume of lateral ventricle (LatVent)
Thickness of entorhinal cotrex (EntCtx)	Thickness of fusiform gyrus (Fusiform)
Thickness of inferior parietal gyrus (InfParietal)	Thickness of inferior temporal gyrus (InfTemporal)
Thickness of middle temporal gyrus (MidTemporal)	Thickness of parahippocampal gyrus (Parahipp)
Thickness of posterior cingulate (PostCing)	Thickness of postcentral gyrus (Postcentral)
Thickness of precentral gyurs (Precentral)	Thickness of precuneus (Precuneus)
Thickness of superior frontal gyrus (SupFrontal)	Thickness of superior parietal gyurs (SupParietal)
Thickness of superior temporal gyrus (SupTemporal)	Thickness of supramarginal gyrus (Supramarg)
Thickness of temporal pole (TemporalPole)	
Mean thickness of caudal anterior cingulate, isthmus cingulate, posterior cingulate, and rostral anterior cingulate (MeanCing)	
Mean thickness of caudal midfrontal, rostral midfrontal, superior frontal, lateral orbitofrontal, and medial orbitofrontal gyri and frontal pole (MeanFront)	
Mean thickness of inferior temporal, middle temporal, and superior temporal gyri (MeanLatTemp)	
Mean thickness of fusiform, parahippocampal, and lingual gyri, temporal pole and transverse temporal pole (MeanMedTemp)	
Mean thickness of inferior and superior parietal gyri, supramarginal gyrus, and precuneus (MeanPar)	
Mean thickness of precentral and postcentral gyri (MeanSensMotor)	
Mean thickness of inferior temporal, middle temporal, superior temporal, fusiform, parahippocampal , and lingual gyri, temporal pole and transverse temporal pole (MeanTemp)	

[a]Each of the 28 phenotypes mentioned corresponds to two phenotypes, one for the left side and the other for the right side.

processed using PLINK [87].

The LVREML software and all data processing and analysis scripts for **Article I** are available at `https://github.com/michoel-lab/lvreml`. LVREML is also available as a Python PyPi package and can be installed using the command `pip install LVREML`.

The code to reproduce the analysis for **Article II** are available at `https://github.com/michoel-lab/Reverse-Pred-GWAS`.

The scripts for reproducing the results in **Article III** are available at `https://github.com/michoel-lab/rfPhen2Gen`.

Chapter 5

Discussion

Combined analysis of genetic and clinical data helps us in discovering the genetic factors that contribute to phenotypic traits (e.g. molecular phenotypes, gene-expression levels, protein expression levels etc.) or common diseases in humans. Analysis of the traits individually ignores the correlations among them. Therefore multi-trait approaches are more advantageous in cases where multiple correlated traits are to be studied simultaneously. Multi-trait approaches also help in reducing the multiple-testing burden because only a single test needs to be performed for association to a set of traits.

In **Article I**, we presented a random-effects model for estimating the contribution of known and latent variance components in gene expression data simultaneously. The known confounders in our study are represented by the genetic variants (SNPs), whereas the latent confounders are hidden factors that need to be inferred by the model. Our results show that eQTL analyses are confounded due to a subset of SNPs that explain a lot of variation in genome-wide gene expression. We propose LVREML, a method that is conceptually analogous to estimating fixed and random effects in linear mixed models, to correct for the confounding factors.

However, in order to figure out which traits are affected by these SNPs, we need to reverse the functional relation between genotypes and traits. We fit a multivariate regression model that predicts genotypes from multiple traits simultaneously instead of the usual approach to regress traits on genotypes. The current approaches that make use of this alternative approach suffer from a few limitations. Firstly, the present approaches use linear models that search for the linear combination of traits that is most strongly associated to the genetic variant, but there is no a priori biological reason why only linear combinations should be considered. Secondly, while

L2-regularization allows dealing with high-dimensional traits, it does not address the problem of variable selection. For instance, in the case of gene expression, we expect that trans-eQTLs are potentially associated with many but not all genes. Thirdly, a systematic biological validation and comparison of the available methods was lacking.

In **Article II**, we addressed these questions by considering a wider range of machine learning methods (in particular, random forests (RF) and support vector machines (SVM)) for reverse genotype prediction from gene expression traits. The basic hypothesis of our study was that true trans-eQTL associations are mediated by transcription regulatory networks. However, our results support the basic hypotheses only partially. We observed that the genotype prediction performance varied across genetic variants. But there was no relation between genotype prediction performance and the number of gene expression traits affected by a variant, nor with the accuracy of predicting individual trans-eQTL target genes from model feature importances or coefficients. The significance of this is that it shows that in the absence of ground-truth information, low RMSE does not always predict variants for which model features will overlap best with true trans-associated genes. This was further illustrated by the fact that random forest regression performed best at the genotype prediction task but performed worst on the trans-eQTL prediction task. The only systematic relation we observed, both in the simulated and the yeast data, was a negative correlation between genotype prediction performance and the number of model features. This suggests that variants with good prediction performance can achieve this performance with a relatively small number of traits.

Although RMSE cannot necessarily be used for the selection of variants with good trans-eQTL prediction performance, our results showed that model feature importances or coefficients were generally predictive of how likely a given gene is a true trans-eQTL target of a given variant. We observed strong predictive performance in simulated data, with more than 75% of variants obtaining an AUROC greater than 80%. But also in yeast data, 15-20% of variants obtained an AUROC greater than 70%.

One of the important goals of multi-trait GWAS is distinguishing between variants associated with a high or low number of traits. Interestingly, we found that only random forest, but not SVR or ridge regression, resulted in models with a wide variation in the number of selected features across variants.

In summary, we showed in **Article II** that traditional multi-trait GWAS methods such as CCA can be described as reverse genotype prediction methods and that machine learning based genotype prediction models are a promising alternative to existing

linear methods. Moreover, feature coefficients of machine learning models correlated with the strength of association between variants and individual traits and were predictive of true trans-eQTL target genes.

However, one aspect of multi-trait GWAS that was not considered in the study was the statistical inference. This is straightforward for the linear methods, where the null distribution of the model fit score under the assumption of no association can be approximated analytically to obtain a p-value for the significance of any observed score. However, obtaining a p-value for the significance of any observed score for non-linear methods such as random forest requires a large number of permutation tests for each variant separately. This is computationally infeasible when a large number of variants need to be studied. Therefore, a possible solution can be to use the approximate methods such as [88]. However, we did not observe a relationship between model fit and strength or extent of true biological relations. Therefore, the relevance of performing statistical inference on this test statistic, at least for trans-eQTL identification, remains to be clarified.

Based on the results in **Article II**, we decided to explore the use of random forest regression for the prediction of genotypes from another type of high-dimensional phenotypes, namely brain imaging features. Moreover, to the best of our knowledge, a genome-wide analysis of machine learning methods for reverse genotype prediction in human GWAS has not yet been conducted. Therefore, in this study, we predicted genotypes of 518,485 SNPs spanning the whole human genome (selected after the quality control procedure) from 56 brain imaging quantitative traits using data from the ADNI database.

We observed that lasso and random forest regression, but not ridge regression, identified a SNP in the APOE gene as the best performing variant. APOE genotype is the most well known genetic risk factor for Alzheimer disease. Moreover, when compared across the whole genome, random forests produced a distinct list of selected SNPs, based on RMSE prediction performance, than the linear methods (ridge and lasso regression), which were highly similar to each other. Further literature search and existing GWAS data showed that the top SNPs identified by random forests are all located in or near genes that have been previously associated with brain-related disorders. This supports our argument of using non-linear multi-variate GWAS methods for the identification of genetic variants distinct from those selected by conventional linear methods. Extending the analysis to the top 1000 SNPs predicted by random forests, our results showed clustering of image features. This shows that a group of variants not colocated in the genome tend to associate with similar brain regions or features. A more limited analysis on 876 SNPs showed that permutation p-values

and random forest regression RMSE values, but not classification accuracies, showed a high degree of correlation. This is important because the null distribution of the test statistic (RMSE) is unknown, and the only way to quantify statistical significance is by computing permutation p-values. However, this is computationally infeasible across the whole genome. Therefore, a possible solution could be to learn a model to predict p-values from RMSE values from a suitable set of training SNPs; and use these to obtain approximate permutation p-values genome-wide.

A major challenge in this study was the lack of ground truth data. Our results showed that groups of variants not colocated on the genome tend to associate with similar brain regions or features. However, even-though reverse genotype prediction correctly picks up these correlations between the phenotypic traits, the corresponding correlations and shared effects between SNPs are currently ignored since reverse genotype prediction approaches tend to learn prediction models for each SNP individually. Thus, a logical extension of our approach would be to use multi-task regression, i.e. to predict multiple SNPs simultaneously. However, this raises important computational challenges, and it may be infeasible to predict SNPs simultaneously on a genome-wide scale. We also observed that permutation p-values and random forest regression RMSE values showed a high degree of correlation. A possible solution could therefore be to learn a model to predict p-values from RMSE values from a suitable set of training SNPs to be used to obtain approximate permutation p-values genome-wide. Another limitation of the current study is that in the presence of highly correlated traits, the feature weights obtained by different methods are not necessarily robust. It would be interesting to investigate other measures of feature importance for random forest models, beyond the default ones based on Gini importances, such as model-agnostic methods like permutation importance [89].

Chapter 6

Conclusion and future prospects

In this thesis, we observed that some genetic variants explain a high proportion of variation in genome-wide gene expression. To identify which traits are affected by these variants, we explored machine learning approaches for the genetic analysis of high-dimensional phenotypic data. Our findings showed that genotype prediction performance using different machine learning methods varied across genetic variants. This helps in identifying genomic variants that have an effect on a large number of high-dimensional phenotypic traits, such as gene expression and brain-imaging features. Random forests, in particular, performed better as a generic method that requires very little parameter tuning. Most of the present GWAS approaches in the reverse genotype prediction setting rely on linear methods. However, as discussed earlier, there is no *a-priori* biological reason behind this choice.

In this thesis, we showed that non-linear methods could also be used for the purpose of reverse genotype prediction. Furthermore, our results showed that the genetic variants identified by non-linear machine learning methods like random forest were distinct from the variants identified by linear methods. We also observed that feature weights (feature importances in case of the random forest) in machine learning models can be used to identify biologically relevant variant-trait associations.

However, comparing the relative importance of variants in these models in a GWAS-like manner using a single test statistic is still an open challenge. Moreover, in the presence of highly correlated traits, the feature weights are not necessarily robust. Therefore, an interesting future research area can be to consider other measures of feature importance for random forest models beyond the default ones based on Gini importance, such as model-agnostic methods like permutation importance [89].

Another challenge the current study poses is the computation of p-values for the sig-

nificance of each genetic variant because computing a large number of permutation tests for a large number of variants is not feasible. Therefore, another interesting future prospect is to consider approximate methods [88] to overcome this hurdle.

Finally, an important promising future prospect of our study is to explore additional non-linear machine learning methods such as neural networks, including deep neural networks for predicting genotypes using MRI recordings of the brain directly instead of using a priori extracted features [90].

Bibliography

[1] Andreas Ziegler, Inke R Konig, and Friedrich Pahlke. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an E-learning platform*. John Wiley & Sons, 2010.

[2] Robert C Elston, Jaya M Satagopan, and Shuying Sun. Genetic terminology. In *Statistical Human Genetics*, pages 1–9. Springer, 2012.

[3] Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *New England journal of medicine*, 363(2):166–176, 2010.

[4] David M Umbach and Clarice R Weinberg. The use of case-parent triads to study joint effects of genotype and exposure. *The American Journal of Human Genetics*, 66(1):251–261, 2000.

[5] Wikimedia Commons David Eccles (Gringer). A single nucleotide polymorphism is a change of a nucleotide at a single base-pair location on dna., 2014. URL `https://upload.wikimedia.org/wikipedia/commons/archive/2/2e/20210508130700%21Dna-SNP.svg`.

[6] International HapMap Consortium et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

[7] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21, 2021.

[8] Gabor T Marth, Ian Korf, Mark D Yandell, Raymond T Yeh, Zhijie Gu, Hamideh Zakeri, Nathan O Stitziel, LaDeana Hillier, Pui-Yan Kwok, and Warren R Gish. A general approach to single-nucleotide polymorphism discovery. *Nature genetics*, 23(4):452–456, 1999.

[9] Geraldine M Clarke, Carl A Anderson, Fredrik H Pettersson, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133, 2011.

[10] David Baltimore. Our genome unveiled. *Nature*, 409(6822):815–816, 2001.

[11] Richard P Horgan and Louise C Kenny. 'omic'technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, 13(3): 189–195, 2011.

[12] Matthew Avison. *Measuring gene expression*. Taylor & Francis, 2008.

[13] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[14] Andreas Von Bubnoff. Next-generation sequencing: the race is on. *Cell*, 132(5): 721–723, 2008.

[15] David B Allison, Bonnie Thiel, Pamela St Jean, Robert C Elston, Ming C Infante, and Nicholas J Schork. Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *The American Journal of Human Genetics*, 63(4):1190–1201, 1998.

[16] Manuel AR Ferreira and Shaun M Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–133, 2009.

[17] Tessel E Galesloot, Kristel Van Steen, Lambertus ALM Kiemeney, Luc L Janss, and Sita H Vermeulen. A comparison of multivariate genome-wide association methods. *PloS one*, 9(4):e95923, 2014.

[18] Alexandra C Nica and Emmanouil T Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362, 2013.

[19] Nayang Shan, Zuoheng Wang, and Lin Hou. Identification of trans-eqtls using mediation analysis with multiple mediators. *BMC bioinformatics*, 20(3):87–97, 2019.

[20] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4):e1000888, 2010.

[21] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[22] Bertrand Servin and Matthew Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, 3(7):e114, 2007.

[23] Gonçalo R Abecasis, Stacey S Cherny, William O Cookson, and Lon R Cardon. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, 30(1):97–101, 2002.

[24] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

[25] Clive J Hoggart, John C Whittaker, Maria De Iorio, and David J Balding. Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS genetics*, 4(7):e1000130, 2008.

[26] Chen-Hung Kao, Zhao-Bang Zeng, and Robert D Teasdale. Multiple interval mapping for quantitative trait loci. *Genetics*, 152(3):1203–1216, 1999.

[27] Sang Hong Lee, Julius HJ Van Der Werf, Ben J Hayes, Michael E Goddard, and Peter M Visscher. Predicting unobserved phenotypes for complex traits from whole-genome snp data. *PLoS genetics*, 4(10):e1000231, 2008.

[28] Zhao-Bang Zeng. Precision mapping of quantitative trait loci. *Genetics*, 136(4): 1457–1468, 1994.

[29] James H Degnan, Jessica Lasky-Su, Benjamin A Raby, Mousheng Xu, Cliona Molony, Eric E Schadt, and Christoph Lange. Genomics and genome-wide association studies: an integrative approach to expression qtl mapping. *Genomics*, 92 (3):129–133, 2008.

[30] Anatole Ghazalpour, Sudheer Doss, Hyun Kang, Charles Farber, Ping-Zi Wen, Alec Brozell, Ruth Castellanos, Eleazar Eskin, Desmond J Smith, Thomas A Drake, et al. High-resolution mapping of gene expression using association in an outbred mouse stock. *PLoS genetics*, 4(8):e1000149, 2008.

[31] Jennifer Listgarten, Carl Kadie, Eric E Schadt, and David Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38):16465–16470, 2010.

[32] Jianlong Qi, Hassan Foroughi Asl, Johan Björkegren, and Tom Michoel. krux: matrix-based non-parametric eqtl discovery. *BMC bioinformatics*, 15(1):1–7, 2014.

[33] Vasyl Zhabotynsky, Licai Huang, Paul Little, Yi-Juan Hu, Fernando Pardo-Manuel de Villena, Fei Zou, and Wei Sun. eqtl mapping using allele-specific count data is computationally feasible, powerful, and provides individual-specific estimates of genetic effects. *PLoS Genetics*, 18(3):e1010076, 2022.

[34] Judea Pearl. Causal inference. *Causality: objectives and assessment*, pages 39–58, 2010.

[35] Miguel A Hernán, Sonia Hernández-Díaz, Martha M Werler, and Allen A Mitchell. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American journal of epidemiology*, 155 (2):176–184, 2002.

[36] Eric S Lander and Nicholas J Schork. Genetic dissection of complex traits. *Science*, 265(5181):2037–2048, 1994.

[37] Jae Hoon Sul, Lana S Martin, and Eleazar Eskin. Population structure in genetic studies: Confounding factors and mixed models. *PLoS genetics*, 14(12):e1007309, 2018.

[38] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

[39] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[40] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[41] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. Improved linear mixed models for genome-wide association studies. *Nature methods*, 9(6):525–526, 2012.

[42] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208, 2006.

[43] William Astle and David J Balding. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, 2009.

[44] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835, 2011.

[45] Charles E McCulloch and Shayle R Searle. *Generalized, linear, and mixed models*. John Wiley & Sons, 2004.

[46] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[47] Paul F O'Reilly, Clive J Hoggart, Yotsawat Pomyen, Federico CF Calboli, Paul Elliott, Marjo-Riitta Jarvelin, and Lachlan JM Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS one*, 7(5):e34861, 2012.

[48] Saikat Banerjee, Franco L Simonetti, Kira E Detrois, Anubhav Kaphle, Raktim Mitra, Rahul Nagial, and Johannes Soeding. Reverse regression increases power for detecting trans-eqtls. *bioRxiv*, 2020.

[49] Tom Mitchell. Machine learning, 1997.

[50] Fei Tang and Hemant Ishwaran. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.

[51] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[52] Li Shen, Sungeun Kim, Shannon L Risacher, Kwangsik Nho, Shanker Swaminathan, John D West, Tatiana Foroud, Nathan Pankratz, Jason H Moore, Chantel D Sloan, et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *Neuroimage*, 53(3):1051–1063, 2010.

[53] Ganesh Chauhan, Hieab HH Adams, Joshua C Bis, Galit Weinstein, Lei Yu, Anna Maria Töglhofer, Albert Vernon Smith, Sven J Van Der Lee, Rebecca F Gottesman, Russell Thomson, et al. Association of Alzheimer's disease GWAS loci with MRI markers of brain aging. *Neurobiology of aging*, 36(4):1765–e7, 2015.

[54] Xuan Bi, Liuqing Yang, Tengfei Li, Baisong Wang, Hongtu Zhu, and Heping Zhang. Genome-wide mediation analysis of psychiatric and cognitive traits through imaging phenotypes. *Human brain mapping*, 38(8):4088–4097, 2017.

[55] Zhao-Hua Lu, Zakaria Khondker, Joseph G Ibrahim, Yue Wang, Hongtu Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies. *NeuroImage*, 149:305–322, 2017.

[56] Guiyou Liu, Lifen Yao, Jiafeng Liu, Yongshuai Jiang, Guoda Ma, Zugen Chen, Bin Zhao, Keshen Li, et al. Cardiovascular disease contributes to Alzheimer's disease: evidence from large-scale genome-wide association studies. *Neurobiology of aging*, 35(4):786–792, 2014.

[57] Steven G Potkin, Jessica A Turner, Guia Guffanti, Anita Lakatos, Federica Torri, David B Keator, and Fabio Macciardi. Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations. *Cognitive neuropsychiatry*, 14(4-5):391–418, 2009.

[58] Hua Wang, Feiping Nie, Heng Huang, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics*, 28(2):229–237, 2012.

[59] Hua Wang, Feiping Nie, Heng Huang, Jingwen Yan, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps. *Bioinformatics*, 28(18):i619–i625, 2012.

[60] Meiyan Huang, Thomas Nichols, Chao Huang, Yang Yu, Zhaohua Lu, Rebecca C Knickmeyer, Qianjin Feng, Hongtu Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *Neuroimage*, 118:613–627, 2015.

[61] Meiyan Huang, Chunyan Deng, Yuwei Yu, Tao Lian, Wei Yang, Qianjin Feng, Alzheimer's Disease Neuroimaging Initiative, et al. Spatial correlations exploitation based on nonlocal voxel-wise GWAS for biomarker detection of ad. *NeuroImage: Clinical*, 21:101642, 2019.

[62] Tao Zhou, Kim-Han Thung, Mingxia Liu, and Dinggang Shen. Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model. *IEEE Transactions on Biomedical Engineering*, 66(1):165–175, 2018.

[63] Ahmad R Hariri, Emily M Drabant, and Daniel R Weinberger. Imaging genetics: perspectives from studies of genetically driven variation in serotonin function and corticolimbic affective processing. *Biological psychiatry*, 59(10):888–897, 2006.

[64] Caroline C Brun, Natasha Leporé, Xavier Pennec, Agatha D Lee, Marina Barysheva, Sarah K Madsen, Christina Avedissian, Yi-Yu Chou, Greig I De Zubicaray, Katie L McMahon, et al. Mapping the regional influence of genetics on brain structure variability—a tensor-based morphometry study. *Neuroimage*, 48(1):37–49, 2009.

[65] Li Shen, Andrew J Saykin, Moo K Chung, and Heng Huang. Morphometric analysis of hippocampal shape in mild cognitive impairment: An imaging genetics study. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, pages 211–217. IEEE, 2007.

[66] Steven G Potkin, Guia Guffanti, Anita Lakatos, Jessica A Turner, Frithjof Kruggel, James H Fallon, Andrew J Saykin, Alessandro Orro, Sara Lupoli, Erika Salvi, et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PloS one*, 4 (8):e6501, 2009.

[67] Sergio E Baranzini, Joanne Wang, Rachel A Gibson, Nicholas Galwey, Yvonne Naegelin, Frederik Barkhof, Ernst-Wilhelm Radue, Raija LP Lindberg, Bernard MG Uitdehaag, Michael R Johnson, et al. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Human molecular genetics*, 18(4):767–778, 2009.

[68] Youngsang Cho, Joon-Kyung Seong, Yong Jeong, Sung Yong Shin, Alzheimer's Disease Neuroimaging Initiative, et al. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. *Neuroimage*, 59(3):2217–2230, 2012.

[69] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. Clinical diagnosis of Alzheimer's disease. 34(7): 939–939, 1984. doi: 10.1212/WNL.34.7.939.

[70] Alzheimer's Association et al. 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3):367–429, 2018.

[71] Tapan Khan. *Biomarkers in Alzheimer's disease*. Academic Press, 2016.

[72] Wikimedia Commons Garrondo. Diagram showing changes of the brain caused by alzheimer's disease, 2008. URL `https://upload.wikimedia.org/wikipedia/commons/a/a5/Alzheimer%27s_disease_brain_comparison.jpg`.

[73] Wouter Van Rheenen, Wouter J Peyrot, Andrew J Schork, S Hong Lee, and Naomi R Wray. Genetic correlations of polygenic disease traits: from theory to practice. *Nature Reviews Genetics*, 20(10):567–581, 2019.

[74] Jianlong Qi, Hassan Foroughi Asl, Johan Björkegren, and Tom Michoel. krux: matrix-based non-parametric eqtl discovery. *BMC bioinformatics*, 15(1):1–7, 2014.

[75] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T Dermitzakis, and Olivier Delaneau. Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2016.

[76] Frank Wolfgang Albert, Joshua S Bloom, Jake Siegel, Laura Day, and Leonid Kruglyak. Genetics of trans-regulatory variation in gene expression. *Elife*, 7: e35471, 2018.

[77] Pedro T Monteiro, Jorge Oliveira, Pedro Pais, Miguel Antunes, Margarida Palma, Mafalda Cavalheiro, Mónica Galocha, Cláudia P Godinho, Luís C Martins, Nuno Bourbon, et al. Yeastract+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic acids research*, 48(D1):D642–D649, 2020.

[78] Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, et al. Ensembl 2020. *Nucleic acids research*, 48(D1):D682–D688, 2020.

[79] Andrea Pinna, Nicola Soranzo, Ina Hoeschele, and Alberto de la Fuente. Simulating systems genetics data with sysgensim. *Bioinformatics*, 27(17):2459–2462, 2011.

[80] Bruce Fischl. FreeSurfer. *Neuroimage*, 62(2):774–781, 2012.

[81] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825): 357–362, 2020.

[82] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[83] Jeff Reback, Wes McKinney, Joris Van Den Bossche, Tom Augspurger, Phillip Cloud, Simon Hawkins, Adam Klein, Matthew Roeschke, Jeff Tratner, Chang She, et al. pandas-dev/pandas: Pandas 1.1. 1. *Zenodo*, 2020.

[84] Michael L Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[85] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Austin, TX, 2010.

[86] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[87] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[88] Theo A Knijnenburg, Lodewyk FA Wessels, Marcel JT Reinders, and Ilya Shmulevich. Fewer permutations, more accurate p-values. *Bioinformatics*, 25(12):i161–i168, 2009.

[89] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10): 1340–1347, 2010.

[90] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

Chapter 7

Scientific results

Article I

G3, 2022, 12(2), jkab410
https://doi.org/10.1093/g3journal/jkab410
Advance Access Publication Date: 1 December 2021
Investigation

Restricted maximum-likelihood method for learning latent variance components in gene expression data with known and unknown confounders

Muhammad Ammar Malik and Tom Michoel *

Computational Biology Unit, Department of Informatics, University of Bergen, Bergen 5020, Norway

*Corresponding author: tom.michoel@uib.no

Abstract

Random effects models are popular statistical models for detecting and correcting spurious sample correlations due to hidden confounders in genome-wide gene expression data. In applications where some confounding factors are known, estimating simultaneously the contribution of known and latent variance components in random effects models is a challenge that has so far relied on numerical gradient-based optimizers to maximize the likelihood function. This is unsatisfactory because the resulting solution is poorly characterized and the efficiency of the method may be suboptimal. Here, we prove analytically that maximum-likelihood latent variables can always be chosen orthogonal to the known confounding factors, in other words, that maximum-likelihood latent variables explain sample covariances not already explained by known factors. Based on this result, we propose a restricted maximum-likelihood (REML) method that estimates the latent variables by maximizing the likelihood on the restricted subspace orthogonal to the known confounding factors and show that this reduces to probabilistic principal component analysis on that subspace. The method then estimates the variance–covariance parameters by maximizing the remaining terms in the likelihood function given the latent variables, using a newly derived analytic solution for this problem. Compared to gradient-based optimizers, our method attains greater or equal likelihood values, can be computed using standard matrix operations, results in latent factors that do not overlap with any known factors, and has a runtime reduced by several orders of magnitude. Hence, the REML method facilitates the application of random effects modeling strategies for learning latent variance components to much larger gene expression datasets than possible with current methods.

Keywords: gene expression; random effects model; latent factors; eQTLs

Introduction

Following the success of genome-wide association studies (GWAS) in mapping the genetic architecture of complex traits and diseases in human and model organisms (Mackay *et al.* 2009; Hindorff *et al.* 2009; Manolio 2013), there is now a great interest in complementing these studies with molecular data to understand how genetic variation affects epigenetic and gene expression states (Albert and Kruglyak 2015; Franzén *et al.* 2016; GTEx Consortium 2017). In GWAS, it is well-known that population structure or cryptic relatedness among individuals may lead to confounding that can alter significantly the outcome of the study (Astle and Balding 2009). When dealing with molecular data, this is further exacerbated by the often unknown technical or environmental influences on the data generating process. This problem is not confined to population-based studies—in single-cell analyses of gene expression, hidden subpopulations of cells and an even greater technical variability cause significant expression heterogeneity that needs to be accounted for (Buettner *et al.* 2015).

In GWAS, linear mixed models have been hugely successful in dealing with confounding due to population structure (Yu *et al.* 2006; Astle and Balding 2009; Kang *et al.* 2010; Lippert *et al.* 2011; Zhou and Stephens 2012). In these models, it is assumed that an individual's trait value is a linear function of fixed and random effects, where the random effects are normally distributed with a covariance matrix determined by the genetic similarities between individuals, hence accounting for confounding in the trait data. Random effect models have also become popular in the correction for hidden confounders in gene expression data (Kang *et al.* 2008; Listgarten *et al.* 2010; Fusi *et al.* 2012), generally outperforming approaches based on principal component analysis (PCA), the singular value decomposition (SVD), or other hidden factor models (Leek and Storey 2007; Stegle *et al.* 2010, 2012). In this context, estimating the latent factors and the sample-to-sample correlations they induce on the observed high-dimensional data is the critical problem to solve.

If it is assumed that the observed correlations between samples are entirely due to latent factors, it can be shown that the resulting random effects model is equivalent to probabilistic PCA, which can be solved analytically in terms of the dominant eigenvectors of the sample covariance matrix (Tipping and Bishop 1999; Lawrence 2005). However, in most applications, some

Received: September 07, 2021. **Accepted:** November 11, 2021

confounding factors are known in advance (*e.g.*, batch effects, genetic factors in population-based studies, or cell-cycle stage in single-cell studies), and the challenge is to estimate simultaneously the contribution of the known as well as the latent factors. This has so far relied on the use of numerical gradient-based quasi-Newton optimizers to maximize the likelihood function (Fusi *et al.* 2012; Buettner *et al.* 2015). This is unsatisfactory because the resulting solution is poorly characterized, the relation between the known and latent factors is obscured, and due to the high-dimensionality of the problem, "limited memory" optimizers have to be employed whose theoretical convergence guarantees are somewhat weak (Liu and Nocedal 1989; Lin *et al.* 2017).

Intuitively, latent variables should explain sample covariances not already explained by known confounding factors. Here, we demonstrate analytically that this intuition is correct: latent variables can always be chosen orthogonal to the known factors without reducing the likelihood or variance explained by the model. Based on this result, we propose a method that is conceptually analogous to estimating fixed and random effects in linear mixed models using the restricted maximum-likelihood (REML) method, where the variance parameters of the random effects are estimated on the restricted subspace orthogonal to the maximum-likelihood estimates of the fixed effects (Gumedze and Dunne 2011). Our method, called LVREML, similarly estimates the latent variables by maximizing the likelihood on the restricted subspace orthogonal to the known factors, and we show that this reduces to probabilistic PCA on that subspace. It then estimates the variance–covariance parameters by maximizing the remaining terms in the likelihood function given the latent variables, using a newly derived analytic solution for this problem. Similarly to the REML method for conventional linear mixed models, the LVREML solution is not guaranteed to maximize the total likelihood function. However, we prove analytically that for any given number p of latent variables, the LVREML solution attains minimal unexplained variance among all possible choices of p latent variables, arguably a more intuitive and easier to understand criterion.

The inference of latent variables that explain observed sample covariances in gene expression data is usually pursued for two reasons. First, the latent variables, together with the known confounders, are used to construct a sample-to-sample covariance matrix that is used for the downstream estimation of variance parameters for individual genes and improved identification of trans-eQTL associations (Fusi *et al.* 2012; Stegle *et al.* 2012). Second, the latent variables are used directly as "endophenotypes" that are given a biological interpretation and whose genetic architecture is of standalone interest (Parts *et al.* 2011; Stegle *et al.* 2012). This study contributes to both objectives. First, we show that the covariance matrix inferred by LVREML is *identical* to the one inferred by gradient-based optimizers, while computational runtime is reduced by orders of magnitude (*e.g.*, a 10^4-fold speed-up on gene expression data from 600 samples). Second, latent variables inferred by LVREML by design do not overlap with already known covariates and thus represent *new* aggregate expression phenotypes of potential interest. In contrast, we show that existing methods infer latent variables that overlap significantly with the known covariates (cosine similarities of up to 30%) and thus represent partially redundant expression phenotypes.

Materials and methods

Mathematical methods

All model equations, mathematical results, and detailed proofs are described in a separate Supplementary material document.

Data

We used publicly available genotype and RNA sequencing data from 1012 segregants from a cross between two yeast strains (Albert *et al.* 2018), consisting of gene expression levels for 5720 genes and (binary) genotype values for 42,052 SNPs. Following Albert *et al.* (2018), we removed batch and optical density effects from the expression data using categorical regression. The expression residuals were centered such that each sample had mean zero to form the input matrix $\mathbf{Y}$ to the model (cf. Supplementary Section S2). L2-normalized genotype PCs were computed using the SVD of the genotype data matrix with centered (mean zero) samples and used to form input matrices $\mathbf{Z}$ to the model (cf. Supplementary Section S2). Data preprocessing scripts are available at https://github.com/michoel-lab/lvreml.

LVREML analyses

The LVREML software, as well as a script that details the LVREML analyses of the yeast data, is available at https://github.com/michoel-lab/lvreml.

PANAMA analyses

We obtained the PANAMA software from the LIMIX package available at https://github.com/limix/limix-legacy.

The following settings were used to ensure that exactly the same normalized data were used by both methods: (1) For parameter **Y**, the same gene expression matrix, with each sample normalized to have zero mean, was used as input for LVREML, setting the **standardize** parameter to false. (2) The parameter **Ks** requires a list of covariance matrices for each known factor. Therefore, for each column z_i of the matrix $\mathbf{Z}$ used by LVREML, we generated a covariance matrices $\mathbf{Ks}_i = z_i z_i^T$. The **use Kpop** parameter, which is used to supply a population structure covariance matrix to PANAMA in addition to the known covariates, was set to false.

To be able to calculate the log-likelihoods and extract other relevant information from the PANAMA results, we made the following modifications to the PANAMA code: (1) The covariance matrices returned by PANAMA are by default normalized by dividing the elements of the matrix by the mean of its diagonal elements. To make these covariance matrices comparable to LVREML, this normalization was omitted by commenting out the lines in the original PANAMA code where this normalization was being performed. (2) PANAMA does not return the variance explained by the known confounders unless the **use Kpop** parameter is set to true. Therefore, the code was modified so that it would still return the variance explained by the known confounders. (3) The $\mathbf{K}$ matrix returned by PANAMA does not include the effect of the noise parameter σ^2. Therefore, the code was modified to return the $\sigma^2 1$ matrix, which was then added to the returned $\mathbf{K}$, *i.e.*, $\mathbf{K}_{new} = \mathbf{K} + \sigma^2 1$, to be able to use eq. (2) to compute the log-likelihood. The modified code is available as a fork of the LIMIX package at https://github.com/michoel-lab/limix-legacy

Results

REML solution for a random effects model with known and latent variance components

Our model to infer latent variance components in a gene expression data matrix is the same model that was popularized in the PANAMA software (Fusi *et al.* 2012) and scLVM software (Buettner *et al.* 2015), where a linear relationship is assumed between expression levels and the known and latent factors, with random

noise added (Supplementary Section S2). In matrix notation, the model can be written as

$$\mathbf{Y} = \mathbf{ZV} + \mathbf{XW} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times m}$ is a matrix of gene expression data for m genes in n samples, and $\mathbf{Z} \in \mathbb{R}^{n\times d}$ $\mathbf{X} \in \mathbb{R}^{n\times p}$ are matrices of values for d known and p latent confounders in the same n samples. The columns v_i and w_i of the random matrices $\mathbf{V} \in \mathbb{R}^{d\times m}$ and $\mathbf{W} \in \mathbb{R}^{p\times m}$ are the effects of the known and latent confounders, respectively, on the expression level of gene i and are assumed to be jointly normally distributed:

$$p([\begin{smallmatrix} v_i \\ w_i \end{smallmatrix}]) = \mathcal{N}(0, [\begin{smallmatrix} \mathbf{B} & \mathbf{D} \\ \mathbf{D}^T & \mathbf{A} \end{smallmatrix}])$$

where $\mathbf{B} \in \mathbb{R}^{d\times d}$, $\mathbf{A} \in \mathbb{R}^{p\times p}$, and $\mathbf{D} \in \mathbb{R}^{d\times p}$ are the covariances of the known–known, latent–latent, and known–latent confounder effects, respectively. Lastly, $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times m}$ is a matrix of independent samples of a Gaussian distribution with mean zero and variance σ^2, independent of the confounding effects.

Previously, this model was considered with independent random effects ($\mathbf{B}$ and $\mathbf{A}$ diagonal and $\mathbf{D} = 0$; Fusi *et al.* 2012; Buettner *et al.* 2015). As presented here, the model is more general and accounts for possible lack of independence between the effects of the known covariates. Furthermore, allowing the effects of the known and latent factors to be dependent ($\mathbf{D} \neq 0$) is precisely what will allow the latent variables to be orthogonal to the known confounders (Supplementary Section S6). An equivalent model with $\mathbf{D} = 0$ can be considered but requires nonorthogonal latent variables to explain part of the sample covariance matrix, resulting in a mathematically less tractable framework. Finally, it remains the case that we can always choose $\mathbf{A}$ to be diagonal, because the latent factors have an inherent rotational symmetry that allows any non-diagonal model to be converted to an equivalent diagonal model (Supplementary Section S5). By definition, the known covariates correspond to measured or "natural" variables, and hence, they have no such rotational symmetry.

Using standard mixed-model calculations to integrate out the random effects (Supplementary Section S2), the log-likelihood of the unknown model parameters given the observed data can be written as

$$\mathcal{L}(\mathbf{X}, \mathbf{A}, \mathbf{B}, \sigma^2 | \mathbf{Y}, \mathbf{Z}) = -\log\det(\mathbf{K}) - \mathrm{tr}(\mathbf{K}^{-1}\mathbf{C}), \tag{2}$$

where

$$\mathbf{K} = \mathbf{ZBZ}^T + Z\mathbf{DX}^T + \mathbf{XD}^T\mathbf{Z}^T + \mathbf{XAX}^T + \sigma^2 1 \tag{3}$$

and $\mathbf{C} = (\mathbf{YY}^T)/m$ is the sample covariance matrix. Maximizing the log-likelihood (2) over positive definite matrices $\mathbf{K}$ without any further constraints would result in the estimate $\hat{\mathbf{K}} = \mathbf{C}$ (note that $\mathbf{C}$ is invertible because we assume that the number of genes m is greater than the number of samples n; Anderson and Olkin 1985).

If $\mathbf{K}$ is constrained to be of the form $\mathbf{K} = \mathbf{XAX}^T + \sigma^2 1$ for a given number of latent factors $p < n$, then the model is known as probabilistic PCA and the likelihood is maximized by identifying the latent factors with the eigenvectors of $\mathbf{C}$ corresponding to the p largest eigenvalues (Tipping and Bishop 1999; Lawrence 2005). In matrix form, the probabilistic PCA solution can be written as

$$\hat{\mathbf{K}} = \mathbf{P}_1\mathbf{CP}_1 + \hat{\sigma}^2\mathbf{P}_2, \tag{4}$$

where $\mathbf{P}_1$ and $\mathbf{P}_2$ are mutually orthogonal projection matrices on the space spanned by the first p and last $n-p$ eigenvectors of $\mathbf{C}$, respectively, and the maximum-likelihood estimate $\hat{\sigma}^2$ is the average variance explained by the $n-p$ excluded dimensions (Supplementary Section S5).

If $\mathbf{K}$ is constrained to be of the form $\mathbf{K} = \mathbf{ZBZ}^T + \sigma^2 1$, the model is a standard random effects model with the same design matrix $\mathbf{Z}$ for the random effects v_i for each gene i. In general, there exists no analytic solution for the maximum-likelihood estimates of the (co)variance parameter matrix $\mathbf{B}$ in a random effects model (Gumedze and Dunne 2011). However, in the present context, it is assumed that the data for each gene are an independent sample of the *same* random effects model. Again using the fact that $\mathbf{C} = (\mathbf{YY}^T)/m$ is invertible due to the number of genes being greater than the number of samples, the maximum-likelihood solution for $\mathbf{B}$, and hence $\mathbf{K}$, can be found analytically in terms of $\mathbf{C}$ and the SVD of $\mathbf{Z}$. It turns out to be of the same form (4), except that $\mathbf{P}_1$ now projects onto the subspace spanned by the known covariates (the columns of $\mathbf{Z}$; Supplementary Section S4).

In the most general case where $\mathbf{K}$ takes the form (3), we show first that every model of the form (1) can be rewritten as a model of the same form where the hidden factors are orthogonal to the known covariates, $\mathbf{X}^T\mathbf{Z} = 0$. The reason is that any overlap between the hidden and known covariates can be absorbed in the random effects v_i by a linear transformation and, therefore, simply consists of a reparameterization of the covariance matrices $\mathbf{B}$ and $\mathbf{D}$ (Supplementary Section S6). Once this orthogonality is taken into account, the log-likelihood (2) decomposes as a sum $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, where $\mathcal{L}_2$ is identical to the log-likelihood of probabilistic PCA on the *reduced space* that is the orthogonal complement to the subspace spanned by the known covariates (columns of $\mathbf{Z}$). Analogous to the REML method for ordinary linear mixed models, where variance parameters of the random effects are estimated in the subspace orthogonal to the maximum-likelihood estimates of the fixed effects (Patterson and Thompson 1971; Gumedze and Dunne 2011), we estimate the latent variables $\mathbf{X}$ by maximizing only the likelihood term $\mathcal{L}_2$ corresponding to the subspace where these $\mathbf{X}$ live (Supplementary Section S6). Once the REML estimates $\hat{\mathbf{X}}$ are determined, they become "known" covariates, allowing the covariance parameter matrices to be determined by maximizing the remaining terms $\mathcal{L}_1$ in the likelihood function using the analytic solution for a model with known covariates $(\mathbf{Z}, \hat{\mathbf{X}})$ (Supplementary Section S6).

By analogy with the REML method, we call our method the REML method for solving the latent variable model (1), abbreviated "LVREML". While the LVREML solution is not guaranteed to be the absolute maximizer of the total likelihood function, it is guaranteed analytically that for any given number p of latent variables, the LVREML solution attains minimal unexplained variance among all possible choices of p latent variables (Supplementary Section S6).

LVREML, a flexible software package for learning latent variance components in gene expression data

We implemented the REML method for solving model (1) in a software package LVREML, available with Matlab and Python interfaces at https://github.com/michoel-lab/lvreml. LVREML takes as input a gene expression matrix $\mathbf{Y}$, a covariate matrix $\mathbf{Z}$, and a parameter ρ, with $0 < \rho < 1$. This parameter is the desired

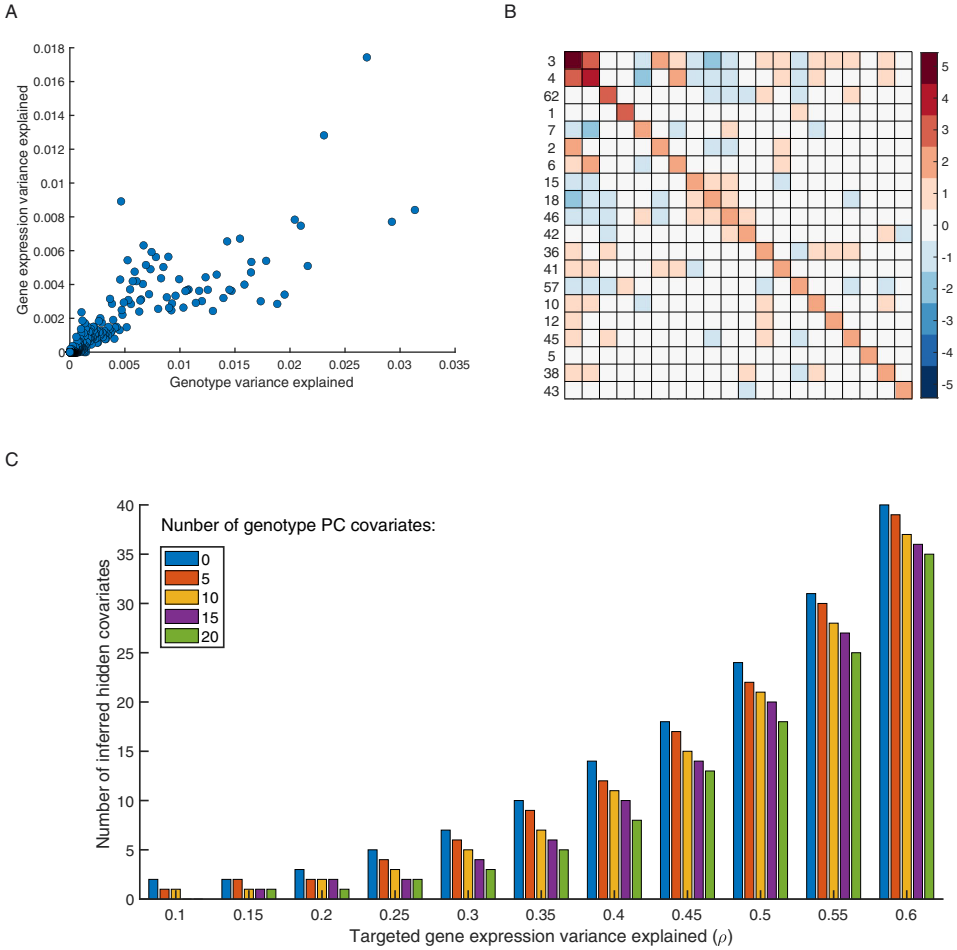


Figure 1 (A) Gene expression variance explained by individual genotype PCs in univariate models *vs* their genotype variance explained. (B) Heatmap of the estimated covariance matrix **B** [cf. (3)] among the effects on gene expression of the top 20 genotype PCs (by gene expression variance explained in univariate models, cf. A, *y*-axis); the row labels indicate the genotype PC index, ranked by genotype variance explained (cf. A, *x*-axis). (C) Number of hidden covariates inferred by LVREML as a function of the parameter ρ (the targeted total amount of variance explained by the known and hidden covariates), with θ (the minimum variance explained by a known covariate) set to retain 0, 5, 10, or 20 known covariates (genotype PCs) in the model. For visualization purposes only the range of ρ upto $\rho = 0.6$ is shown, for the full range, see Supplementary Figure S1.

proportion of variation in $\mathbf{Y}$ that should be explained by the combined known and latent variance components. Given ρ, the number of latent factors p is determined automatically (Supplementary Section S7). LVREML centers the data $\mathbf{Y}$ such that each sample has mean value zero, to ensure that no fixed effects on the mean need to be included in the model (Supplementary Section S3).

When the number of known covariates (or more precisely the rank of $\mathbf{Z}$) exceeds the number of samples, as happens in eQTL studies where a large number of SNPs can act as covariates (Fusi *et al.* 2012), a subset of n linearly independent covariates will always explain *all* of the variation in $\mathbf{Y}$. In Fusi *et al.* (2012), a heuristic approach was used to select covariates during the likelihood optimization, making it difficult to understand *a priori* which covariates will be included in the model and why. In contrast, LVREML includes a function to perform initial screening of the covariates, solving for each one the model (1) with a single known covariate to compute the variance $\hat{\beta}^2$ explained by that covariate alone (Supplementary Section S4). This estimate is then used to include in the final model only those covariates for which $\hat{\beta}^2 \geq \theta \mathrm{tr}(\mathbf{C})$, where $\theta > 0$ is the second free parameter of the method, namely the minimum amount of variation a known covariate needs to explain on its own to be included in the model (Supplementary Section S7). In the case of genetic covariates, we further propose to apply this selection criterion not to individual SNPs, but to principal components (PCs) of the genotype data matrix. Since PCA is a linear transformation of the genotype data, it does not alter model (1). Moreover, selecting PCs as covariates ensures that the selected covariates are linearly independent and are consistent with the fact that genotype PCs are known to reveal population structure in expression data (Brown *et al.* 2018).

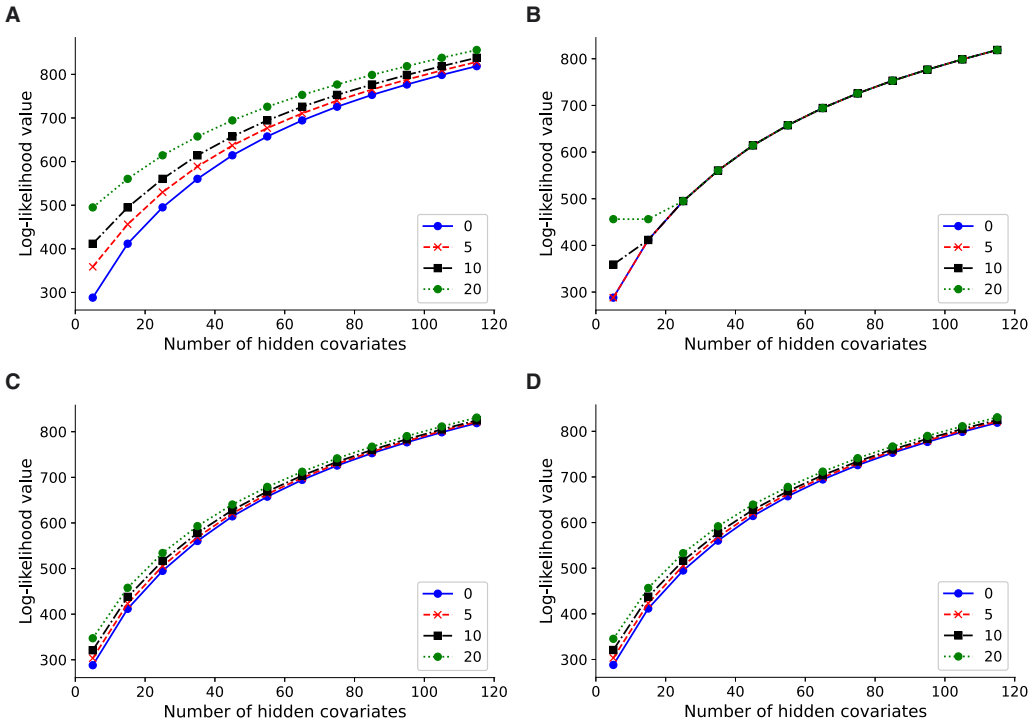


Figure 2 Log-likelihood values for LVREML (A, C) and PANAMA (B, D) using 0, 5, 10, and 20 PCs of the expression data (A, B) or genotype data (C, D) as known covariates. The results shown are for 600 randomly subsampled segregants; corresponding results for 200, 400, and in the case of LVREML 1012 segregants are shown in Supplementary Figure S2.

To test LVREML and illustrate the effect of its parameters, we used genotype data for 42,052 genetic markers and RNA sequencing expression data for 5720 genes in 1012 segregants from a cross between two strains of budding yeast (Albert *et al.* 2018), one of the largest (in terms of sample size), openly available eQTL studies in any organism (see *Materials and methods*). We first performed PCA on the genotype data. The dominant genotype PCs individually explained 2–3% of variation in the genotype data, and 1–2% of variation in the expression data, according to the single-covariate model [Supplementary Section S4, Supplementary Equation (S16), and Figure 1A]. Although genotype PCs are orthogonal by definition, their effects on gene expression are not independent, as shown by the non-zero off-diagonal entries in the maximum-likelihood estimate of the covariance matrix $\mathbf{B}$ [cf. (3); Figure 1B]. To illustrate how the number of inferred hidden covariates varies as a function of the input parameter ρ, we determined values of the parameter θ to include between 0 and 20 genotype PCs as covariates in the model. As expected, for a fixed number of known covariates, the number of hidden covariates increases with ρ, as more covariates are needed to explain more of the variation in $\mathbf{Y}$, and decreases with the number of known covariates, as fewer hidden covariates are needed when the known covariates already explain more of the variation in $\mathbf{Y}$ (Figure 1C).

When setting the parameter θ, or equivalently, deciding the number of known covariates to include in the model, care must be taken due to a mathematical property of the model: the maximizing solution exists only if the minimum amount of variation in $\mathbf{Y}$ explained by a known covariate (or more precisely, by a principal axis in the space spanned by the known covariates) is greater than the maximum-likelihood estimate of the residual variance $\hat{\sigma}^2$ (see Theorems 1 and 4 in Supplementary Sections S4 and S6). If noninformative variables are included among the known covariates, or known covariates are strongly correlated, then the minimum variation explained by them becomes small, and potentially smaller than the residual variance, whose initial "target" value is $1 - \rho$. Because LVREML considers the known covariates as fixed, it lowers the value of $\hat{\sigma}^2$ by including more hidden covariates in the model, until the existence condition is satisfied. In such cases, the total variance explained by the known and hidden covariates will be greater than the target value of the input parameter ρ. Visually, the presence of noninformative dimensions in the linear subspace spanned by the known covariates (due to noninformative or redundant variables) is shown by a saturation of the number of inferred hidden covariates with decreasing ρ (Supplementary Figure S1B), providing a clear cue that the relevance or possible redundancy of (some of) the known covariates for explaining variation in the expression data needs to be reconsidered.

LVREML attains likelihood values higher than or equal to PANAMA

To compare the analytic solution of LVREML against the original model with gradient-based optimization algorithm, as implemented in the

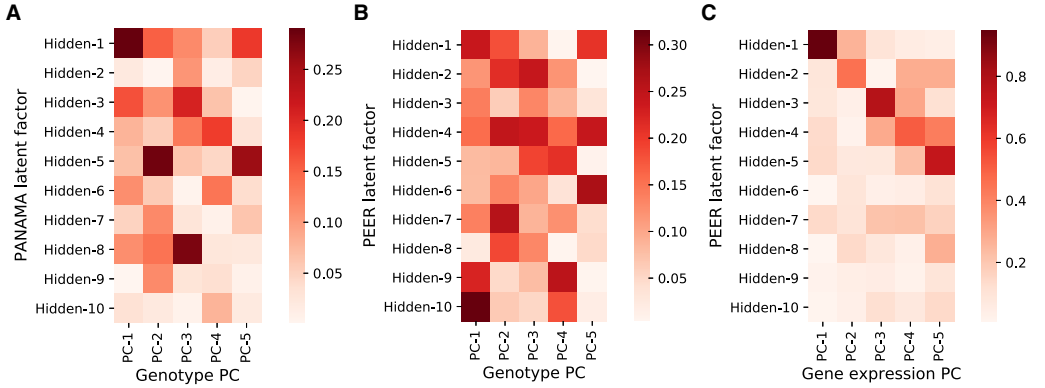


Figure 3 Cosine similarity between known covariates (five genotype PCs) given to the model and hidden factors inferred by PANAMA (A) and PEER (B), and cosine similarity between gene expression PCs and hidden factors inferred by PEER (C) when no known covariates are given to the model. Results are for randomly subsampled data of 200 segregants.

PANAMA software (Fusi *et al.* 2012), we performed a controlled comparison where 0, 5, 10, and 20 dominant PCs of the expression data $\mathbf{Y}$ were used as artificial known covariates. Because of the mathematical properties of the model and the LVREML solution, if the first d expression PCs are included as known covariates, LVREML will return the next p expression PCs as hidden factors. Hence, the log-likelihood of the LVREML solution with d expression PCs as known covariates and p hidden factors will coincide with the log-likelihood of the solution with zero known covariates and $d+p$ hidden factors (that is, probabilistic PCA with $d+p$ hidden factors). Figure 2A shows that this is the case indeed: the log-likelihood curves for 0, 5, 10, and 20 PCs as known covariates are shifted horizontally by a difference of exactly 5 (from 0, to 5, to 10) or 10 (from 10 to 20) hidden factors.

In contrast, PANAMA did not find the optimal shifted probabilistic PCA solution, and its likelihood values largely coincided with the solution with zero known covariates, irrespective of the number of known covariates provided (Figure 2B). In other words, PANAMA did not use the knowledge of the known covariates to explore the orthogonal space of axes of variation not yet explained by the known covariates, instead arriving at a solution where p hidden factors appear to explain no more of the variation than $p-d$ PCs orthogonal to the d known PCs. To verify this, we compared the PANAMA hidden factors to PCs given as known covariates, and found that in all cases where the curves in Figure 2B align, the first d hidden factors coincided indeed with the d known covariates (data not shown).

When genotype PCs were used as known confounders (using the procedure explained above), the shift in log-likelihood values was less pronounced, consistent with the notion that the genotype PCs explain less of the expression variation than the expression PCs. In this case, the likelihood values of LVREML and PANAMA coincided (Figure 2, C and D), indicating that both methods found the same optimal covariance matrix.

The explanation for the difference between Figure 2, A and C is as follows. In Figure 2A, LVREML uses p hidden covariates to explain the same amount of variation as $d+p$ expression PCs. The dominant expression PCs are partially explained by population structure (genotype data). Hence, when d genotype PCs are given as known covariates, LVREML infers p orthogonal latent variables that explain the "missing" portions of the expression PCs not explained by genotype data. This results in a model that explains more expression variation than the p dominant expression PCs, but less than $p+d$ expression PCs, hence the reduced shift in Figure 2C.

It is unclear why PANAMA did not find the correct solution when expression PCs were used as known covariates (Figure 2B), but this behavior was consistent across multiple subsampled datasets of varying sizes (Supplementary Figure S2) as well as in other datasets (data not shown).

PANAMA and PEER infer hidden factors that are partially redundant with the known covariates

Although PANAMA inferred models with the same covariance matrix estimate $\hat{\mathbf{K}}$ and hence the same likelihood values as LVREML when genotype PCs where given as known covariates, the inferred hidden covariates differed between the methods.

As explained, hidden covariates inferred by LVREML are automatically orthogonal to the known covariates and represent linearly independent axes of variation. In contrast, the latent variables inferred by PANAMA overlapped with the known genotype covariates supplied to the model, with cosine similarities of up to 30% (Figure 3A). In PANAMA, covariances among the effects of the known confounders are assumed to be zero. When the optimal model (*i.e.*, maximum-likelihood $\hat{\mathbf{K}}$) in fact has effects with non-zero covariance (as in Figure 1B), the optimization algorithm in PANAMA will automatically select hidden confounders that overlap with the known confounders to account for these non-zero covariances (Supplementary Section S6), thus resulting in the observed overlap. Hence, the common interpretation of PANAMA factors as new determinants of gene expression distinct from known genetic factors is problematic.

To test whether the overlap between inferred and already known covariates also occurs in other methods or is specific to PANAMA, we ran the PEER software (Stegle *et al.* 2012) on a reduced dataset of 200 randomly selected samples from the yeast data (PEER runtimes made it infeasible to run on larger sample sizes). PEER is a popular software that uses a more elaborate hierarchical model to infer latent variance components (Stegle *et al.* 2010). PEER hidden factors again showed cosine similarities of up to 30% (Figure 3B), suggesting that its hidden factors also cannot be interpreted as completely new determinants of gene expression. We also tested the hidden factors returned by PEER when no

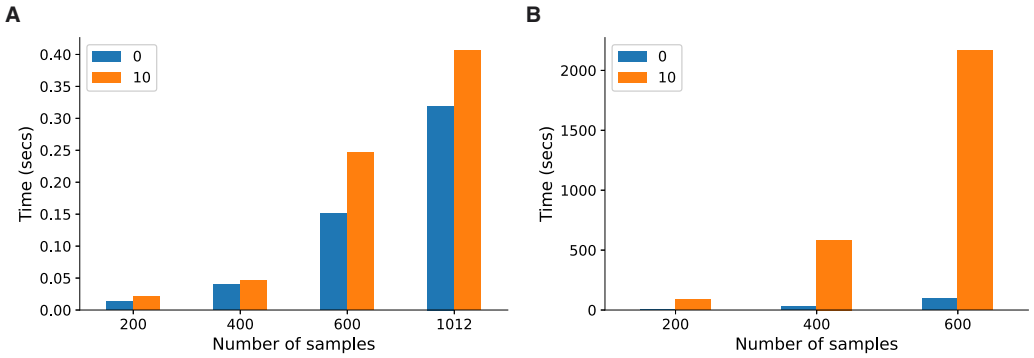


Figure 4 Runtime comparison between LVREML (A) and PANAMA (B), with parameters set to infer 85 hidden covariates with either 0 known covariates or including 10 genotype PCs as known covariates, at multiple sample sizes. Running PANAMA on the full dataset of 1012 segregants was infeasible. For runtime comparisons at other parameter settings, see Supplementary Figure S3.

known covariates are added to the model. In this case, model (1) reduces to probabilistic PCA and both LVREML and PANAMA correctly identify the dominant expression PCs as hidden factors (Figure 2, A and B). Despite its more complex model, which does not permit an analytic solution even in the absence of known covariates, PEER hidden factors in fact do overlap strongly with the same dominant expression PCs (cosine similarities between 60% and 80%), indicating that the added value of the more complicated model structure may be limited, at least in this case.

LVREML is orders of magnitude faster than PANAMA

An analytic solution does not only provide additional insight into the mathematical properties of a model but can also provide significant gains in computational efficiency. The LVREML solution can be computed using standard matrix operations from linear algebra, for which highly optimized implementations exist in all programming languages. Comparison of the runtime of the Python implementations of LVREML and PANAMA on the yeast data at multiple sample sizes showed around 10 thousand-fold speed-up factors, from several minutes for a single PANAMA run to a few tens of milliseconds for LVREML (Figure 4). Interestingly, the computational cost of LVREML did not increase much when known covariates were included in the model, compared to the model without known covariates that is solved by PCA (Figure 4A). In contrast, runtime of PANAMA blows up massively as soon as covariates are included (Figure 4B). Nevertheless, even in the case of no covariates, PANAMA is around 600 times slower than the direct, eigenvector decomposition-based solution implemented in LVREML. Finally, the runtime of LVREML does not depend on the number of known or inferred latent factors, whereas increasing either parameter in PANAMA leads to an increase in runtime (Supplementary Figure S3).

Discussion

We presented a random effects model to estimate simultaneously the contribution of known and latent variance components in gene expression data, which is closely related to models that have been used previously in this context (Lawrence 2005; Stegle *et al.* 2010, 2012; Fusi *et al.* 2012; Buettner *et al.* 2015). By including additional parameters in our model to account for non-zero covariances among the effects of known covariates and latent factors, we were able to show that latent factors can always be taken orthogonal to, and therefore linearly independent of, the known covariates supplied to the model. This is important, because inferred latent factors are not only used to correct for correlation structure in the data but also as new, data-derived "endophenotypes", that is, determinants of gene expression whose own genetic associations are biologically informative (Parts *et al.* 2011; Stegle *et al.* 2012). As shown in this paper, the existing models and their numerical optimization result in hidden factors that in fact overlap significantly with the known covariates, and hence their value in uncovering "new" determinants of gene expression must be questioned.

To solve our model, we did not rely on numerical, gradient-based optimizers, but rather on an analytic REML solution. This solution relies on a decomposition of the log-likelihood function that allows us to identify hidden factors as PCs of the expression data matrix reduced to the orthogonal complement of the subspace spanned by the known covariates. This solution is guaranteed to minimize the amount of unexplained variation in the expression data for a given number of latent factors and is analogous to the widely used REML solution for conventional linear mixed models, where variance parameters of random effects are estimated in the subspace orthogonal to the maximum-likelihood estimates of the fixed effects.

Having an analytic solution is not only important for understanding the mathematical properties of a statistical model, but can also lead to significant reduction of the computational cost for estimating parameter values. Here, we obtained a 10,000-fold speed-up compared to an existing software that uses gradient-based optimization. On a yeast dataset with 1012 samples, our method could solve the covariance structure and infer latent factors in less than half a second, whereas it was not feasible to run an existing implementation of gradient-based optimization on more than 600 samples.

The experiments on the yeast data showed that in real-world scenarios, LVREML and the gradient-based optimizer implemented in the PANAMA software resulted in the same estimates for the sample covariance matrix. Although the latent variables inferred by both methods are different (orthogonal *vs* partially overlapping with the population structure covariates), we anticipate that downstream linear association analyses will nevertheless give similar results as well. For instance, established protocols (Stegle *et al.* 2012) recommend to use known and latent factors as covariates to increase the

power to detect expression QTLs. Since orthogonal and overlapping latent factors can be transformed into each other through a linear combination with the known confounders, linear association models that use both known and latent factors as covariates will also be equivalent (Supplementary Section S8).

While we have demonstrated that the use of latent variance components that are orthogonal to known confounders leads to significant analytical and numerical advantages, we acknowledge that it follows from a mathematical symmetry of the underlying statistical model that allows us to transform a model with overlapping latent factors to an equivalent model with orthogonal factors. Whether the true but unknown underlying variance components are orthogonal or not, nor their true overlap value with the known confounders, can be established by the models studied in this paper precisely due to this mathematical symmetry. Such limitations are inherent to all latent variable methods.

To conclude, we have derived an analytic REML solution for a widely used class of random effects models for learning latent variance components in gene expression data with known and unknown confounders. Our solution can be computed in a highly efficient manner, identifies hidden factors that are orthogonal to the already known variance components, and results in the estimation of a sample covariance matrix that can be used for the downstream estimation of variance parameters for individual genes. The REML method facilitates the application of random effects modeling strategies for learning latent variance components to much larger gene expression datasets than currently possible.

Data availability

The LVREML software and all data processing and analysis scripts underlying this article are available at https://github.com/michoel-lab/lvreml.

The modified code for running the PANAMA analyses is available as a fork of the LIMIX package at https://github.com/michoel-lab/limix-legacy.

No new data were generated in support of this research.

Expression levels in units of log2(TPM) for all yeast genes and segregants were obtained from https://doi.org/10.7554/eLife.35471.021.

Information on experimental batch and growth covariates for all yeast segregants was obtained from https://doi.org/10.7554/eLife.35471.022.

Genotypes at 42,052 markers for all yeast segregants were obtained from https://doi.org/10.7554/eLife.35471.023.

Supplementary material is available at G3 online.

Funding

This research was supported in part by a grant from the Research Council of Norway (grant number 312045) to T.M.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Albert FW, Bloom JS, Siegel J, Day L, Kruglyak L. 2018. Genetics of trans-regulatory variation in gene expression. eLife. 7:e35471.

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 16:197–212.

Anderson TW, Olkin I. 1985. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. Linear Algebra Appl. 70:147–171.

Astle W, Balding DJ. 2009. Population structure and cryptic relatedness in genetic association studies. Stat Sci. 24:451–471.

Brown BC, Bray NL, Pachter L. 2018. Expression reflects population structure. PLoS Genet. 14:e1007841.

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, *et al.* 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 33:155–160.

Franzén O, Ermel R, Cohain A, Akers N, Di Narzo A, *et al.* 2016. Cardiometabolic risk loci share downstream *cis* and *trans* genes across tissues and diseases. Science. 353:827–830.

Fusi N, Stegle O, Lawrence ND. 2012. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. PLoS Comput Biol. 8:e1002330.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. Nature. 550:204.

Gumedze F, Dunne T. 2011. Parameter estimation and inference in the linear mixed model. Linear Algebra Appl. 435:1920–1944.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, *et al.* 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 106:9362–9367.

Kang HM, Sul JH, Zaitlen NA, Kong S, Freimer NB, *et al.* 2010. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 42:348–354.

Kang HM, Ye C, Eskin E. 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics. 180:1909–1925.

Lawrence N. 2005. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. J Mach Learn Res. 6:1783–1816.

Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 3:e161.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, *et al.* 2011. FaST linear mixed models for genome-wide association studies. Nat Methods. 8:833–835.

Lin H, Mairal J, Harchaoui Z. 2017. A generic quasi-Newton algorithm for faster gradient-based optimization. arXiv preprint arXiv:1610.00960 v2.

Listgarten J, Kadie C, Schadt EE, Heckerman D. 2010. Correction for hidden confounders in the genetic analysis of gene expression. Proc Natl Acad Sci U S A. 107:16465–16470.

Liu DC, Nocedal J. 1989. On the limited memory BFGS method for large scale optimization. Math Program. 45:503–528.

Mackay TF, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet. 10:565–577.

Manolio TA. 2013. Bringing genome-wide association findings into clinical use. Nat Rev Genet. 14:549–558.

Parts L, Stegle O, Winn J, Durbin R. 2011. Joint genetic analysis of gene expression data with inferred cellular phenotypes. PLoS Genet. 7:e1001276.

Patterson HD, Thompson R. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika. 58:545–554.

Stegle O, Parts L, Durbin R, Winn J. 2010. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 6:e1000770.

Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (peer) to obtain increased

power and interpretability of gene expression analyses. Nat Protoc. 7:500–507.

Tipping ME, Bishop CM. 1999. Probabilistic principal component analysis. J R Stat Soc B. 61:611–622.

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, *et al.* 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 38: 203–208.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 44:821–824.

Communicating editor: G. de los Campos

Article II

HIGH-DIMENSIONAL MULTI-TRAIT GWAS BY REVERSE PREDICTION OF GENOTYPES USING MACHINE LEARNING METHODS

A PREPRINT

Muhammad Ammar Malik
Department of Informatics
University of Bergen
PO Box 7803, 5020 Bergen, Norway
muhammad.malik@uib.no

Adriaan Ludl
Department of Informatics
University of Bergen
PO Box 7803, 5020 Bergen, Norway
adriaan.ludl@uib.no

Tom Michoel*
Department of Informatics
University of Bergen
PO Box 7803, 5020 Bergen, Norway
tom.michoel@uib.no

February 10, 2022

ABSTRACT

Motivation: Multi-trait genome-wide association studies (GWAS) use multi-variate statistical methods to identify associations between genetic variants and multiple correlated traits simultaneously, and have higher statistical power than independent univariate analyses of traits. Reverse regression, where genotypes of genetic variants are regressed on multiple traits simultaneously, has emerged as a promising approach to perform multi-trait GWAS in high-dimensional settings where the number of traits exceeds the number of samples.

Results: We analyzed different machine learning methods (ridge regression, naive Bayes/independent univariate, random forests and support vector machines) for reverse regression in multi-trait GWAS, using genotypes, gene expression data and ground-truth transcriptional regulatory networks from the DREAM5 SysGen Challenge and from a cross between two yeast strains to evaluate methods. We found that genotype prediction performance, in terms of root mean squared error (RMSE), allowed to distinguish between genomic regions with high and low transcriptional activity. Moreover, model feature coefficients correlated with the strength of association between variants and individual traits, and were predictive of true trans-eQTL target genes, with complementary findings across methods.

Availability: Code to reproduce the analysis is available at https://github.com/michoel-lab/Reverse-Pred-GWAS.

1 Background

Genome-wide association studies (GWAS) aim to find statistical associations between genetic variants and traits of interest using data from a large number of individuals [1, 2]. When multiple correlated traits are studied simultaneously, joint, multi-trait approaches can be more advantageous than studying the traits

*Corresponding author

individually, due to increased power from taking into account cross-trait covariances and reduced multiple-testing burden by performing a single test for association to a set of traits [3, 4, 5, 6].

The most commonly used multi-trait GWAS approaches are based on a multivariate analysis of variance (MANOVA) or canonical correlation analysis (CCA) [4]. However, these are applicable only to studies where the number of traits is relatively small, especially in comparison to the number of samples. When analyzing the effects of genetic variants on molecular traits (gene or protein expression levels, metabolite concentrations) or imaging features, we have to deal with a large number, often an order of magnitude or more greater than the sample size, of correlated traits simultaneously. For such studies, the standard procedure is still to conduct univariate linear regression or ANOVA tests for each genetic variant against each trait separately. While efficient algorithms exist to undertake this task [7, 8, 9], the massive multiple-testing problem results in a significant loss of statistical power.

An alternative approach to multi-trait GWAS has been to reverse the functional relation between genotypes and traits, and fit a multivariate regression model that predicts genotypes from multiple traits simultaneously, instead of the usual approach to regress traits on genotypes. The first study to do this explicitly used logistic regression and showed a significant increase in power compared to univariate methods, without being dependent on assuming normally distributed genotypes like MANOVA or CCA [10]. Although the method as presented in [10] is still only valid when the number of traits is small, extending multivariate regression methods to high-dimensional settings is straightforward. Thus a recent study used L2-regularized linear regression of single nucleotide polymorphisms (SNPs) on gene expression traits to identify trans-acting expression quantitative trait loci (trans-eQTLs), and showed that this approach aggregates evidence from many small trans-effects while being unaffected by strong expression correlations [11]. In a very different application domain, regularized regression of SNP genotypes on longitudinal image phenotypes was used to identify time-dependent genetic associations with imaging phenotypes [12].

Despite these advances, several limitations and open questions remain unanswered in high-dimensional GWAS. Firstly, linear models search for the linear combination of traits that is most strongly associated to the genetic variant, but there is no *a priori* biological reason why only linear combinations should be considered. Secondly, while L2-regularization allows to deal with high-dimensional traits, it does not address the problem of variable selection. For instance, in the case of gene expression, we expect that trans-eQTLs are potentially associated with *many*, but not *all* genes. Indeed, in [11] a secondary set of univariate tests is carried out to select genes associated to trans-eQTLs identified by the initial multi-variate regression. Thirdly, a systematic biological validation and comparison of the available methods is lacking.

Here we address these questions by considering a wider range of machine learning methods (in particular, random forests (RF) and support vector machines (SVM)) for reverse genotype prediction from gene expression traits. Hypothesizing that true trans-eQTL associations are mediated by transcription regulatory networks, we use simulated data from the DREAM5 Systems Genetics Challenge, and real data from 1,012 segregants of a cross between two budding yeast strains [13] together with the YEASTRACT database of known transcriptional interactions [14], to validate and compare these methods against univariate and L2-regularized linear regression.

2 Approach

As in other multi-trait GWAS methods, we consider one genetic variant at a time, and represent it by a random variable Y. We consider p traits represented by random variables $X_1, X_2, \ldots, X_p$ taking real values. We define a "forward" multi-trait association model probabilistically through a conditional distribution $p(X_1, \ldots, X_p \mid Y)$, which corresponds to the natural direction where variation in Y causes variation in the X_i. Using Bayes' formula, we can write the same model in the reverse causal direction using Y as the dependent variable:

$$P(Y \mid X_1, \ldots, X_p) = P(X_1, \ldots, X_p \mid Y) \frac{P(Y)}{P(X_1, \ldots, X_p)} \tag{1}$$

where $P(Y)$ and $P(X_1, \ldots, X_p)$ are prior distributions. Conversely, a forward model $P(X_1, \ldots, X_p \mid Y)$ can be obtained from a reverse model $P(Y \mid X_1, \ldots, X_p)$ using the same formula.

We have data in the form of independent random samples from the joint distribution $P(Y, X_1, \ldots, X_p)$ in n individuals, represented by a genotype vector $\mathbf{y} \in \mathbb{R}^n$ and trait vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p \in \mathbb{R}^n$, which we gather in a matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$. The log-likelihood of observing the data is the log-probability

density

$$\begin{aligned}\mathcal{L} = \log p(\mathbf{y}, \mathbf{X}) &= \log \prod_{j=1}^{n} p(y_j, x_{j1}, \ldots, x_{jp}) \\ &= \sum_{j=1}^{n} \log p(y_j, x_{j1}, \ldots, x_{jp}),\end{aligned}$$

which can be expressed in terms of the forward or reverse conditional probabilities depending on the type of model being fit. We now review how existing as well as newly proposed, and low-dimensional as well as high-dimensional multi-trait GWAS methods fit within this framework.

2.1 Univariate tests

The simplest method for multi-trait GWAS in the high-dimensional setting consists of testing each trait for association with the genetic variant independently. In this case we fit, by maximum-likelihood, a model $p(x_i \mid y)$ for each trait X_i independently using a linear model

$$p(x_i \mid y) = \mathcal{N}(\mu_y, \sigma^2)$$

a normal distribution with mean μ_y dependent on the genotype value y. This corresponds to the multi-trait model

$$p(x_1, \ldots, x_p \mid y) = \prod_{i=1}^{p} p(x_i \mid y)$$

Using Bayes' rule eq. (1), we obtain

$$\begin{aligned}P(y \mid x_1, \ldots, x_p) &= p(x_1, \ldots, x_p \mid y) \frac{P(y)}{p(x_1, \ldots, x_p)} \\ &\propto P(y) \prod_{i=1}^{p} p(x_i \mid y)\end{aligned}$$

where $P(y)$ is the prior probability (background frequency) of observing genotype class y. This is the formula for a *naive Bayes classifier* of the genotype y given features x_i. In the univariate approach, statistical tests are carried out to determine whether a genotype-dependent model $p(x_i \mid y)$ is more likely or not than a model where the trait is independent of the genotype. This is equivalent to doing a feature selection to determine which traits to include in the naive Bayes classifier.

2.2 Canonical correlation analysis

MV-PLINK [4] is a multivariate method based on Canonical Correlation Analysis (CCA). Given two sets of random variables $(X_1, X_2, \ldots, X_p)$ and
$(Y_1, Y_2, \ldots, Y_q)$, CCA finds linear coefficients $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ that maximize the correlation

$$\rho(\mathbf{a}, \mathbf{b}) = \operatorname{corr}\left(\sum_{i=1}^{p} a_i X_i, \sum_{j=1}^{q} b_j Y_j\right)$$

It can be shown (see SI Section S1) that if $q = 1$, then the maximizing coefficients $\hat{\mathbf{a}}$ are given by $\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where $\mathbf{X}$ and $\mathbf{y}$ are the data sampled from the joint distribution $P(Y, X_1, X_2, \ldots, X_p)$. These are the same coefficients that would be obtained from a *linear regression* model where Y is modelled as a linear function of the predictors $(X_1, X_2, \ldots, X_p)$, or from the maximum-likelihood solution of a reverse probabilistic model

$$p(y \mid x_1, \ldots, x_p) = \mathcal{N}\left(\sum_{i=1}^{p} a_i x_i, \sigma^2\right). \tag{2}$$

2.3 Reverse logistic regression

MultiPhen [10] is a method that is described directly in terms of a model to predict genotypes from multiple traits, using proportional odds *logistic regression*, that is, instead of fitting the genotype class probabilities $P(y = m \mid x_1, \ldots, x_p)$, for $m = 0, 1, 2$ (for biallelic data), the method fits

$$P(y \leq m \mid x_1, \ldots, x_p) = \frac{1}{1 + e^{-\alpha_m - \sum_{i=1}^{p} \beta_i x_i}}$$

Then a likelihood ratio test is used to determine if this model fits the data better than a model where $\beta_1 = \cdots = \beta_p = 0$, thus carrying out a single test for each genetic variant, testing whether the variant is associated with *any* of the traits using the logistic regression model.

2.4 L2-Regularized reverse regression

Expressing CCA for multi-trait GWAS as a linear regression of the variant genotype on the trait values [eq. (2)] immediately leads to a generalization to the high-dimensional setting in the form of regularizing the regression coefficients, that is, augmenting eq. (2) with a prior distribution $p(a_i) = \mathcal{N}(0, \alpha)$, $i = 1, \ldots, p$. Finding the maximum-likelihood values of the regression coefficients is equivalent to L_2-regularized or ridge regression. This is the approach followed by [11], who combined it with a likelihood ratio test to determine whether the fitted model is more likely than a model where the genotype is independent of the traits ($a_i = 0$ for all i) and obtain a single association p-value for each variant.

2.5 Reverse genotype prediction using machine learning methods

From the above, we conclude that existing multi-trait GWAS methods can be described as reverse genotype prediction methods. From this perspective, L_2-regularized linear regression is but one of many established machine learning methods that could be used to predict an outcome variable Y from a high number of predictors or features X_i, $i = 1, \ldots, p$. Hence we propose to consider a wider range of machine learning methods such as random forest regression (RFR) and support vector regression (SVR) [15]. Our overall hypothesis is that genetic variants whose genotypes can be predicted with higher accuracy are more likely to affect some or all of the traits under consideration than variants whose genotypes cannot be predicted well, and that feature weights in the fitted models measure the strength of biological association between a variant and a trait.

3 Methods

3.1 Reverse genotype prediction

For genotype prediction using machine learning models, the expression values were treated as explanatory variables whereas the genotype value of a variant was treated as a response variable. The prediction performance was measured by computing the root mean squared error (RMSE) between the predicted and the actual genotype value of variants.

3.2 Trans-eQTL target prediction

Trans-eQTL target prediction was done using weights assigned to the features by the machine learning methods: feature importance in case of random forest regression (RFR), and coefficients for support vector regression (SVR) and ridge regression (RR). We computed the area under the receiver operating characteristic (AUROC) curve to measure prediction performance by comparing the weights against the true targets in the ground truth for each variant.

3.3 Datasets

3.3.1 Simulated data

The simulated data for our experiments was obtained from DREAM5 Systems Genetic Challenge A (`https://www.synapse.org/#!Synapse:syn2820440/wiki/`), generated by the SysGenSIM software [16].

The DREAM data consists of simulated genotype and transcriptome data of synthetic gene regulatory networks. The dataset consists of 15 sub-datasets, where 5 different networks are provided and for each network 100, 300 and 999 samples are simulated. Every sub-dataset contains 1000 genes. We used the networks with 999 samples only.

In the DREAM data, each genetic variant is associated to a unique causal gene that mediates its effect. We therefore defined ground-truth trans-eQTL targets for each variant as the causal gene's direct targets in the ground-truth network.

In the DREAM data 25% of the variants acted in *cis*, meaning they affected expression of their causal gene directly. The remaining 75% of the variants acted in *trans*. Since the identities of the *cis* and *trans* eQTLs are unknown, we computed the P-values of genotype-gene expression associations between matching variant-gene pairs using Pearson correlation and selected all genes with P-values less than 1/750 to identify cis-acting eQTLs.

3.3.2 Yeast data

The yeast data used in this paper was obtained from [13]. The expression data contains expression values for 5,720 genes in 1,012 segregants. The genotype data consists of binary genotype values for 42,052 genetic markers in the same 1,012 segregants.

Batch and optical density (OD) effects, as given by the covariates provided in [13], were removed from the expression data using categorical regression, as implemented in the *statsmodels* python package. The expression data was then normalized to have zero mean and unit standard deviation.

To match variants to genes, we considered the list of genome-wide significant eQTLs provided by [13] whose confidence interval (of variable size) overlapped with an interval covering a gene plus 1,000 bp upstream and 200 bp downstream of the gene position. This resulted in a list of 2,884 genes and for each of these genes we defined its matching variant as the most strongly associated variant from the list.

Networks of known transciptional regulatory interactions in yeast (S. cerevisiae) were obtained from the YEASTRACT (Yeast Search for Transcriptional Regulators And Consensus Tracking) [14]. Regulation matrices were obtained from `http://www.yeastract.com/formregmatrix.php`. We retrieved the ground-truth matrix containing all reported interactions of the type *DNA binding and expression evidence*. Self regulation was removed from the ground-truths. The Ensembl database (release 83, December 2015) [17] was used to map gene names to their identifiers. After overlaying the ground-truth with the set of genes with matching cis-eQTL, a ground-truth network of 80 transcription factors (TFs) with matching cis-eQTL and 3,394 target genes was obtained.

The expression dataset was then filtered to contain only the genes present in the ground truth network, and ground-truth trans-eQTL sets for the 80 TF-associated cis-eQTL genetic variants were defined as direct targets of the corresponding TFs in the ground-truth network.

3.4 Experimental settings

In all sets of experiments we used 5-fold cross-validation, except for the feature selection experiment in yeast data where we used 80-20 train-test split due to time constraints.

Ridge Regression (RR), Random Forest Regression (RFR), Support Vector Regression (SVR), and Naive Bayes (NB) were implemented using the Python library *scikit-learn*. For RR, the regularization strength (α) was set to 100 and other parameters were set to their defaults. For RR and SVR, the default parameters were used. For NB, we used the Gaussian Naive Bayes from *scikit-learn* library. For trans-eQTL predictions, univariate linear correlation was also used to compare with the regression methods mentioned above.

3.4.1 Feature selection

For each method we took the absolute values of the feature importances/coefficients, scaled so that their sum equals to 1, and sorted these in descending order. These scaled values represent the relative contribution of each feature to the prediction of each variant. We selected the top-scoring features which together contributed 50% of the feature weight sum.

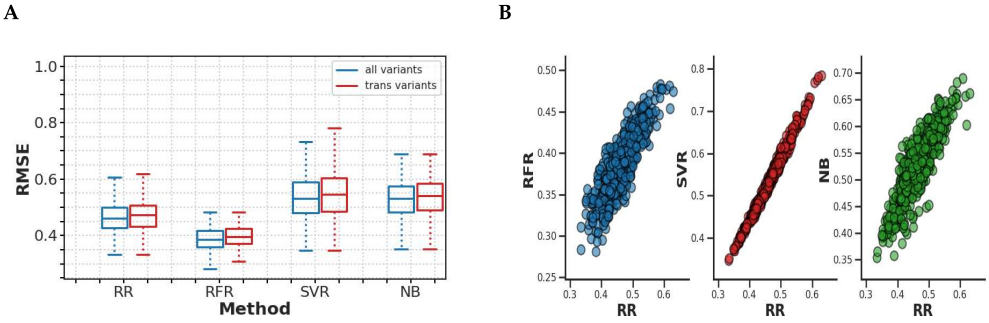


Figure 1: RMSE values for genotype prediction on DREAM5 simulated data. (**A**) Boxplots show the distribution of the RMSE values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). (**B**) Scatter plots show RMSE values of RFR, SVR, and NB vs RR for all variants. The data shown are for DREAM Network 1. The results for Network 2-5 are shown in Supp. Figs. S1-S4.

3.5 Code

The scripts to reproduce the analysis are available at
`https://github.com/michoel-lab/Reverse-Pred-GWAS`.

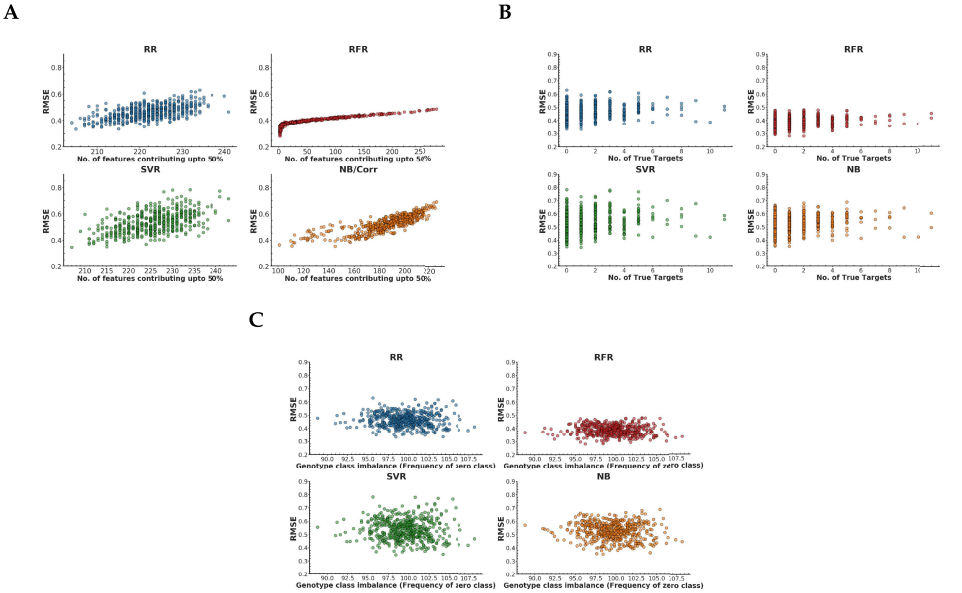


Figure 2: Scatter plots of genotype RMSE values on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). The data shown are for DREAM Network 1. The results for Network 2-5 are shown in Supp. Figs. S5-S8.

4 Results

4.1 Reverse genotype prediction and trans-eQTL analysis in simulated data

In the DREAM5 Systems Genetics Challenge, binary genotypes and steady-state gene expression data for 1,000 genes were simulated for a population of 999 individuals, based on a gene network topology and the individuals' genotypes at a set of genome-wide DNA variants, using non-linear ordinary differential equations (ODEs) [16]. In the simulations, there was a one-to-one mapping between genetic variants and genes, such that the effects of each variant are mediated by exactly one causal gene. 25% of the variants acted in *cis*, meaning they affected expression of their causal gene, but not the value of any of the parameters in the ODE model. The remaining 75% of the variants acted in *trans*, meaning they did not affect expression of their causal gene, but did affect the transcription rate of the causal gene's targets in the network. Simulated data for five networks are available.

4.1.1 Genotype prediction accuracy varies across genetic variants

We trained models to predict the genotypes for variants whose causal gene had at least one target in the ground-truth network (covering between 491-644 genes depending on the network/dataset) using the expression data from all 1,000 genes as predictors, using Random Forest Regression (RFR), Support Vector Regression (SVR), Ridge Regression (RR) and Naive Bayes classification (NB). RMSE was then measured for each predicted variant in the test data. Mean performance across the five train-test folds is reported here.

RFR achieved the best prediction performance (lowest RMSE) overall (RMSE $\sim 0.3 - 0.5$). RR achieved RMSEs in the range of $\sim 0.6 - 0.8$. In contrast to RFR and RR, the RMSE varied widely for SVR and NB ($\sim 0.3 - 0.9$) (Fig. 1A). We did not observe a significant change in the distribution of RMSE values for all the variants versus keeping only *trans*-acting variants (Fig. 1A), i.e. *cis*-acting variants are not significantly easier to predict (by virtue of having a highly correlated *cis*-gene) than variants that only have *trans*-associated genes. While RMSE values are correlated between the methods (Fig. 1B), the correlation is imperfect (with the exception of SVR-RR), such that there is considerable variation in the RMSE-based ranking of variants between the methods.

Taken together these result show that prediction performance varies across genetic variants within each method (i.e. variants can be ranked according to their RMSE) and that RFR can be preferred over the others in terms of average prediction performance, but with considerable variation in relative performance across methods for individual variants.

Next, we compared the genotype prediction performance for the different methods with the number of features contributing upto 50% of the total sum of feature weights (cf. Methods). In general, variants that were more predictable had models with fewer features, and vice versa, irrespective of the prediction method used (Fig. 2A). On the other hand, we did not observe any significant relation between the prediction performance and the number of true targets in the ground truth network (Fig. 2B). We also tested whether RMSE was influenced by the genotype class imbalance. This was not the case for the regression-based methods used here (Fig. 2C).

4.1.2 Feature importances are predictive of true trans-eQTL associations

To evaluate the ability of reverse genotype prediction methods to identify true trans-eQTL targets of a given variant, we defined true trans associations as direct target genes of a variant's causal gene in the ground-truth network and used feature importances/coefficients in the genotype prediction model to predict how likely a gene is to be a trans-eQTL of a given variant (see Methods). Performance was measured using the area under the receiver operating curve (AUROC).

For all methods, more than $\sim$ 55%, resp. $\sim$ 65% of variants with at least one trans-eQTL target in the ground-truth network had AUROC$>$ 0.8, resp. 0.7, with univariate linear correlation and ridge regression performing somewhat better than random forest and SVR (Fig. 3). Ridge regression and univariate correlation methods also had less variation in terms of AUROCs when compared with RFR and SVR, and no significant difference in terms of AUROC was observed when using all the variants versus using only *trans*-acting variants (Fig. 4A). Interestingly, the variants for which high AUROCs were obtained differed between RFR, RR and univariate correlation methods, whereas RR and SVR obtained nearly identical performance on all variants. (Fig. 4B).

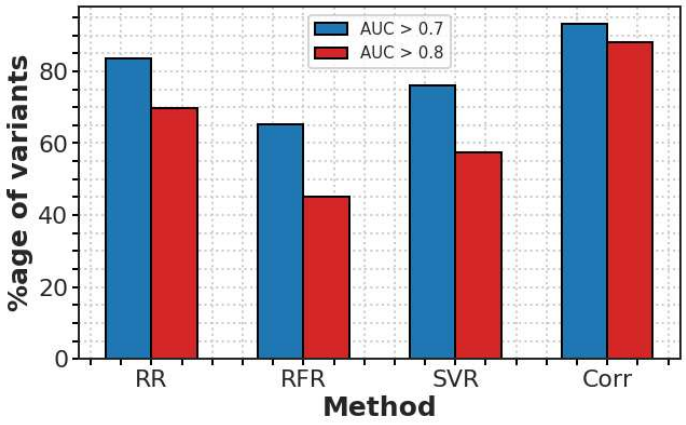


Figure 3: Bar plots show the proportion of variants with trans-eQTL target prediction AUROC > 0.7 (blue) and > 0.8 (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). The data shown are for DREAM Network 1. The results for Network 2-5 are shown in Supp. Fig. S9.

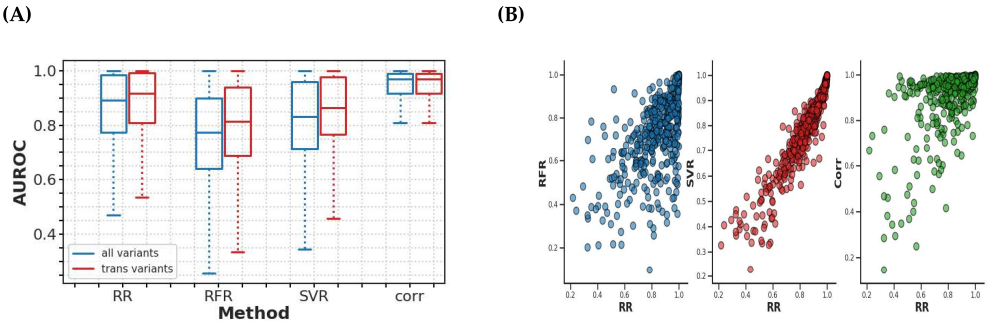


Figure 4: Trans-eQTL target prediction performance on DREAM5 simulated data. **(A)** Boxplots show the distribution of AUROC values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). **(B)** Scatter plots show AUROC values of classification methods RFR, SVR, and Corr vs RR for all variants. The data shown are for DREAM Network 1. The results for Network 2-5 are shown in Supp. Figs. S10-S13.

When compared to potential explanatory factors, no significant relation was observed between AUROC values and number of selected model features (Fig. 5A), number of ground-truth targets (Fig. 5B), or the genotype class balance (Fig. 5C).

4.1.3 Genotype and trans-eQTL prediction performance do not correlate

Finally we tested whether genotype prediction accuracy can be used as a proxy for trans-eQTL prediction accuracy, that is, in the absence of ground-truth networks, can we use genotype prediction accuracy to filter variants whose model feature weights are indicative of true trans-eQTL targets? However, we did not observe any correlation between the genotype prediction performance and trans-eQTL target prediction performance for any of the methods (Fig. 6).

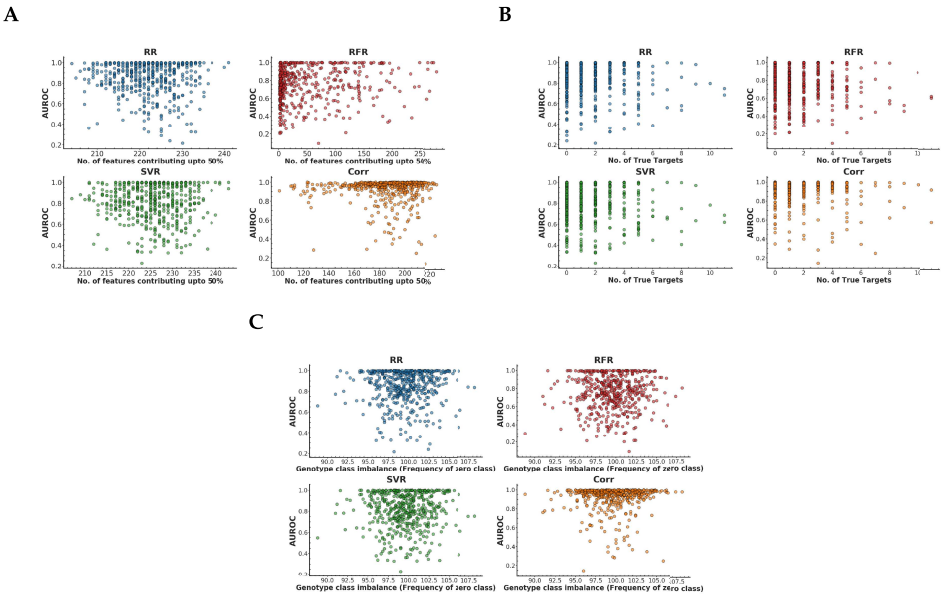


Figure 5: Scatter plots of trans-eQTL target prediction performance (AUROC) on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB). The data shown are for DREAM Network 1. The results for Network 2-5 are shown in Supp. Figs. S14-S17.

4.2 Reverse genotype prediction and trans-eQTL analysis in yeast

In the next set of experiments we repeated the same analysis on yeast dataset. Compared to the simulated data, the yeast data differs in two important aspects. First, ground-truth target information is available for a small set of transcription factors (TFs) only. Secondly, we have no knowledge of the causal gene(s) corresponding to each variant, and need to rely on a local *cis*-association between a variant and a TF to define a ground-truth set of trans-eQTL targets to a variant (cf. Methods).

4.2.1 Genotype prediction accuracy varies across genetic variants

Genotype prediction performance for the yeast data also varied across genetic variants. Similar to DREAM data RFR achieved lowest RMSE values in the yeast data as well (Fig. 7A). We tested whether prediction performance may be explained by local *cis*-associations by removing genes on the same chromosome as the test variant from the list of predictors. In this case we did observe that RMSE increased significantly (i.e. prediction performance decreased) when removing local genes, except for NB, and that after removing local genes, RR, RFR, and SVR have similar average prediction performance (Fig. 7A).

Correlations of RMSEs between methods showed a similar pattern as in the simulated data, with RR and SVR RMSEs being particularly strongly correlated (Fig. 7B).

As in the simulated data, genotype prediction performance decreased (i.e. RMSE increased) with increasing number of model features (Fig. 8A), but did not depend significantly on the number of true targets (Fig. 8B) or genotype class balance (Fig. 8C).

Next we tested whether feature importance weights were predictive of true trans-eQTL associations, defined as genes that were bound by and differentially expressed upon perturbation of a TF for which a given

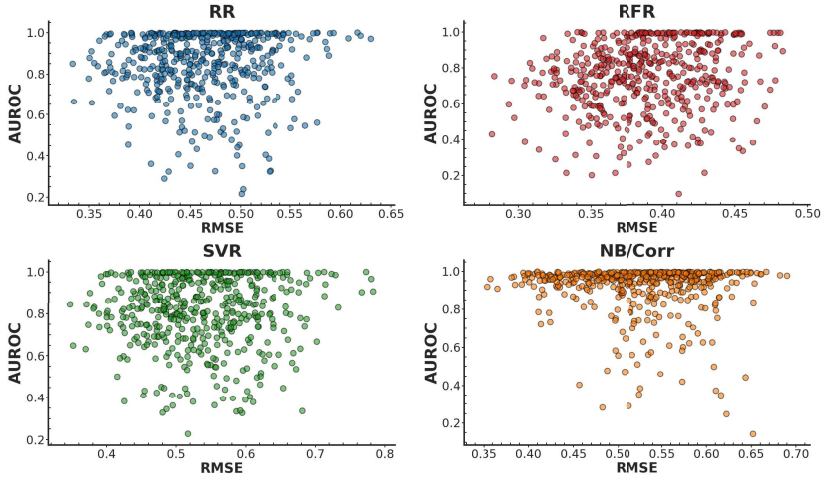


Figure 6: Scatter plots show trans-eQTL target prediction performance (AUROC) vs genotype prediction performance (RMSE) on DREAM5 simulated data for all genetic variants for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB/Corr). The data shown are for DREAM Network 1. The results for Network 2-5 are shown in Supp. Figs. S18-S21.

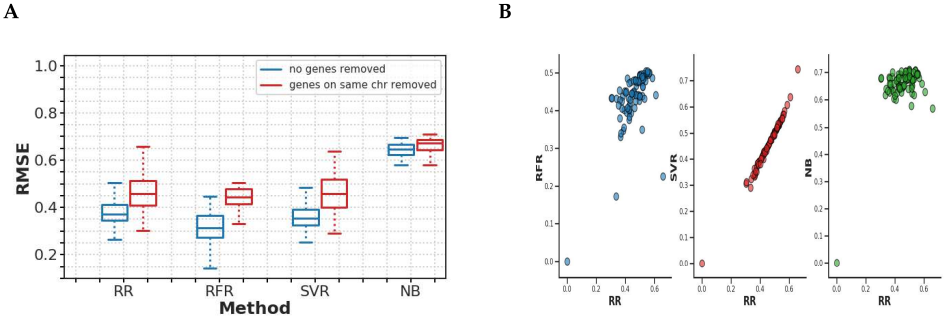


Figure 7: Genotype prediction performance on yeast data. Genotype prediction performance on yeast data. (**A**) Boxplots show the distribution of the performance for all variants using all genes (blue) and excluding genes on the same chromosome as the variant (red) as predictors, for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). (**B**) Scatter plots show RMSE values of classification methods RFR, SVR, and NB vs RR for all variants. Genes on the same chromosome were excluded as predictors for each SNP.

variant is a cis-eQTL (cf. Methods). In this case, feature importances were only modestly predictive, with 20-30%, resp. 10-15%, of TF cis-eQTLs obtaining AUROCs > 0.6, resp. > 0.7, and, as in the simulated data, there were fewer variants with high AUROC for RFR, compared to the other methods (Fig. 9).

We confirmed that the distribution of AUROC values was not affected by removing genes on the same chromosome as a variant of interest from the list of predictors (Fig. 10A). Furthermore, the AUROC values

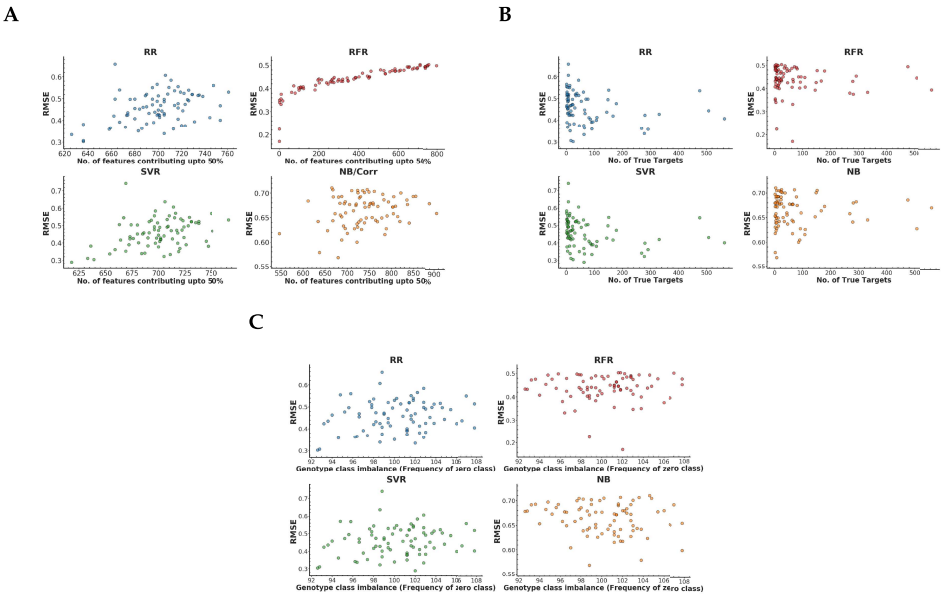


Figure 8: Scatter plots of genotype prediction performance on yeast data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**) and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). Genes on the same chromosome were excluded as predictors for each SNP.

showed no relation with the number of selected features, the number of true targets, or the genotype class balance (Fig. 11).

Although AUROCs generally correlated between methods (Fig. 10B), in line with the correlation of RMSE values, AUROC values tended to be systematically higher for SVR and RR compared to RFR and univariate correlations. Interestingly, univariate correlation and SVR share the same number of TF eQTLs with AUROC > 0.70 (10), only 5 were common and each method had five TFs not found by the other method. (Fig.13).

4.2.2 Genotype and trans-eQTL prediction performance do not correlate

Similar to the DREAM data we again observed poor correlation between genotype and trans-eQTL prediction performance (Fig. 12).

4.2.3 Feature selection in random forest produces a map of transcriptional hotspots

Transcriptional hotspots are regions of the genome associated with widespread changes in gene expression [13]. We learned prediction models for all 2,884 SNPs in the yeast genome that were associated with local changes in gene expression and plotted the RMSE for each predicted SNP against its genome position. RFR showed a wide variation in RMSE values for SNPs, across the whole genome, allowing to delineate genomic ranges with high and low regulatory activity. Whereas RR and SVR showed much less variation, and did not allow to separate high and low activity regions on most chromosomes (Fig. 14). Interestingly, the regions detected by RFR overlapped only partially with traditional hotspot maps based on univariate correlations (Supp. Fig. S22), again suggesting that non-linear methods like random forest may detect biological signals missed by traditional methods.

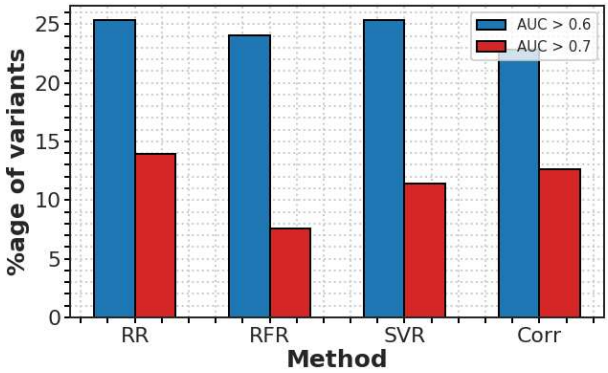


Figure 9: Bar plots show the number of variants with trans-eQTL target prediction AUROC $\geq$ 0.6 (blue) and $\geq$ 0.7 (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). Genes on the same chromosome were excluded as predictors for each SNP.

(A) **(B)**

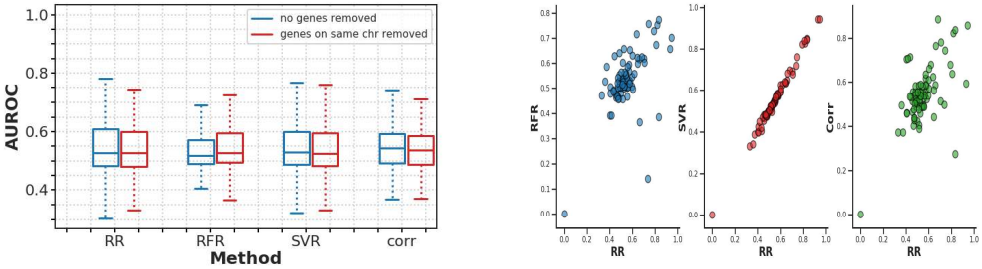


Figure 10: Trans-eQTL target prediction performance on yeast data. **(A)** Boxplots show the distribution of AUROC values using all genes (blue) and excluding genes on the same chromosome (red) as predictors, for random forest regression (RFR), support vector regression (SVM), ridge regression (RR), and univariate correlation (Corr). **(B)** Scatter plots show AUROC values of classification methods RFR, SVR, and univariate correlation (Corr) vs RR for all variants. Genes on the same chromosome were excluded as predictors for each SNP.

5 Discussion

In this study we analyzed the use of machine learning methods for genotype prediction in high-dimensional multi-trait GWAS. The basic hypotheses of reverse genotype prediction from multiple trait combinations are that variants whose genotypes can be predicted with higher accuracy are more likely to have a true effect on a large number of the measured traits, and that feature importances or coefficients in the trained models indicate the strength of association between variants and individual traits. However, existing studies have not presented conclusive evidence for these hypotheses, because they only performed downstream analysis for the highest scoring variants, and only considered linear models. Here we performed an in-depth validation of various machine learning methods for reverse genotype prediction in the context of trans-eQTL analysis, including univariate, ridge regression, random forest, and support vector regression, using both simulated and real transcriptional regulatory networks to define ground-truth sets of trans-eQTL target sets.

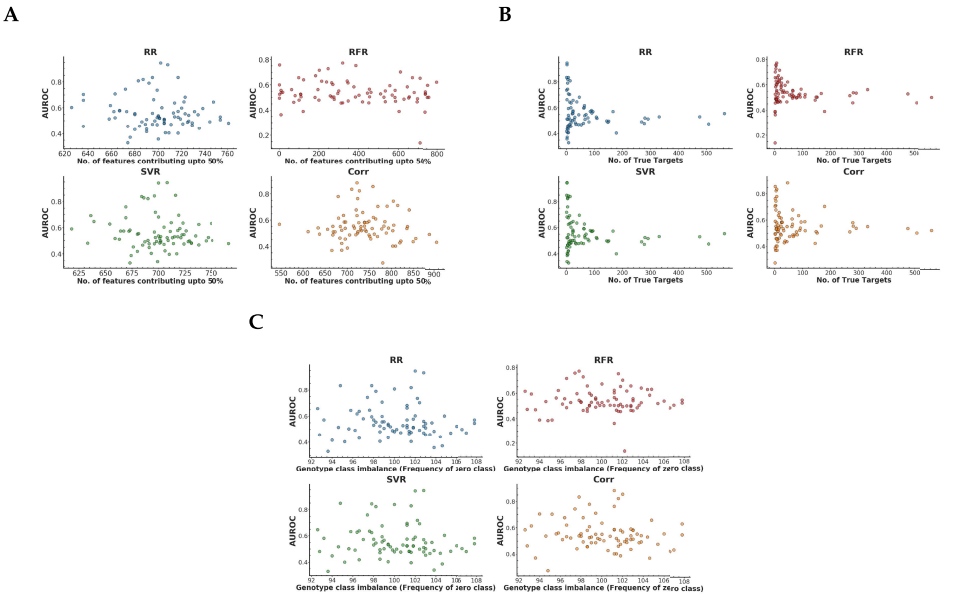


Figure 11: Scatter plots of trans-eQTL target prediction performance (AUROC) on yeast data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**) and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). Genes on the same chromosome were excluded as predictors for each SNP.

Our results support the basic hypotheses only partially. In particular, although genotype prediction performance indeed varied across genetic variants, there was no relation between genotype prediction performance and the number of gene expression traits affected by a variant, nor with the accuracy of predicting individual trans-eQTL target genes from model feature importances or coefficients. This is important, because it shows that in the absence of ground-truth information, we cannot use RMSE to select variants for which model features will overlap best with true trans-associated genes. This was further illustrated by the fact that random forest regression performed best at the genotype prediction task, but performed worst on the trans-eQTL prediction task.

The only systematic relation we observed, both in the simulated and the yeast data, was a negative correlation between genotype prediction performance and number of model features, suggesting that if a variant can be predicted well, it can be done with a relatively small number of traits.

While RMSE cannot be used to select variants with good trans-eQTL prediction performance, we did observe that model feature importances or coefficients were generally predictive of how likely a given gene is a true trans-eQTL target of a given variant. Predictive performance was very strong in simulated data, with more than 75% of variants obtaining an AUROC greater than 80%, but also in yeast, 15-20% of variants obtained an AUROC greater than 70%.

An important goal of multi-trait GWAS is to distinguish between variants that are associated with high vs low number of traits. Interestingly, we found that only random forest, but not SVR or ridge regression, resulted in models with a wide variation in the number of selected features across variants. However this involved use of a simple, heuristic strategy for feature selection, and further research to finetune this result will be required.

One aspect of multi-trait GWAS not considered in this study is statistical inference. For linear methods, the null distribution of the model fit score under the assumption of no association can be approximated

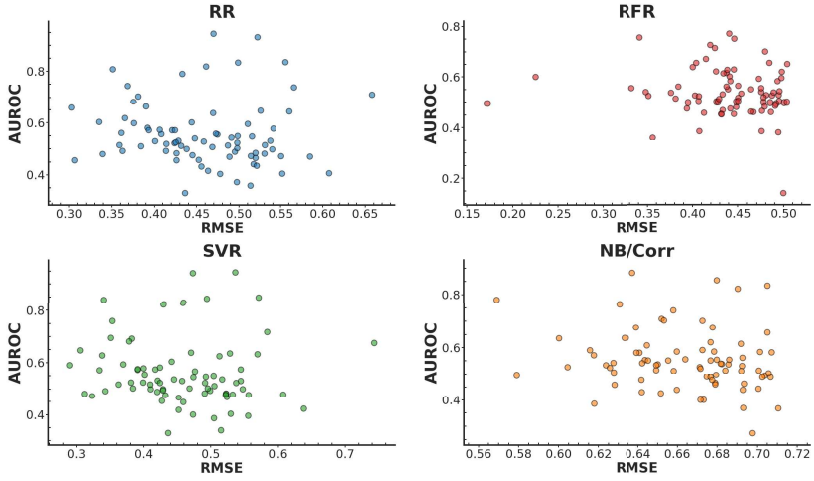


Figure 12: Scatter plots show trans-eQTL target prediction accuracy (AUROC) vs genotype prediction accuracy for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes/univariate correlation (NB/Corr) on yeast data. Genes on the same chromosome were excluded as predictors for each SNP.

analytically to obtain a p-value for the significance of any observed score. Non-linear methods such as random forest or SVM require a large number of permutations for each variant separately to obtain a p-value, which becomes computationally infeasible for a large number of variants. However approximate methods may yet overcome this hurdle [18]. More importantly though, since our results indicate that model fit is not related to either the strength or extent of true biological relations, the relevance of performing statistical inference on this test statistic is in doubt.

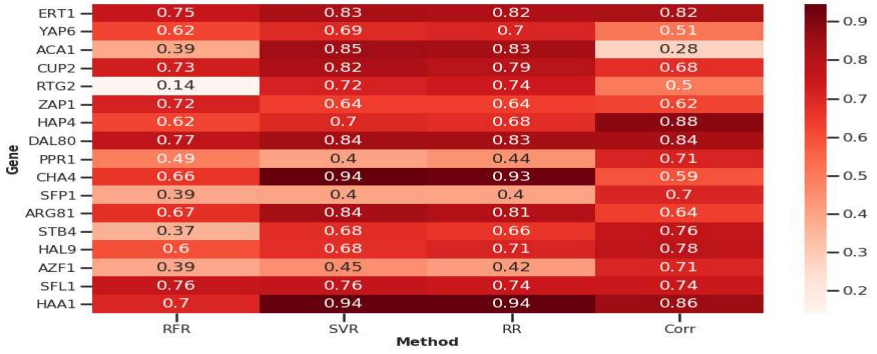


Figure 13: AUROC values for genes where at least one of the four methods (RFR, SVR, RR, Corr) gives AUROC above 0.7.

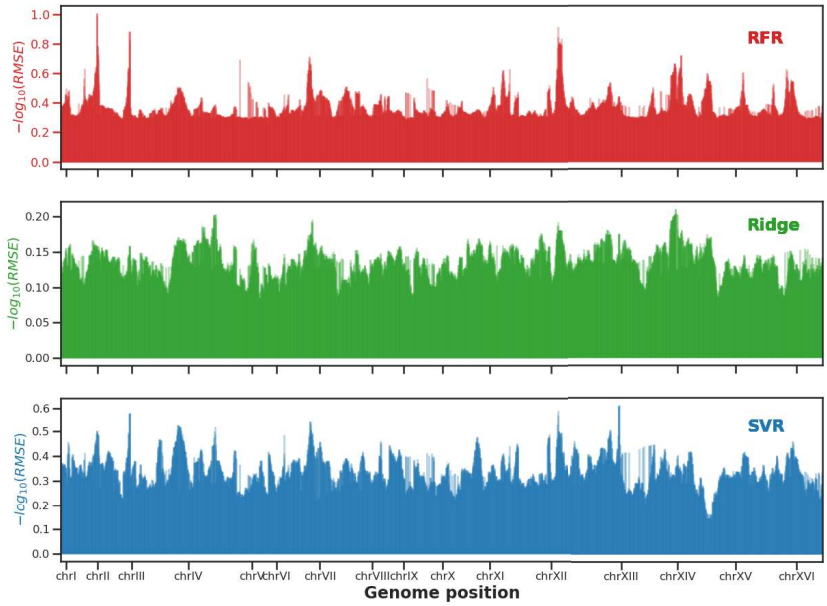


Figure 14: Expression hotspot maps showing the negative log transformed RMSE values vs genome position for 2884 SNPs in the yeast genome, for random forest (RF, top), ridge regression (Ridge, middle), and support vector regression (SVR, bottom). Genes on the same chromosome were excluded as predictors for each SNP.

Another area of future research concerns the generalization to other organisms, in particular human. We focused on realistic simulated data from the DREAM project and data from the eukaryotic model organism yeast, due to the availability of data from a study with extraordinarily large sample size and extensive, high-quality ground-truth transcriptional interaction data. The availability of ground-truth associations also motivated our choice of studying gene expression traits. It will be of interest to expand this work to other types of traits, including protein and metabolite levels, as well as high-dimensional phenotypic traits such as images.

In summary, feature importance weights in machine learning models that predict genotypes from high-dimensional sets of traits identify biologically relevant variant-trait associations, but comparing the relative importance of variants through these models in a GWAS-like manner using a single test statistic remains an open challenge.

References

[1] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nature reviews genetics*, vol. 9, no. 5, pp. 356–369, 2008.

[2] T. A. Manolio, "Bringing genome-wide association findings into clinical use," *Nature Reviews Genetics*, vol. 14, no. 8, pp. 549–558, 2013.

[3] D. B. Allison, B. Thiel, P. S. Jean, R. C. Elston, M. C. Infante, and N. J. Schork, "Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages," *The American Journal of Human Genetics*, vol. 63, no. 4, pp. 1190–1201, 1998.

[4] M. A. Ferreira and S. M. Purcell, "A multivariate test of association," *Bioinformatics*, vol. 25, no. 1, pp. 132–133, 2009.

[5] T. E. Galesloot, K. Van Steen, L. A. Kiemeney, L. L. Janss, and S. H. Vermeulen, "A comparison of multivariate genome-wide association methods," *PloS one*, vol. 9, no. 4, p. e95923, 2014.

[6] W. Van Rheenen, W. J. Peyrot, A. J. Schork, S. H. Lee, and N. R. Wray, "Genetic correlations of polygenic disease traits: from theory to practice," *Nature Reviews Genetics*, vol. 20, no. 10, pp. 567–581, 2019.

[7] A. A. Shabalin, "Matrix eqtl: ultra fast eqtl analysis via large matrix operations," *Bioinformatics*, vol. 28, no. 10, pp. 1353–1358, 2012.

[8] J. Qi, H. F. Asl, J. Björkegren, and T. Michoel, "krux: matrix-based non-parametric eqtl discovery," *BMC bioinformatics*, vol. 15, no. 1, pp. 1–7, 2014.

[9] H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, and O. Delaneau, "Fast and efficient qtl mapper for thousands of molecular phenotypes," *Bioinformatics*, vol. 32, no. 10, pp. 1479–1485, 2016.

[10] P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. Coin, "Multiphen: joint model of multiple phenotypes can increase discovery in gwas," *PloS one*, vol. 7, no. 5, p. e34861, 2012.

[11] S. Banerjee, F. L. Simonetti, K. E. Detrois, A. Kaphle, R. Mitra, R. Nagial, and J. Soeding, "Reverse regression increases power for detecting trans-eqtls," *bioRxiv*, 2020.

[12] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, and A. D. N. Initiative, "From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps," *Bioinformatics*, vol. 28, no. 18, pp. i619–i625, 2012.

[13] F. W. Albert, J. S. Bloom, J. Siegel, L. Day, and L. Kruglyak, "Genetics of trans-regulatory variation in gene expression," *Elife*, vol. 7, p. e35471, 2018.

[14] P. T. Monteiro, J. Oliveira, P. Pais, M. Antunes, M. Palma, M. Cavalheiro, M. Galocha, C. P. Godinho, L. C. Martins, N. Bourbon, *et al.*, "Yeastract+: a portal for cross-species comparative genomics of transcription regulation in yeasts," *Nucleic acids research*, vol. 48, no. D1, pp. D642–D649, 2020.

[15] J. Friedman, T. Hastie, R. Tibshirani, *et al.*, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

[16] A. Pinna, N. Soranzo, I. Hoeschele, and A. de la Fuente, "Simulating systems genetics data with sysgensim," *Bioinformatics*, vol. 27, no. 17, pp. 2459–2462, 2011.

[17] A. D. Yates, P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, *et al.*, "Ensembl 2020," *Nucleic acids research*, vol. 48, no. D1, pp. D682–D688, 2020.

[18] T. A. Knijnenburg, L. F. Wessels, M. J. Reinders, and I. Shmulevich, "Fewer permutations, more accurate p-values," *Bioinformatics*, vol. 25, no. 12, pp. i161–i168, 2009.

Article III

RFPHEN2GEN: A MACHINE LEARNING BASED ASSOCIATION STUDY OF BRAIN IMAGING PHENOTYPES TO GENOTYPES

A PREPRINT

Muhammad Ammar Malik
Department of Informatics
University of Bergen
PO Box 7803, 5020 Bergen, Norway
muhammad.malik@uib.no

Alexander Lundervold
Department of Computer Science
Western Norway University of Applied Sciences
PO Box 7030, 5020 Bergen, Norway
alexander.selvikvag.lundervold@hvl.no

Tom Michoel
Department of Informatics
University of Bergen
PO Box 7803, 5020 Bergen, Norway
tom.michoel@uib.no

April 4, 2022

ABSTRACT

Imaging genetic studies aim to find associations between genetic variants and imaging quantitative traits. Traditional genome-wide association studies (GWAS) are based on univariate statistical tests, but when multiple traits are analyzed together they suffer from a multiple-testing problem and from not taking into account correlations among the traits. An alternative approach to multi-trait GWAS is to reverse the functional relation between genotypes and traits, by fitting a multivariate regression model to predict genotypes from multiple traits simultaneously. However, current reverse genotype prediction approaches are mostly based on linear models. Here, we evaluated random forest regression (RFR) as a method to predict SNPs from imaging QTs and identify biologically relevant associations. We learned machine learning models to predict 518,484 SNPs using 56 brain imaging QTs. We observed that genotype regression error is a better indicator of permutation p-value significance than genotype classification accuracy. SNPs within the known Alzheimer disease (AD) risk gene APOE had lowest RMSE for lasso and random forest, but not ridge regression. Moreover, random forests identified additional SNPs that were not prioritized by the linear models but are known to be associated with brain-related disorders. Feature selection identified well-known brain regions associated with AD, like the hippocampus and amygdala, as important predictors of the most significant SNPs. In summary, our results indicate that non-linear methods like random forests may offer additional insights into phenotype-genotype associations compared to traditional linear multi-variate GWAS methods.

Keywords genome-wide association studies · neuroimaging genetics · alzheimer's disease · multi-trait GWAS · genotype prediction

1 INTRODUCTION

Neuroimaging genetics, also known as imaging genomics or imaging genetics, is a useful tool to investigate the associations between genetic variants and variation in brain structure among individuals [1]. The discovery of biomarkers jointly from imaging and genetic data helps us to better understand the underlying pathological processes of neuropsychiatric and neurodegenerative diseases [2, 3]. Moreover, neuroimaging may help us discover the genetic pathways through which genes affect the above-mentioned diseases, by identifying associations between causal genes and vari-

ations in brain regions [4,5]. And lastly, imaging genetics studies have been shown to have increased statistical power when compared with conventional case-control studies and therefore have decreased sample size requirement [6].

Recently a large number of neuroimaging studies have been conducted to explore the association between neurodegenerative disease and brain structure [1,7–10]. Some of these studies have focused on understanding the genetic causes of these diseases (for example Alzheimer's disease), whereas the other genome-wide association studies (GWAS) focus on identifying the genetic variations that influence brain structure and function. A common issue with most imaging genetics studies is the reduction in either imaging or genetic data (or sometimes both). For example, whole-brain studies have mostly focused on a small number of genetic variants [11–14], whereas whole-genome studies have focused on a limited number of imaging quantitative traits (QTs) [15,16]. This restriction in either genotype or phenotype data can greatly hinder our ability to identify important associations.

A typical procedure to investigate genotype-phenotype associations is to conduct univariate linear regression or analysis of variance (ANOVA) tests for each genetic variant against each trait separately. However, when multiple traits are studied simultaneously this approach ignores correlations among traits and leads to a high multiple-testing burden. Other approaches for multi-trait GWAS are based on multivariate analysis of variance (MANOVA) or canonical correlation analysis (CCA) [17]. But these are applicable only to studies with a small number of traits. A promising alternative approach to multi-trait GWAS has been to reverse the functional relation between genotypes and traits and fit a multivariate regression model that predicts genotypes from multiple traits simultaneously, instead of the usual approach to regress traits on genotypes [18]. In fact, in [19], we showed that more traditional multi-trait GWAS methods such as CCA can also be described as reverse genotype prediction methods.

Reverse genotype prediction has also been considered in the context of imaging genetics. For example, in [8] a task-correlated longitudinal sparse regression approach was used to investigate associations between phenotype markers and Alzheimer's disease relevant SNPs belonging to top 40 AD relevant genes.

Thus far, efforts to extend reverse genotype prediction methods to the high-dimensional setting (in either SNP or feature dimension) have mostly focused on gene expression traits. For instance, a recent study used L2-regularized linear regression of SNPs on gene expression traits to identify trans-acting expression quantitative trait loci (trans-eQTLs), and showed that this approach aggregates evidence from many small trans-effects while being unaffected by strong expression correlations [20].

However, several limitations and open questions remain in multi-trait GWAS. For example, existing studies mostly rely on linear models that search for linear combinations of traits associated to the SNP, but there is no *a priori* biological evidence to support the use of only linear combinations. Moreover, even though L2-regularization allows to deal with high-dimensional traits, it does not address the problem of feature selection. For instance, in [20] a secondary set of univariate tests is carried out to select genes associated to trans-eQTLs identified by the initial multivariate regression.

In [19], using gene expression data from a cross between two yeast strains, we found that feature coefficients of machine learning models (lasso, ridge, linear SVM, and random forest) correlated with the strength of association between variants and individual traits, and were predictive of true trans-eQTL target genes. However, to the best of our knowledge, a genome-wide analysis of machine learning methods for reverse genotype prediction in human GWAS has not yet been conducted.

In this study, we explored the use of a non-linear machine learning method, specifically random forests, for predicting genotypes from brain imaging phenotypes. Our overall hypothesis is that genetic variants whose genotypes can be predicted with higher accuracy from imaging traits are more likely to affect some or all of the traits under consideration than variants whose genotypes cannot be predicted well, and that feature weights in the fitted models measure the

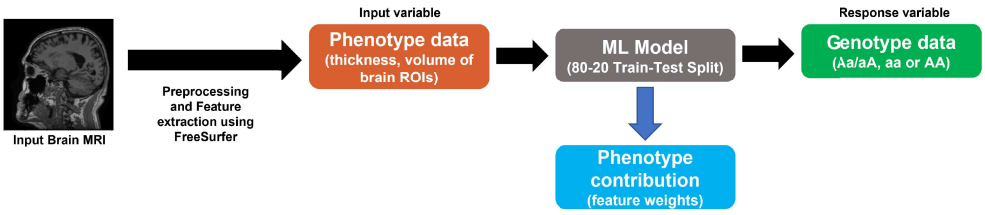


Figure 1: Flowchart of the approach used in this study

strength of biological association between a SNP and a QT. We compare the results of random forest with two well-known linear methods, lasso and ridge regression.

2 MATERIALS AND METHODS

2.1 Dataset

Data used in this study was obtained from the ADNI database (`adni.loni.ucla.edu`). The baseline T1-weighted MRI images from the four phases of the ADNI study, the Illumina SNP genotyping data, demographic information, APOE genotype, and baseline diagnosis was downloaded from the ADNI database. The demographic information of the samples used in the study can be found in Table 1. Details about the standardized imaging protocols used in ADNI can be found in `https://adni.loni.usc.edu/methods/documents/mri-protocols/`.

Table 1: Demographic information of the samples used in the study. CN: Controls, MCI: Mild Cognitive Impairment, AD: Alzheimer's Disease

	CN	**MCI**	**AD**
No. of subjects	211	359	178
Gender(M/F)	112/99	234/125	94/84
Baseline age(years:mean $\pm$ SD)	75.7$\pm$4.9	74.7$\pm$7.3	75.4$\pm$7.3
Education(years:mean $\pm$ SD)	16.0$\pm$2.8	15.7$\pm$3.0	14.6$\pm$3.2
Race (Caucasian/Non-Caucasian)	191/20	325/34	161/17

2.2 MRI data and imaging phenotype extraction

We extracted subcortical segmentation and cortical parcellation from the T1-weighted images using FreeSurfer v6.0 [21] to obtain imaging phenotypes. Following [1] we defined 56 volumetric and cortical thickness values mentioned in(Table 2).

2.3 SNP genotypes

The SNP data from ADNI database were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA, USA). The genotype data consists of 620,901 SNPs. The SNP data was screened using the following quality control (QC) steps: (1) call rate check per subject ($\geq 90\%$) and per SNP marker ($\geq 90\%$), (2) gender check (3) marker removal according to the minor allele frequency (MAF) $\geq 5\%$ and (4) Hardy-Weingberg equilibrium (HWE) test of $p \leq 10^{-6}$. The remaining missing genotype values were imputed as the modal values. After the QC procedure, 749 subjects and 518,484 SNPs remained in the data. The APOE gene is one of the important causal genes for AD, but the previously identified APOE SNPs (rs429358/rs7412) were not available on the Illumina array. Therefore, the APOE genotype was coded from the `ADNIMERGE.csv` file prepared by the ADNI study by using the number of APOE-ε4 risk alleles.

2.4 Genotype prediction model

For genotype prediction using machine learning models, the phenotype values were treated as explanatory variables whereas the genotype values of SNPs were treated as the response variables, learning separate models for each SNP (Fig. 1). We used both regression and classification to predict SNPs. The prediction performance in the regression setting was measured by computing the root mean squared error (RMSE) between the predicted and the actual genotype value of a variant. The classification performance was measured as the ratio between the number of correctly predicted samples and the total number of samples in the test set.

We compared the performance of random forest regression and classification (RFR), with linear machine learning methods, in particular ridge regression (RR) and lasso regression (LR).

2.5 Phenotype contribution

We used the feature weights of the machine learning models for measuring the association between SNPs and specific brain regions (in the case of RFR feature importances were used). The absolute values of the feature weights were used and normalized such that they sum up to 1.

Table 2: List of FreeSurfer phenotypes defined as volume or cortical thickness of various region of interests (ROI)[a]

Phenotype description (Phenotype ID)	
Volume of amygdala (AmygVol)	Volume of cerebral cortex (CerebCtx)
Volume of cerebral white matter (CerebWM)	Volume of hippocampus (HippVol)
Volume of inferior lateral ventricle (InfLatVent)	Volume of lateral ventricle (LatVent)
Thickness of entorhinal cotrex (EntCtx)	Thickness of fusiform gyrus (Fusiform)
Thickness of inferior parietal gyrus (InfParietal)	Thickness of inferior temporal gyrus (InfTemporal)
Thickness of middle temporal gyrus (MidTemporal)	Thickness of parahippocampal gyrus (Parahipp)
Thickness of posterior cingulate (PostCing)	Thickness of postcentral gyrus (Postcentral)
Thickness of precentral gyurs (Precentral)	Thickness of precuneus (Precuneus)
Thickness of superior frontal gyrus (SupFrontal)	Thickness of superior parietal gyurs (SupParietal)
Thickness of superior temporal gyrus (SupTemporal)	Thickness of supramarginal gyrus (Supramarg)
Thickness of temporal pole (TemporalPole)	
Mean thickness of caudal anterior cingulate, isthmus cingulate, posterior cingulate, and rostral anterior cingulate (MeanCing)	
Mean thickness of caudal midfrontal, rostral midfrontal, superior frontal, lateral orbitofrontal, and medial orbitofrontal gyri and frontal pole (MeanFront)	
Mean thickness of inferior temporal, middle temporal, and superior temporal gyri (MeanLatTemp)	
Mean thickness of fusiform, parahippocampal, and lingual gyri, temporal pole and transverse temporal pole (MeanMedTemp)	
Mean thickness of inferior and superior parietal gyri, supramarginal gyrus, and precuneus (MeanPar)	
Mean thickness of precentral and postcentral gyri (MeanSensMotor)	
Mean thickness of inferior temporal, middle temporal, superior temporal, fusiform, parahippocampal , and lingual gyri, temporal pole and transverse temporal pole (MeanTemp)	

[a]Each of the 28 phenotypes mentioned corresponds to two phenotypes, one for the left side and the other for the right side.

2.6 Permutation tests

We conducted permutation tests [22] using 100 permutations for a subset of 876 SNPs to determine the statistical significance of RMSE and classification accuracies. *P*-values were calculated as the fraction of permutation values that are at least as extreme as the original statistic (RMSE or classification accuracy) derived from non-permuted data, that is

$$p = \frac{i+1}{N} \tag{1}$$

where N denotes the number of permutations, and i denotes the number of times the performance measure of the permuted SNP was found to be better than the unpermuted measure for that SNP.

2.7 Rank-based p-values

Because conducting permutation tests for all 518,484 SNPs was computationally prohibitive, we computed the p-values for all the SNPs in the genome by ranking the SNPs based on the obtained RMSE values (for tied ranks the average of the rank was assigned) and dividing the rank by the total number of SNPs (Eq. 2).

$$p = \frac{k}{N} \tag{2}$$

where N denotes the total number of SNPs, and k denotes the rank of the particular SNP. These rank-based p-values for visualizing relative prediction performance of SNPs in Manhattan plots.

2.8 Experimental settings

The machine learning models including lasso regression, ridge regression and random forest regression and classification were implemented using the *scikit-learn* Python library. The dataset was divided into 80-20 training-test split. Both the input and output data were normalized to have unit standard deviation and zero mean. The volumes of the ROIs (phenotype data) were corrected for age, gender, education, and baseline Intracranial Volume (ICV_bl) as estimated by FreeSurfer v6.0.

3 Results

3.1 Genotype classification vs genotype regression

SNP genotypes take discrete values (0,1,2) counting the number of alternative alleles in an individual. Since it is generally assumed that the effect of a SNP on a quantitative trait increases or decreases with the number of alternative alleles, genotype prediction could either use classification models (emphasizing the discrete nature) or regression models (emphasizing the ordinal nature). To compare and select between regression or classification for the genotype prediction task, we performed permutation tests to convert prediction performance (*root mean squared error* in case of regression and *accuracy* in case of classification) to p-values.

Since performing permutation tests across the whole genome was computationally unfeasible, we considered 876 SNPs belonging to the top 40 AD-related genes as mentioned on `alzgene.org`. For the permutation tests, the target labels were randomly permuted 100 times and the results were compared with unpermuted data. Based on eq. 1 the best possible p-value in this scenario is 0.01, i.e. none of the permuted sets performed better than the unpermuted set.

The correlation between *p-values* and the prediction task performance was found to be higher for regression (**Spearman correlation coefficient 0.79**) than for classification (**Spearman correlation coefficient -0.55**) (Fig. 2). Further investigation showed that the poor correspondence between classification accuracy and p-values was due to strong class imbalance (Supp. Fig. S1): if alternative alleles of a SNP are relatively rare (few individuals in the 1 and/or 2 class), classification accuracy can be high by randomly assigning individuals to classes based on the class frequencies such that classification accuracy from real features is no better than from random features. In contrast, as indicated in Fig. 2, genotype regression was less affected by class imbalance differences among SNPs.

Since RMSE is a better indicator of non-random prediction performance, regardless of minor allele frequency differences between SNPs, than classification accuracy, we decided to proceed with regression analysis for the remainder of our experiments.

3.2 Genotype prediction across the whole genome

To the best of our knowledge, none of the previous reverse genotype prediction studies were conducted across the whole human genome. We performed reverse genotype regression for all 518,484 SNPs that passed QC using Lasso, Ridge, and Random Forest regression on 56 volumetric and cortical thickness image features. For visualization purposes, we converted the RMSE value to a rank-based p-value for each SNP (eq. 2).

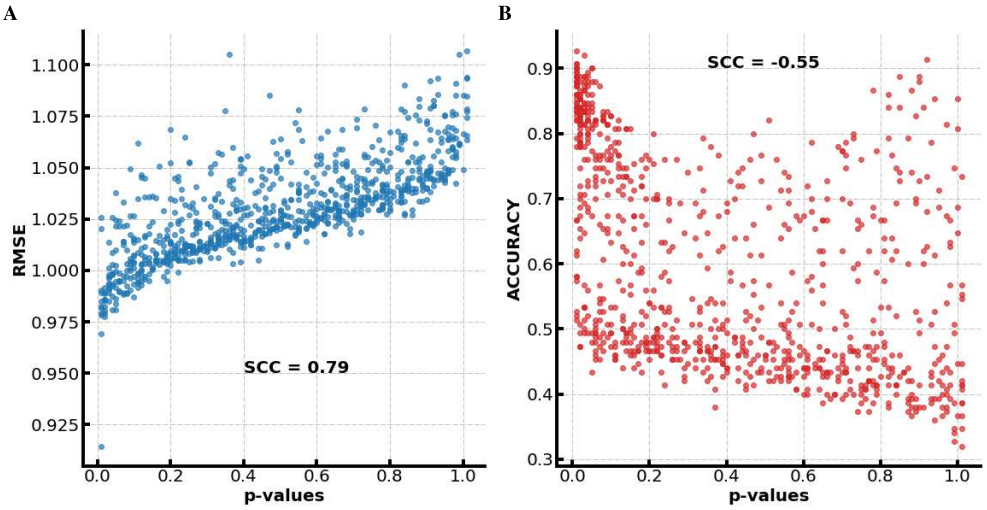


Figure 2: Permutation tests for Random Forest. **(A)** Regression (RMSE), **(B)** Classification (Accuracy). SCC: Spearman Correlation Coefficient

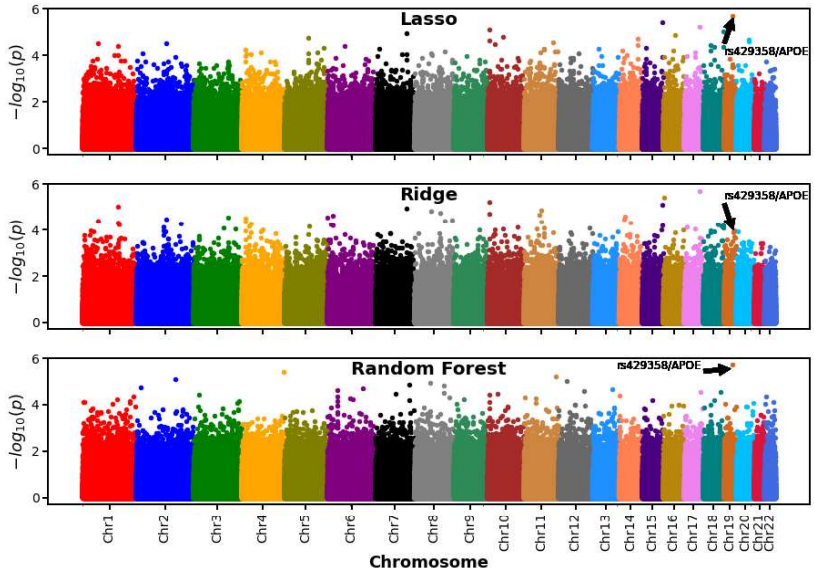


Figure 3: Negative log transformed p-values plotted across the whole genome for Random Forest Regression (RFR), Ridge Regression (RR), Lasso Regression (LR).

In support of our overall hypothesis, we observe that genotype prediction performance is variable across the genome and that some SNPs can be predicted with lower RMSE than the random background (Fig. 3). Moreover, the top SNP achieving the lowest RMSE value for random forests (RMSE=0.915) and lasso (RMSE=0.932) was rs429358/APOE, the best known causal gene for Alzheimer's disease. In contrast, ridge regression picked up rs1864685/SLC39A11 as the best SNP (RMSE=0.926), with rs429358/APOE being ranked 27th (RMSE=0.950).

Despite this difference on the APOE prediction, overall the linear methods (lasso and ridge) produced nearly the same ranking of SNPs (Spearman correlation coefficient 0.91), whereas random forests predictions were clearly distinct from the linear ones (Spearman correlation coefficients 0.25 and 0.29 with ridge and lasso, respectively) (Supp. Fig. S2). Different peaks across the methods further illustrate point to the fact that distinct SNPs are identified using different methods, also near the top of the ranking (Supp. Fig. S3).

To test whether any of these results were affected by population structure in the data, we repeated the analysis using only Caucasian subjects, and observed no significant difference in the RMSE distribution (Supp. Fig. S4).

Since lasso and ridge regression gave comparable results, and ridge regression has been well studied as a multivariate GWAS method [20], we focused the remainder of our analysis on the random forest predictions.

3.3 Identification of genetically associated imaging markers

One of the aims of this study is to test if a subset of genetic variants and imaging phenotypes related to brain-related disorders can be identified from the feature weights of the best performing SNPs. Figure 4 shows the clustermap of feature weights of all the imaging features and the top 1000 SNPs identified by the random forest method. A clear pattern can be noticed where groups of SNPs not colocated on the genome associate to similar groups of imaging features. This is consistent with previous findings from univariate approaches where genetic variants affecting one cortical region were found to often also affect other cortical regions [23]. In particular, brain regions previously

known to be associated with brain-related disorders, such as the hippocampus, amygdala, temporal lobe, entorhinal cortex and lateral ventricles, clustered together.

The hippocampus has been shown to be related to memory and cognition [24]. The amygdala is known to play a primary role in decision making and memory [25], and its atrophy is associated with Alzheimer's disease [26]. The temporal lobe is largely responsible for creating and preserving both conscious and long-term memory [27]. The temporal pole, also known as Broadman area 38, is among the earliest affected by Alzheimer's disease, frontotemporal dementia and frontotemporal lobar degeneration [28]. The entorhinal cortex is one of the first regions to be affected in Alzheimer's disease [29] and moreover, the lateral ventricles are known to be enlarged in patients with Alzheimer's [30], schizophrenia [31], bipolar and major depressive disorders [32].

3.4 Characterization of top genetic variants

To further investigate the relevance of our findings, we analyzed the top 10 SNPs that were identified by the random forest method in detail using literature search and data from the GWAS catalog [33] (Table 3.4). As previously mentioned, the SNP rs429358 belonging to the APOE gene was identified as the top SNP. APOE genotype is the most well-known genetic risk factor for AD [34]. The top 3 features identified for APOE are the volumes of the hippocampus, amygdala and the thickness of the entorhinal cortex, all of these are known to be among the first regions to be affected in Alzheimer's disease [25, 29].

While, apart from APOE, the rest of the top SNPs do not belong to known genes associated with AD, all of them have been associated previously with other brain related functions (Table 3.4). For example, F11-AS1, SYBU, GACAT3, and LAMA2 are known to be associated with traits like cognitive decline, cortical thickness and surface area, and brain volume measurements [35, 36]. Moreover, FLI1 and TMEM213 have been shown to be associated with paired helical filaments (PHF)-tau measurements [37]. The tau protein is also a well-known Alzheimer's disease biomarker [38].

4 Discussion

In this paper, we analyzed the use of machine learning methods for associating brain imaging phenotypes to genotypes, in a reverse genotype prediction setting, using data from 56 imaging quantitative traits and more than half a million SNPs. The basic hypotheses of reverse genotype prediction from multiple trait combinations are that variants whose genotypes can be predicted with higher accuracy are more likely to have a true effect on one or more of the measured traits, and that feature importances or coefficients in the trained models indicate the strength of association between variants and individual traits. However, existing studies only considered linear models. Here we performed a comparison of linear models including ridge regression and lasso regression with a non-linear random forest method, to find associations between brain imaging phenotypes and genotypes.

Our results show that lasso and random forest regression, but not ridge regression, identified a SNP in the APOE gene as the best performing variant. When compared across the whole genome, random forests produced a distinct list of

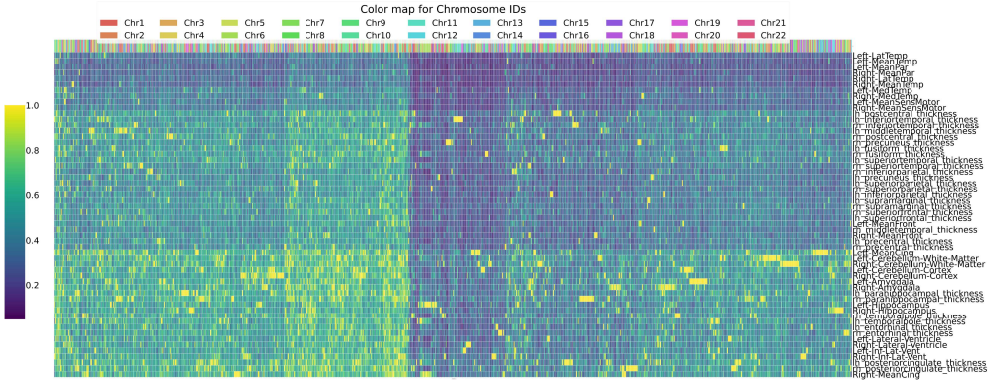


Figure 4: Heat maps of feature weights for top 1000 SNPs identified by random forest.

Table 3: Top SNPs identified by random forest

SNP	**Gene**	**Associated traits**	**Features identified**
rs429358	APOE	HDL cholesterol, LDL cholesterol, Alzheimer's disease, Fatty liver	Hippocampus, Amygdala, Entorhinal cortex
rs2084729	F11-AS1	cognitive decline, cortical thickness, cortical surface area, brain volume measurements	Entorhinal cortex, Amygdala, Posterior cingulate cortex
rs622735	FLI1	PHF-tau measurments	Parahippocampal gyrus, Superior Temporal gyrus
rs3749030	TTC21B	intelligence, cognitive function, measurements	MeanCing, Cerebellum White Matter, Postcentral gyrus
rs11048593	ITPR2	HDL cholesterol, total cholestrol, LDL cholesterol levels, unipolar depression	Cerebellum White Matter, MeanCing, Amygdala
rs6473635	PXDNL	cognitive impairment measurement, unipolar depression, bipolar disorder, schizophrenia	MeanCing, Temporal pole gyrus, Amygdala
rs896189	TMEM213	PHF-tau measurments	inferior parietal gyrus, postcentral gyrus, fusiform gyrus
rs12678956	SYBU	Schizophrenia, cortical surface area	Inferior lateral ventricle, posterior cingulate cortex, fusiform gyrus
rs6714955	GACAT3	cortical surface area measurement, neuroimaging measurement, brain measurement	posterior cingulate cortex, Cerebellum Cortex, MeanCing
rs9482965	LAMA2	cognitive ability, cortical surface area, cortical thickness	MeanTemp, posterior cingulate cortex, Inferior lateral ventricle

selected SNPs, based on RMSE prediction performance, than the linear methods, which were highly similar to each other. Literature search and existing GWAS data showed that the top SNPs identified by random forests showed are all located in or near genes that have been previously associated with brain-related disorders, supporting the use of non-linear multi-variate GWAS methods to identify distinct genetic variants than those selected by conventional linear methods. Further in-depth study of the genes identified in this analysis may contribute to a better understanding of their association with brain function.

Extending the analysis to the top 1,000 SNPs predicted by random forests, we observed a clustering of image features, showing that groups of variants, not colocated on the genome, tend to associate with similar brain regions or features.

While reverse genotype prediction correctly picks up these correlations between the phenotypic traits, the corresponding correlations and shared effects between SNPs are currently ignored, since reverse genotype prediction approaches tend to learn prediction models for each SNP individually. Thus, a logical extension of our approach would be to use multi-task regression, i.e. to predict multiple SNPs simultaneously. However, this raises important computational challenges and it may be infeasible to predict SNPs simultaneously on a genome-wide scale.

A disadvantage of using machine learning methods, in particular non-linear ones, is that the null distribution of the test statistic (RMSE) is unknown and the only way to quantify statistical significance is to compute permutation p-values, which was computationally infeasible across the whole genome. However, a more limited analysis on 876 SNPs showed that permutation p-values and random forest regression RMSE values, but not classification accuracies, showed a high degree of correlation. A possible solution could therefore be to learn a model to predict p-values from RMSE values from a suitable set of training SNPs, to be used to obtain approximate permutation p-values genome-wide.

Another limitation of the current study is that in the presence of highly correlated traits, the feature weights obtained by different methods are not necessarily robust. It would be interesting to investigate other measures of feature importance for random forest models, beyond the default ones based on gini importances, such as model-agnostic methods like permutation importance [39].

Further future research could investigate additional non-linear machine learning methods such as neural networks, including deep neural networks for predicting genotypes using MRI recordings of the brain directly instead of extracted features [40]. Moreover, since dementia is a progressive disorder, another interesting avenue to pursue would be to use longitudinal data.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding

This work was supported in part by a grant from the Research Council of Norway (grant number 312045) to T.M, and in part by a grant from the Trond Mohn Research Foundation (grant number BFS2018TMT07) to A.S.L.

References

[1] Li Shen, Sungeun Kim, Shannon L Risacher, Kwangsik Nho, Shanker Swaminathan, John D West, Tatiana Foroud, Nathan Pankratz, Jason H Moore, Chantel D Sloan, et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *Neuroimage*, 53(3):1051–1063, 2010.

[2] Ganesh Chauhan, Hieab HH Adams, Joshua C Bis, Galit Weinstein, Lei Yu, Anna Maria Töglhofer, Albert Vernon Smith, Sven J Van Der Lee, Rebecca F Gottesman, Russell Thomson, et al. Association of Alzheimer's disease GWAS loci with MRI markers of brain aging. *Neurobiology of aging*, 36(4):1765–e7, 2015.

[3] Xuan Bi, Liuqing Yang, Tengfei Li, Baisong Wang, Hongtu Zhu, and Heping Zhang. Genome-wide mediation analysis of psychiatric and cognitive traits through imaging phenotypes. *Human brain mapping*, 38(8):4088–4097, 2017.

[4] Zhao-Hua Lu, Zakaria Khondker, Joseph G Ibrahim, Yue Wang, Hongtu Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies. *NeuroImage*, 149:305–322, 2017.

[5] Guiyou Liu, Lifen Yao, Jiafeng Liu, Yongshuai Jiang, Guoda Ma, Zugen Chen, Bin Zhao, Keshen Li, et al. Cardiovascular disease contributes to Alzheimer's disease: evidence from large-scale genome-wide association studies. *Neurobiology of aging*, 35(4):786–792, 2014.

[6] Steven G Potkin, Jessica A Turner, Guia Guffanti, Anita Lakatos, Federica Torri, David B Keator, and Fabio Macciardi. Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations. *Cognitive neuropsychiatry*, 14(4-5):391–418, 2009.

[7] Hua Wang, Feiping Nie, Heng Huang, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics*, 28(2):229–237, 2012.

[8] Hua Wang, Feiping Nie, Heng Huang, Jingwen Yan, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. From phenotype to genotype: an association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs. *Bioinformatics*, 28(18):i619–i625, 2012.

[9] Meiyan Huang, Thomas Nichols, Chao Huang, Yang Yu, Zhaohua Lu, Rebecca C Knickmeyer, Qianjin Feng, Hongtu Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *Neuroimage*, 118:613–627, 2015.

[10] Meiyan Huang, Chunyan Deng, Yuwei Yu, Tao Lian, Wei Yang, Qianjin Feng, Alzheimer's Disease Neuroimaging Initiative, et al. Spatial correlations exploitation based on nonlocal voxel-wise GWAS for biomarker detection of ad. *NeuroImage: Clinical*, 21:101642, 2019.

[11] Tao Zhou, Kim-Han Thung, Mingxia Liu, and Dinggang Shen. Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model. *IEEE Transactions on Biomedical Engineering*, 66(1):165–175, 2018.

[12] Ahmad R Hariri, Emily M Drabant, and Daniel R Weinberger. Imaging genetics: perspectives from studies of genetically driven variation in serotonin function and corticolimbic affective processing. *Biological psychiatry*, 59(10):888–897, 2006.

[13] Caroline C Brun, Natasha Leporé, Xavier Pennec, Agatha D Lee, Marina Barysheva, Sarah K Madsen, Christina Avedissian, Yi-Yu Chou, Greig I De Zubicaray, Katie L McMahon, et al. Mapping the regional influence of genetics on brain structure variability—a tensor-based morphometry study. *Neuroimage*, 48(1):37–49, 2009.

[14] Li Shen, Andrew J Saykin, Moo K Chung, and Heng Huang. Morphometric analysis of hippocampal shape in mild cognitive impairment: An imaging genetics study. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, pages 211–217. IEEE, 2007.

[15] Steven G Potkin, Guia Guffanti, Anita Lakatos, Jessica A Turner, Frithjof Kruggel, James H Fallon, Andrew J Saykin, Alessandro Orro, Sara Lupoli, Erika Salvi, et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PloS one*, 4(8):e6501, 2009.

[16] Sergio E Baranzini, Joanne Wang, Rachel A Gibson, Nicholas Galwey, Yvonne Naegelin, Frederik Barkhof, Ernst-Wilhelm Radue, Raija LP Lindberg, Bernard MG Uitdehaag, Michael R Johnson, et al. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Human molecular genetics*, 18(4):767–778, 2009.

[17] Manuel AR Ferreira and Shaun M Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–133, 2009.

[18] Paul F O'Reilly, Clive J Hoggart, Yotsawat Pomyen, Federico CF Calboli, Paul Elliott, Marjo-Riitta Jarvelin, and Lachlan JM Coin. Multiphen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS one*, 7(5):e34861, 2012.

[19] Muhammad Ammar Malik, Adriaan-Alexander Ludl, and Tom Michoel. High-dimensional multi-trait GWAS by reverse prediction of genotypes. *arXiv preprint arXiv:2111.00108*, 2021.

[20] Saikat Banerjee, Franco L Simonetti, Kira E Detrois, Anubhav Kaphle, Raktim Mitra, Rahul Nagial, and Johannes Söding. Tejaas: reverse regression increases power for detecting trans-eqtls. *Genome biology*, 22(1):1–16, 2021.

[21] Bruce Fischl. FreeSurfer. *Neuroimage*, 62(2):774–781, 2012.

[22] Eugene Edgington and Patrick Onghena. *Randomization tests*. Chapman and Hall/CRC, 2007.

[23] Alexey A Shadrin, Tobias Kaufmann, Dennis van der Meer, Clare E Palmer, Carolina Makowski, Robert Loughnan, Terry L Jernigan, Tyler M Seibert, Donald J Hagler, Olav B Smeland, et al. Vertex-wise multivariate genome-wide association study identifies 780 unique genetic loci associated with cortical morphology. *NeuroImage*, 244:118603, 2021.

[24] Larry R Squire and John T Wixted. The cognitive neuroscience of human memory since hm. *Annual review of neuroscience*, 34:259–288, 2011.

[25] Katrin Amunts, O Kedo, M Kindler, P Pieperhoff, H Mohlberg, NJ Shah, U Habel, F Schneider, and K Zilles. Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. *Anatomy and embryology*, 210(5):343–352, 2005.

[26] Stéphane P Poulin, Rebecca Dautoff, John C Morris, Lisa Feldman Barrett, Bradford C Dickerson, Alzheimer's Disease Neuroimaging Initiative, et al. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Research: Neuroimaging*, 194(1):7–13, 2011.

[27] Clifford R Jack, Ronald C Petersen, Yuecheng Xu, Peter C O'Brien, Glenn E Smith, Robert J Ivnik, Eric G Tangalos, and Emre Kokmen. Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurology*, 51(4):993–999, 1998.

[28] Steven E Arnold, Bradley T Hyman, Jill Flory, Antonio R Damasio, and Gary W Van Hoesen. The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with alzheimer's disease. *Cerebral cortex*, 1(1):103–116, 1991.

[29] Usman A Khan, Li Liu, Frank A Provenzano, Diego E Berman, Caterina P Profaci, Richard Sloan, Richard Mayeux, Karen E Duff, and Scott A Small. Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical Alzheimer's disease. *Nature neuroscience*, 17(2):304–311, 2014.

[30] Sean M Nestor, Raul Rupsingh, Michael Borrie, Matthew Smith, Vittorio Accomazzi, Jennie L Wells, Jennifer Fogarty, Robert Bartha, and Alzheimer's Disease Neuroimaging Initiative. Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain*, 131(9):2443–2454, 2008.

[31] Ian C Wright, Sophia Rabe-Hesketh, Peter WR Woodruff, Anthony S David, Robin M Murray, and Edward T Bullmore. Meta-analysis of regional brain volumes in schizophrenia. *American Journal of Psychiatry*, 157(1):16–25, 2000.

[32] Matthew J Kempton, Zainab Salvador, Marcus R Munafò, John R Geddes, Andrew Simmons, Sophia Frangou, and Steven CR Williams. Structural neuroimaging studies in major depressive disorder: meta-analysis and comparison with bipolar disorder. *Archives of general psychiatry*, 68(7):675–690, 2011.

[33] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901, 2017.

[34] Lindsay A Farrer, L Adrienne Cupples, Jonathan L Haines, Bradley Hyman, Walter A Kukull, Richard Mayeux, Richard H Myers, Margaret A Pericak-Vance, Neil Risch, and Cornelia M Van Duijn. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis. *Jama*, 278(16):1349–1356, 1997.

[35] Dennis van der Meer, Oleksandr Frei, Tobias Kaufmann, Alexey A Shadrin, Anna Devor, Olav B Smeland, Wesley K Thompson, Chun Chieh Fan, Dominic Holland, Lars T Westlye, et al. Understanding the genetic determinants of the brain with mostest. *Nature communications*, 11(1):1–9, 2020.

[36] Katrina L Grasby, Neda Jahanshad, Jodie N Painter, Lucía Colodro-Conde, Janita Bralten, Derrek P Hibar, Penelope A Lind, Fabrizio Pizzagalli, Christopher RK Ching, Mary Agnes B McMahon, et al. The genetic architecture of the human cerebral cortex. *Science*, 367(6484):eaay6690, 2020.

[37] Hui Wang, Jingyun Yang, Julie A Schneider, Philip L De Jager, David A Bennett, and Hong-Yu Zhang. Genome-wide interaction analysis of pathological hallmarks in Alzheimer's disease. *Neurobiology of aging*, 93:61–68, 2020.

[38] Charles R Harrington, Elizabeta B Mukaetova-Ladinska, Richard Hills, Patricia C Edwards, E Montejo De Garcini, Michael Novak, and CM1905817 Wischik. Measurement of distinct immunochemical presentations of tau protein in Alzheimer disease. *Proceedings of the National Academy of Sciences*, 88(13):5842–5846, 1991.

[39] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[40] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

Appendices

Supplementary Information for Article I

Restricted maximum-likelihood method for learning latent variance components in gene expression data with known and unknown confounders

— Supplementary Information —

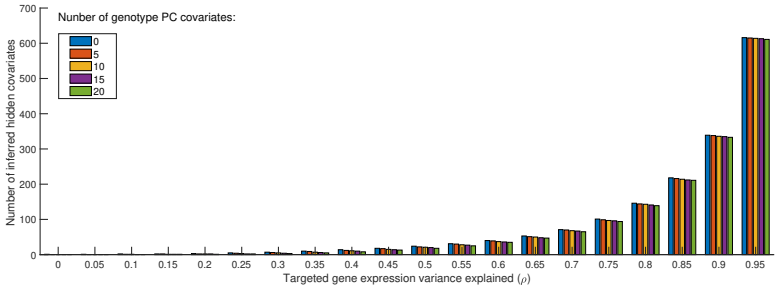
Muhammad Ammar Malik and Tom Michoel*

Computational Biology Unit, Department of Informatics, University of Bergen, PO Box 7803, 5020 Bergen, Norway

* Corresponding author, email: tom.michoel@uib.no

Supplementary Figures

A



B

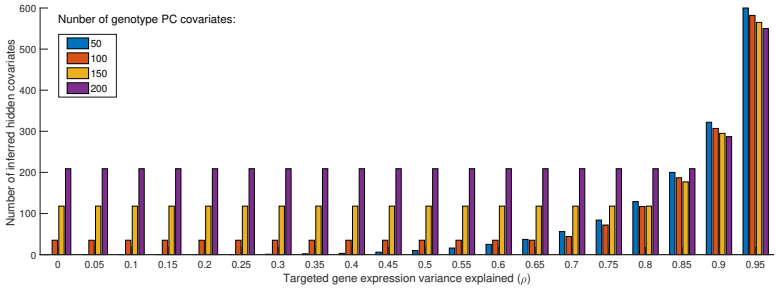


Figure S1: **A.** Number of hidden covariates inferred by LVREML as a function of the parameter ρ (the targeted total amount of variance explained by the known and hidden covariates), with θ (the minimum variance explained by a known covariate) set to retain 0, 5, 10, or 20 known covariates (genotype PCs) in the model. **B.** Same as panel **A**, with θ set to retain 50, 100, 150, or 200 genotype PCs in the model. The saturation of the number of hidden covariates with decreasing ρ for models with 100, 150, and 200 known covariates is a visual indicator that some of the dimensions in the linear subspace spanned by the known covariates do not explain sufficient variation in the expression data, and the relevance or possible redundancy of (some of) the known covariates for explaining variation in the expression data needs to be reconsidered.

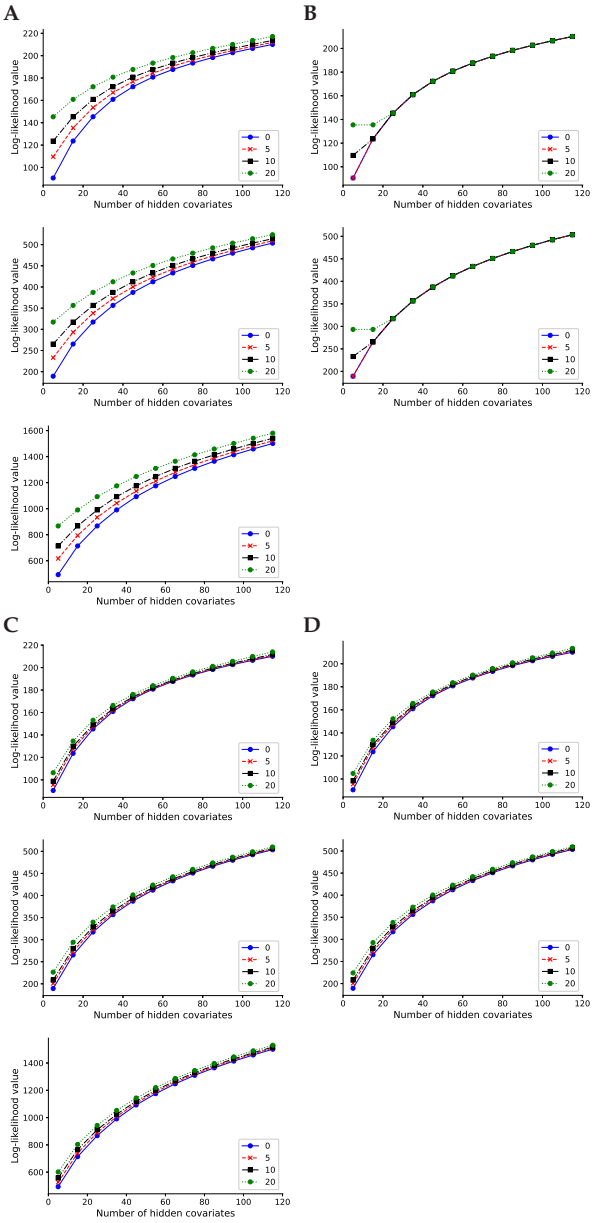


Figure S2: Log-likelihood values for LVREML (**A,C**) and PANAMA (**B,D**) using 0, 5, 10, and 20 PCs of the expression data (**A**,**B**) or genotype data (**C**,**D**) as known covariates, at sample sizes of 200, 400, and in the case of LVREML 1,012 segregants (top to bottom).

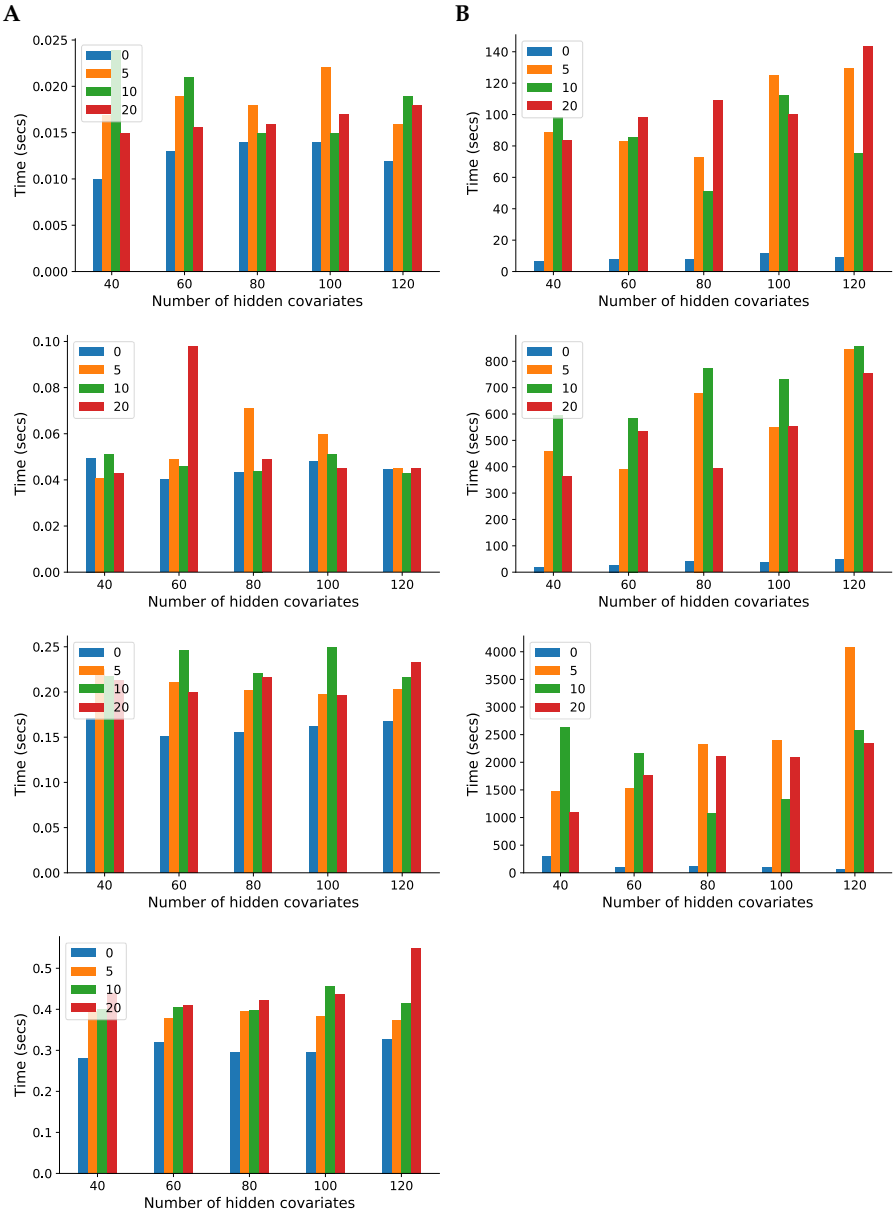


Figure S3: Runtime comparison on between LVREML (**A**) and PANAMA (**B**), with parameters set to infer 85 hidden covariates with 0, 5, 10, or 20 genotype PCs included as known covariates, at sample sizes of 200, 400, and in the case of LVREML 1,012 segregants (top to bottom).

Supplementary Methods

S1 Preliminary results

In the sections below, we will repeatedly use the following results. The first result concerns linear transformations of normally distributed variables and can be found in most textbooks on statistics or probability theory:

Lemma 1. *Let $x \in \mathbb{R}^n$ be a random, normally distributed vector,*

$$p(x) = \mathcal{N}(\mu, \Psi),$$

with $\mu \in \mathbb{R}^n$, and $\Psi \in \mathbb{R}^{n\times n}$ a positive definite covariance matrix. For any linear transformation $y = \mathbf{M}x$ with $\mathbf{M} \in \mathbb{R}^{m\times n}$, we have

$$p(y) = \mathcal{N}(\mathbf{M}\mu, \mathbf{M}\Psi\mathbf{M}^T).$$

□

If the linear transformation $y = \mathbf{M}x$ in this Lemma is overdetermined, that is, if $m > n$, then the transformed covariance matrix $\Psi' = \mathbf{M}\Psi\mathbf{M}^T$ will have a lower rank n than its dimension m, that is, $\Psi' \in \mathbb{R}^{m\times m}$ is a positive *semi*-definite matrix (i.e., has one or more zero eigenvalues). Thus we can extend the definition of normal distributions to include *degenerate* distributions with positive *semi*-definite covariance matrix, by interpreting them as the distributions of overdetermined linear combinations of normally distributed vectors. A degenerate one-dimensional normal distribution is simply defined as a δ-distribution, that is, for $x \in \mathbb{R}$

$$p(x) = \mathcal{N}(\mu, 0) = \delta(x - \mu),$$

which can be derived as a limit $\sigma^2 \to 0$ of normal distribution density functions $\mathcal{N}(\mu, \sigma^2)$.

The second result is one that is attributed to von Neumann [1]:

Lemma 2. *Let $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{n\times n}$ be two positive definite matrices. Then*

$$\mathrm{tr}(\mathbf{P}^{-1}\mathbf{Q}) \geq \sum_{i=1}^{n} \pi_i^{-1}\chi_i, \tag{S1}$$

where $\pi_1 \geq \cdots \geq \pi_n$ and $\chi_1 \geq \cdots \geq \chi_n$ are the ordered eigenvalues of $\mathbf{P}$ and $\mathbf{Q}$, respectively, and equality in eq. (S1) *is achieved if and only if the eigenvector of $\mathbf{P}$ corresponding to π_i is equal to the eigenvector of $\mathbf{Q}$ corresponding to χ_{n-i+1}, $i = 1, \ldots, n$.* □

S2 The model

We will use the following notation:

- $\mathbf{Y} \in \mathbb{R}^{n\times m}$ is a matrix of gene expression data for m genes in n samples. The ith column of $\mathbf{Y}$ is denoted $y_i \in \mathbb{R}^n$ and corresponds to the vector of expression values for gene i. We assume that the data in each sample are centred, $\sum_{i=1}^{m} y_i = 0 \in \mathbb{R}^n$.
- $\mathbf{Z} \in \mathbb{R}^{n\times d}$ is a matrix of values for d known confounders in the same n samples. The kth column of $\mathbf{Z}$ is denoted $z_k \in \mathbb{R}^n$ and corresponds to the data for confounding factor k.
- $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a matrix of values for p latent variables to be determined in the same n samples. The jth column of $\mathbf{X}$ is denoted $x_j \in \mathbb{R}^n$.

To identify the hidden correlation structure of the expression data, we assume a linear relationship between expression levels and the known and latent variables, with random noise added:

$$y_i = \mathbf{Z}v_i + \mathbf{X}w_i + \epsilon_i, \tag{S2}$$

where $v_i \in \mathbb{R}^d$ and $w_i \in \mathbb{R}^p$ are jointly normally distributed random vectors,

$$p\left(\begin{bmatrix} v_i \\ w_i \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} \mathbf{B} & \mathbf{D} \\ \mathbf{D}^T & \mathbf{A} \end{bmatrix}\right) \tag{S3}$$

with $\mathbf{B} \in \mathbb{R}^{d\times d}$, $\mathbf{D} \in \mathbb{R}^{d\times p}$ and $\mathbf{A} = \mathrm{diag}(\alpha_1^2, \ldots, \alpha_p^2)$, such that

$$\Psi = \begin{bmatrix} \mathbf{B} & \mathbf{D} \\ \mathbf{D}^T & \mathbf{A} \end{bmatrix}$$

is a positive semi-definite matrix; the errors $\epsilon_i \in \mathbb{R}^n$ are assumed to be independent and normally distributed,

$$p(\epsilon_i) = \mathcal{N}(0, \sigma^2 \mathbb{1}).$$

Note that our aim is to identify variance components shared across genes, and hence σ^2 is assumed to be the same for all i. By assumption, the errors are also independent of the effect sizes, and hence we can write

$$p\left(\begin{bmatrix} v_i \\ w_i \\ \epsilon_i \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} \mathbf{B} & \mathbf{D} & 0 \\ \mathbf{D}^T & \mathbf{A} & 0 \\ 0 & 0 & \sigma^2\mathbb{1} \end{bmatrix}\right). \tag{S4}$$

By Lemma 1, y_i is normally distributed with distribution

$$p(y_i) = \mathcal{N}(0, \mathbf{K}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\mathbf{K})}} \exp\left(-\frac{1}{2}\langle y_i, \mathbf{K}^{-1} y_i\rangle\right), \tag{S5}$$

where

$$\mathbf{K} = \begin{bmatrix} \mathbf{Z} & \mathbf{X} & \mathbb{1} \end{bmatrix} \begin{bmatrix} \mathbf{B} & \mathbf{D} & 0 \\ \mathbf{D}^T & \mathbf{A} & 0 \\ 0 & 0 & \sigma^2\mathbb{1} \end{bmatrix} \begin{bmatrix} \mathbf{Z}^T \\ \mathbf{X}^T \\ \mathbb{1} \end{bmatrix} = \mathbf{Z}\mathbf{B}\mathbf{Z}^T + \mathbf{Z}\mathbf{D}\mathbf{X}^T + \mathbf{X}\mathbf{D}^T\mathbf{Z} + \mathbf{X}\mathbf{A}\mathbf{X}^T + \sigma^2\mathbb{1},$$

and we used the notation $\langle u, v\rangle = u^T v$ to denote the inner product between two vectors in $\mathbb{R}^n$.

Defining matrices $\mathbf{V} \in \mathbb{R}^{d\times m}$ and $\mathbf{W} \in \mathbb{R}^{p\times m}$, whose columns are the random effect vectors v_i and w_i, respectively, eq. (S2) can be written in matrix notation as

$$\mathbf{Y} = \mathbf{ZV} + \mathbf{XW} + \boldsymbol{\epsilon}$$

Under the assumption that the columns y_i of $\mathbf{Y}$ are independent samples of the distribution (S5), the likelihood of observing $\mathbf{Y}$ given covariate data $\mathbf{Z}$, (unknown) latent variable data $\mathbf{X}$ and values for the hyper-parameters $\Theta = \{\sigma^2, \mathbf{A}, \mathbf{B}, \mathbf{D}\}$, is given by

$$p(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}, \Theta) = \prod_{i=1}^{m} p(y_i \mid 0, \mathbf{K}).$$

Note that in standard mixed-model calculations, the distribution (S5) is often arrived at by integrating out the random effects. This is equivalent to application of Lemma 1.

To conclude, the log-likelihood is, upto an additive constant, and divided by half the number of genes:

$$\mathcal{L} = -\frac{2}{m}\Big[\frac{m}{2}\log\det(\mathbf{K}) + \frac{1}{2}\sum_{i=1}^{m}\langle y_i, \mathbf{K}^{-1}y_i\rangle\Big] = -\log\det(\mathbf{K}) - \operatorname{tr}\big(\mathbf{K}^{-1}\mathbf{C}\big),$$

where

$$\mathbf{C} = \frac{\mathbf{YY}^T}{m}$$

is the empirical covariance matrix.

S3 Systematic effects on the mean

Eq. (S2) only considers random effects, which leads to a model for studying systematic effects on the covariance between samples. We could also include fixed effects to model systematic effects on mean expression level. However, by centering the data, $\sum_{i=1}^{m} y_i = 0$, the maximum-likelihood estimate of such fixed effects is always zero. To see this, let $\mathbf{T} \in \mathbb{R}^{n\times c}$ be a matrix of c covariates with fixed effects $\beta \in \mathbb{R}^c$ shared across genes (we are only interested in discovering systematic biases in the data). Then the minus log-likelihood (2) becomes

$$\mathcal{L} = \log\det(\mathbf{K}) + \frac{1}{m}\sum_{i=1}^{m}\langle y_i - \mathbf{T}\beta, \mathbf{K}^{-1}(y_i - \mathbf{T}\beta)\rangle$$

Optimizing with respect to β leads to the equation

$$\hat{\beta} = (\mathbf{T}^T\mathbf{K}^{-1}\mathbf{T})^{-1}\mathbf{T}^T\bar{y}$$

where

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i = 0.$$

S4 Solution of the model without latent variables

We start by considering the problem of finding the maximum-likelihood solution in the absence of any latent variables, i.e. minimizing eq. (2) with

$$\mathbf{K} = \mathbf{Z}\mathbf{B}\mathbf{Z}^T + \sigma^2\mathbb{1} \tag{S6}$$

with respect to $\mathbf{B}$ and σ^2.

Note first of all that we may assume the set of confounding factors $\{z_1, \ldots, z_d\}$ to be linearly independent, because if not, the expression in eq. (S2) can be rearranged in terms of a linearly independent subset of factors whose coefficients are still normally distributed due to elementary properties of the multivariate normal distribution, see for instance the proof of Lemma 5 below. Linear independence of $\{z_1, \ldots, z_d\}$ implies that we must have $d \leq n$ and $\mathrm{rank}(\mathbf{Z}) = d$.

The singular value decomposition allows to decompose $\mathbf{Z}$ as $\mathbf{Z} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{n\times n}$, $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbb{1}$, $\mathbf{\Gamma} \in \mathbb{R}^{n\times d}$ diagonal with $\gamma_k^2 \equiv \Gamma_{kk} > 0$ for $k \in \{1, \ldots, d\}$ [this uses $\mathrm{rank}(\mathbf{Z}) = d$], and $\mathbf{V} \in \mathbb{R}^{d\times d}$, $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbb{1}$. There is also a 'thin' SVD, $\mathbf{Z} = \mathbf{U}_1\mathbf{\Gamma}_1\mathbf{V}^T$, where $\mathbf{U}_1 \in \mathbb{R}^{n\times d}$, $\mathbf{U}_1^T\mathbf{U}_1 = \mathbb{1}$, $\mathbf{\Gamma}_1 \in \mathbb{R}^{d\times d}$ diagonal with diagonal elements γ_k^2. In block matrix notation, $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ and

$$\mathbf{Z} = \begin{pmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{pmatrix}\begin{pmatrix}\mathbf{\Gamma}_1 \\ 0\end{pmatrix}\mathbf{V} \tag{S7}$$

Note that unitarity of $\mathbf{U}$ implies $\mathbf{U}_1^T\mathbf{U}_2 = 0$.

Denote by $\mathcal{H}_Z$ the space spanned by the columns (i.e. covariate vectors) of $\mathbf{Z}$. The projection matrix $\mathbf{P}_Z$ onto $\mathcal{H}_Z$ is given by

$$\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T = \mathbf{U}_1\mathbf{\Gamma}_1\mathbf{V}^T(\mathbf{V}\mathbf{\Gamma}_1^{-2}\mathbf{V}^T)\mathbf{V}\mathbf{\Gamma}_1\mathbf{U}_1^T = \mathbf{U}_1\mathbf{U}_1^T.$$

Using the basis of column vectors of $\mathbf{U}$, we can write any matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ as a partitioned matrix

$$\mathbf{U}^T\mathbf{M}\mathbf{U} = \begin{pmatrix}\mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22}\end{pmatrix} \tag{S8}$$

where

$$\mathbf{M}_{ij} = \mathbf{U}_i^T\mathbf{M}\mathbf{U}_j. \tag{S9}$$

The following results for partitioned matrices are derived easily or can be found in [2]:

$$\mathrm{tr}(\mathbf{M}) = \mathrm{tr}(\mathbf{M}_{11}) + \mathrm{tr}(\mathbf{M}_{22}) \tag{S10}$$

$$\det(\mathbf{M}) = \det\big(\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}\big)\ \det(\mathbf{M}_{22}) \tag{S11}$$

Using this notation, the following result solves the model without latent variables:

Theorem 1. *Let* $\mathbf{C} \in \mathbb{R}^{n\times n}$ *be a positive definite matrix such that*

$$\lambda_{\min}(\mathbf{C}_{11}) > \frac{\operatorname{tr}(\mathbf{C}_{22})}{n-d}, \tag{S12}$$

where $\lambda_{\min}(\cdot)$ *denotes the smallest eigenvalue of a matrix. Then the maximum-likelihood solution*

$$\hat{\mathbf{K}} = \operatorname*{argmin}_{\{\mathbf{K}:\ \mathbf{K}=\mathbf{Z}\mathbf{B}\mathbf{Z}^T+\sigma^2\mathbb{1}\}} \log\det\mathbf{K} + \operatorname{tr}\big(\mathbf{K}^{-1}\mathbf{C}\big), \tag{S13}$$

subject to $\mathbf{B}$ *being positive semi-definite and* $\sigma^2 \geq 0$*, is given by*

$$\hat{\mathbf{B}} = \mathbf{V}\mathbf{\Gamma}_1^{-1}(\mathbf{C}_{11} - \hat{\sigma}^2\mathbb{1})\mathbf{\Gamma}_1^{-1}\mathbf{V}^T \tag{S14}$$

$$\hat{\sigma}^2 = \frac{\operatorname{tr}(\mathbf{C}_{22})}{n-d} \tag{S15}$$

Proof. Using eq. (S7), we can write

$$\begin{aligned}\mathbf{K} = \mathbf{Z}\mathbf{B}\mathbf{Z}^T + \sigma^2\mathbb{1} &= \mathbf{U}_1\mathbf{\Gamma}_1\mathbf{V}^T\mathbf{B}\mathbf{V}\mathbf{\Gamma}_1\mathbf{U}_1^T + \sigma^2(\mathbf{U}_1\mathbf{U}_1^T + \mathbf{U}_2\mathbf{U}_2^T)\\ &= \mathbf{U}_1\mathbf{\Gamma}_1\mathbf{V}^T\big(\mathbf{B} + \sigma^2\mathbf{V}\mathbf{\Gamma}_1^{-2}\mathbf{V}^T\big)\mathbf{V}\mathbf{\Gamma}_1\mathbf{U}_1^T + \sigma^2\mathbf{U}_2\mathbf{U}_2^T.\end{aligned}$$

Hence, in the block matrix notation (S8), we have

$$\begin{aligned}\mathbf{K}_{11} &= \mathbf{\Gamma}_1\mathbf{V}^T\big(\mathbf{B} + \sigma^2\mathbf{V}\mathbf{\Gamma}_1^{-2}\mathbf{V}^T\big)\mathbf{V}\mathbf{\Gamma}_1\\ \mathbf{K}_{22} &= \sigma^2\mathbb{1}\\ \mathbf{K}_{12} &= \mathbf{K}_{21} = 0.\end{aligned}$$

It follows that

$$\mathbf{K}^{-1} = \begin{pmatrix}\mathbf{K}_{11}^{-1} & 0\\ 0 & \mathbf{K}_{22}^{-1}\end{pmatrix}$$

and, using eqs. (S10) and (S11),

$$\begin{aligned}\log\det(\mathbf{K}) &= \log\det(\mathbf{K}_{11}) + \log\det(\mathbf{K}_{22}) = \log\det(\mathbf{K}_{11}) + (n-d)\log(\sigma^2)\\ \operatorname{tr}(\mathbf{K}^{-1}\mathbf{C}) &= \operatorname{tr}(\mathbf{K}_{11}^{-1}\mathbf{C}_{11}) + \operatorname{tr}(\mathbf{K}_{22}^{-1}\mathbf{C}_{22}) = \operatorname{tr}(\mathbf{K}_{11}^{-1}\mathbf{C}_{11}) + \frac{\operatorname{tr}(\mathbf{C}_{22})}{\sigma^2}.\end{aligned}$$

Let $\mathbf{C}_{11}$ have eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ with corresponding eigenvectors $u_1, \ldots, u_d \in \mathbb{R}^d$. Applying Lemma 2 to the term $\operatorname{tr}(\mathbf{K}_{11}^{-1}\mathbf{C}_{11})$, it follows that for the minimizer $\hat{\mathbf{K}}$, $\hat{\mathbf{K}}_{11}$ must have eigenvalues $\kappa_1 \geq \cdots \geq \kappa_d$ with the same eigevectors $u_1, \ldots, u_d$ as $\mathbf{C}_{11}$. Expressing the minus log-likelihood in terms of these eigenvalues results in

$$\mathcal{L}(\hat{\mathbf{K}}) = \sum_{i=1}^{d}\log(\kappa_i) + \sum_{i=1}^{d}\kappa_i^{-1}\lambda_i + (n-d)\log(\sigma^2) + \frac{\operatorname{tr}(\mathbf{C}_{22})}{\sigma^2}.$$

Minimizing with respect to the parameters κ_i and σ^2 (i.e., setting their derivatives to zero) results in the solution $\hat{\kappa}_i = \lambda_i$ for all i and $\hat{\sigma}^2 = \frac{\operatorname{tr}(\mathbf{C}_{22})}{n-d}$. In other words, $\hat{\mathbf{K}}_{11}$ has the same eigenvalues and eigenvectors as $\mathbf{C}_{11}$, that is,

$$\hat{\mathbf{K}}_{11} = \mathbf{C}_{11}.$$

This equation is satisfied if

$$\hat{\mathbf{B}} + \hat{\sigma}^2 \mathbf{V}\boldsymbol{\Gamma}_1^{-2}\mathbf{V}^T = \mathbf{V}\boldsymbol{\Gamma}_1^{-1}\mathbf{C}_{11}\boldsymbol{\Gamma}_1^{-1}\mathbf{V}^T$$

or

$$\hat{\mathbf{B}} = \mathbf{V}\boldsymbol{\Gamma}_1^{-1}(\mathbf{C}_{11} - \hat{\sigma}^2\mathbb{1})\boldsymbol{\Gamma}_1^{-1}\mathbf{V}^T$$

$\hat{\mathbf{B}}$ is positive semi-definite if and only if for all $v \in \mathbb{R}^d$

$$0 < \langle v, \hat{\mathbf{B}}v\rangle = \langle w, (\mathbf{C}_{11} - \hat{\sigma}^2\mathbb{1})w\rangle,$$

where $w = \boldsymbol{\Gamma}_1\mathbf{V}v$. Because $\mathbf{V}$ is unitary and $\boldsymbol{\Gamma}_1$ diagonal with strictly positive elements, $\langle v, \hat{\mathbf{B}}v\rangle > 0$ for all $v \in \mathbb{R}^d$ if and only if $\langle w, (\mathbf{C}_{11} - \hat{\sigma}^2\mathbb{1})w\rangle > 0$ for all $w \in \mathbb{R}^d$, or

$$0 < \min_{w\in\mathbb{R}^d} \frac{\langle w, \mathbf{C}_{11}w\rangle}{\langle w, w\rangle} - \hat{\sigma}^2 = \lambda_{\min}(\mathbf{C}_{11}) - \hat{\sigma}^2.$$

□

Eq. (S12) is a condition on the amount of variation in $\mathbf{Y}$ explained by the confounders $\mathbf{Z}$, with $\lambda_{\min}(\mathbf{C}_{11})$ being (proportional to) the minimum amount of variation explained by any of the dimensions spanned by the columns of $\mathbf{Z}$, and $\frac{1}{n-d}\operatorname{tr}(\mathbf{C}_{22})$ being the average amount of variation explained by the dimensions orthogonal to the columns of $\mathbf{Z}$. Failure of this condition simply means that there must be other, latent variables that explain more variation than the known ones, which is precisely what we are seeking to detect.

A useful special case of Theorem 1 occurs when the number of confounders equals one. In this case, we are seeking maximum-likelihood solutions for $\mathbf{K}$ of the form

$$\mathbf{K} = \beta^2 zz^T + \sigma^2\mathbb{1},$$

where $z \in \mathbb{R}^n$ is the confounding data vector. Let $\gamma^2 = \|z\|^2$ and $u = \frac{1}{\gamma}z$. Then $\mathbf{P}_z = uu^T$ is the projection matrix onto z, $\mathbf{C}_{11} = \langle u, \mathbf{C}u\rangle$, and $\operatorname{tr}(\mathbf{C}_{22}) = \operatorname{tr}((1 - \mathbf{P}_z)\mathbf{C}) = \operatorname{tr}(\mathbf{C}) - \langle u, \mathbf{C}u\rangle$. By Theorem 1, we have

$$\begin{aligned}
\hat{\beta}^2 &= \frac{1}{\gamma^2}\left\{\langle u, \mathbf{C}u\rangle - \frac{\operatorname{tr}\big([1 - \mathbf{P}_z]\mathbf{C}\big)}{n-1}\right\} \\
&= \frac{1}{\gamma^2}\left\{\frac{n}{n-1}\langle u, \mathbf{C}u\rangle - \frac{\operatorname{tr}(\mathbf{C})}{n-1}\right\} \qquad \text{(S16)} \\
\hat{\sigma}^2 &= \frac{\operatorname{tr}\big([1 - \mathbf{P}_z]\mathbf{C}\big)}{n-1} = \frac{\operatorname{tr}(\mathbf{C}) - \langle u, \mathbf{C}u\rangle}{n-1},
\end{aligned}$$

provided

$$\langle u, \mathbf{C}u\rangle > \frac{\operatorname{tr}(\mathbf{C})}{n}.$$

S5 Solution of the model without known covariates

Next, consider a model without known covariates, i.e. with posterior sample covariance matrix $\mathbf{K} = \mathbf{K}_X(\{\alpha_j, x_j\}) + \sigma^2\mathbb{1}$, where

$$\mathbf{K}_X(\{\alpha_j, x_j\}) = \sum_{j=1}^{p} \alpha_j^2 x_j x_j^T.$$

This model is equivalent to probabilistic principal component analysis [3,4], and its maximum-likelihood solution is given by the first p eigenvectors or principal components with largest eigenvalues of $\mathbf{C}$. Here we present a more direct proof of this fact than what can be found in the literature.

Lemma 3. *Without loss of generality, we may assume that the latent variables have unit norm, are linearly independent, and are mutually orthogonal.*

Proof. If the latent variables do not have unit norm, define $c_j = \|x_j\|^{-1}$, $\alpha_j' = \alpha_j/c_j$ and $x_j' = c_j x_j$ for all j. It follows immediately that $\|x_j'\| = 1$ and

$$\mathbf{K}_X(\{\alpha_j, x_j\}) = \mathbf{K}_X(\{\alpha_j', x_j'\}).$$

Next assume that the latent variables are not linearly independent, i.e. that $\operatorname{rank}(\mathbf{K}_X) = r < p$. Because $\mathbf{K}_X$ is a symmetric matrix, we must have $\mathbf{K}_X = \sum_{l=1}^{r} t_l t_l^T$ for some set of linearly independent vectors $t_l \in \mathbb{R}^n$. Define $\alpha_l' = \|t_l\|$ and $x_l' = t_l/\|t_l\|$. Then x_l' has unit norm and

$$\mathbf{K}_X(\{\alpha_l', x_l'\}) = \mathbf{K}_X(\{\alpha_j, x_j\}).$$

Finally, recall that

$$\mathbf{K}_X(\{\alpha_j, x_j\}) = \mathbf{X}\mathbf{A}\mathbf{X}^T = (\mathbf{X}\mathbf{A}^{\frac{1}{2}})(\mathbf{X}\mathbf{A}^{\frac{1}{2}})^T,$$

where $\mathbf{A} = \operatorname{diag}(\alpha_1^2, \ldots, \alpha_p^2)$. Because we may now assume that $\operatorname{rank}(\mathbf{X}) = p$, and because $\alpha_j > 0$ for all j, the matrix $\mathbf{X}\mathbf{A}^{\frac{1}{2}}$ has singular value decomposition

$$\mathbf{X}\mathbf{A}^{\frac{1}{2}} = \mathbf{U}\Xi\mathbf{V}^T$$

with $\mathbf{U} \in \mathbb{R}^{n\times p}$, $\mathbf{U}^T\mathbf{U} = \mathbb{1}$, $\Xi \in \mathbb{R}^{p\times p}$ diagonal with diagonal elements $\Xi_{jj} = \xi_j > 0$, and $\mathbf{V} \in \mathbb{R}^{p\times p}$, $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbb{1}$. Hence

$$\mathbf{K}_X(\{\alpha_j, x_j\}) = \mathbf{U}\Xi^2\mathbf{U}^T = \sum_{j=1}^{p} \xi_j^2 u_j u_j^T = \mathbf{K}_X(\{\xi_j, u_j\}),$$

with u_j the orthonormal columns of $\mathbf{U}$, $\langle u_j, u_j'\rangle = (\mathbf{U}^T\mathbf{U})_{jj'} = \delta_{j,j'}$. □

We will also need the following simple result:

Lemma 4. *Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ be a decreasing sequence of positive numbers, and let $1 \leq p < n$. If there exists $j > p$ such that $\lambda_p > \lambda_j$, then*

$$\lambda_p > \frac{1}{n-p} \sum_{j=p+1}^{n} \lambda_j. \tag{S17}$$

Proof. Eq. (S17) follows from

$$\lambda_p - \frac{1}{n-p}\sum_{j=p+1}^{n}\lambda_j = \frac{1}{n-p}\sum_{j=p+1}^{n}(\lambda_p - \lambda_j) > 0,$$

because each term on the r.h.s. is non-negative, and at least one is strictly positive. □

Theorem 2. *Let $\mathbf{C} \in \mathbb{R}^{n\times n}$ be a positive definite matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and corresponding eigenvectors $u_1, \ldots, u_n$, and let either $p = n$ or $1 \leq p < n$ such that there exists $j > p$ with $\lambda_p > \lambda_j$. Then the maximum-likelihood solution*

$$\hat{\mathbf{K}} = \operatorname*{argmin}_{\{\mathbf{K}:\ \mathbf{K}=\mathbf{X}\mathbf{A}\mathbf{X}^T+\sigma^2\mathbb{1}\}} \log\det\mathbf{K} + \operatorname{tr}\big(\mathbf{K}^{-1}\mathbf{C}\big),$$

is given by

$$\begin{aligned}\hat{x}_j &= u_j\\ \hat{\alpha}_j^2 &= \lambda_j - \hat{\sigma}^2\\ \hat{\sigma}^2 &= \frac{1}{n-p}\sum_{j=p+1}^{n}\lambda_j.\end{aligned}$$

Proof. By Lemma 3, we can assume that $\mathbf{X}$ has orthonormal columns, and hence there exist $\mathbf{V} \in \mathbb{R}^{n\times(n-p)}$ such that $\mathbf{Q} = (\mathbf{X}, \mathbf{V}) \in \mathbb{R}^{n\times n}$ is unitary, $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbb{1}$. Hence $\mathbf{K} = \mathbf{X}\mathbf{A}\mathbf{X}^T + \sigma^2\mathbb{1}$ has the spectral decomposition

$$\mathbf{K} = \begin{pmatrix}\mathbf{X} & \mathbf{V}\end{pmatrix}\begin{pmatrix}\mathbf{A}^2+\sigma^2\mathbb{1} & 0\\ 0 & \sigma^2\mathbb{1}\end{pmatrix}\begin{pmatrix}\mathbf{X}^T\\ \mathbf{V}^T\end{pmatrix},$$

and hence

$$\mathbf{K}^{-1} = \sum_{j=1}^{p}\frac{1}{\alpha_j^2+\sigma^2}x_j x_j^T + \frac{1}{\sigma^2}\sum_{l=1}^{n-p}v_l v_l^T,$$

where $v_l \in \mathbb{R}^n$ are the columns of $\mathbf{V}$.

Assume that the α_j^2 are ordered, $\alpha_1^2 \geq \cdots \geq \alpha_p^2$. Applying von Neumann's Lemma 2 gives

$$\begin{aligned}\mathcal{L} &= \log\det(\mathbf{K}) + \operatorname{tr}\big(\mathbf{K}^{-1}\mathbf{C}\big)\\ &\geq \sum_{j=1}^{p}\log(\alpha_j^2+\sigma^2) + (n-p)\log(\sigma^2) + \sum_{j=1}^{p}\frac{\lambda_j}{\alpha_j^2+\sigma^2} + \sum_{j=p+1}^{n}\frac{\lambda_j}{\sigma^2},\end{aligned} \tag{S18}$$

with equality if and only if

$$\begin{aligned}x_j &= u_j \text{ for } j = 1,\ldots,p\\ v_l &= u_{p+l} \text{ for } l = 1,\ldots,n-p\end{aligned}$$

Hence, independent of the values for α_j, the maximum-likelihood latent variables are the eigenvectors of $\mathbf{C}$ corresponding to the p largest eigenvalues. Minimizing eq, (S18) w.r.t. α_j^2

and σ^2 then gives

$$\alpha_j^2 = \lambda_j - \sigma^2$$
$$\sigma^2 = \frac{1}{n-p} \sum_{j=p+1}^{N} \lambda_j.$$

By Lemma 4, $\alpha_j^2 > 0$ for all j. □

Note that plugging the maximum-likelihood values in the likelihood function gives

$$\mathcal{L}_{\min} = \sum_{j=1}^{p} \log(\lambda_j) + (n-p) \log\Big(\frac{1}{n-p} \sum_{j=p+1}^{n} \lambda_j\Big) + n \tag{S19}$$

Either p can be set *a priori* small enough such that condition (S17) is satisfied, or else the value of p with smallest $\mathcal{L}_{\min}$ satisfying this condition can be found easily from eq. (S19).

Note also that in the models of [3,4], uniform prior variances are assumed ($\alpha_1^2 = \cdots = \alpha_p^2 = 1$), such that $\mathbf{X}$ is defined upto an arbitrary rotation, because $\mathbf{X}\mathbf{X}^T = (\mathbf{X}\mathbf{R})(\mathbf{X}\mathbf{R})^T$ for any rotation matrix $\mathbf{R}$. In our model, there is no such rotational freedom (if $\mathbf{A}$ is assumed to be diagonal), except if $\mathbf{C}$ has eigenvalues with multiplicities greater than one, when there is some freedom to choose the corresponding eigenvectors.

S6 Solution of the full model

S6.1 Orthogonality of known and hidden confounders

Lemma 5. *Without loss of generality, we may assume that the latent variables are orthogonal to the known confounders:*

$$\mathbf{X}^T\mathbf{Z} = \mathbf{Z}\mathbf{X}^T = 0. \tag{S20}$$

Proof. As in Section S4, let $\mathbf{P}_Z$ again be the projection matrix on the space spanned by the known covariates z_k (i.e. the columns of $\mathbf{Z}$). For any choice of latent variables x_j, we have

$$x_j = \mathbf{P}_Z x_j + (1 - \mathbf{P}_Z)x_j = \sum_{k=1}^{d} m_{kj} z_k + \tilde{x}_j,$$

for some matrix of linear coefficients $\mathbf{M} = (m_{kq}) \in \mathbb{R}^{d \times p}$, and with $\langle s_k, \tilde{x}_j \rangle = 0$ for all k. Or, in matrix notation

$$\mathbf{X} = \mathbf{Z}\mathbf{M} + \tilde{\mathbf{X}}, \quad \text{with} \quad \tilde{\mathbf{X}}^T\mathbf{Z} = \mathbf{Z}^T\tilde{\mathbf{X}} = 0$$

Plugging this in eq. (S2), results in

$$y_i = \mathbf{Z}\tilde{v}_i + \tilde{\mathbf{X}}w_i + \epsilon_i \tag{S21}$$

where $\tilde{v}_i = v_i + \mathbf{M}w_i$. Hence

$$\begin{bmatrix} \tilde{v}_i \\ w_i \\ \epsilon_i \end{bmatrix} = \begin{bmatrix} \mathbb{1} & M & 0 \\ 0 & \mathbb{1} & 0 \\ 0 & 0 & \mathbb{1} \end{bmatrix} \begin{bmatrix} v_i \\ w_i \\ \epsilon_i \end{bmatrix}$$

and hence, using Lemma 1, it follows that

$$p\left(\begin{bmatrix} \tilde{v}_i \\ w_i \\ \epsilon_i \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} \mathbf{B} + \mathbf{M}\mathbf{D}^T + \mathbf{D}\mathbf{M}^T + \mathbf{M}\mathbf{A}\mathbf{M}^T & \mathbf{D} + \mathbf{A}\mathbf{M}^T & 0 \\ \mathbf{D}^T + \mathbf{M}\mathbf{A} & \mathbf{A} & 0 \\ 0 & 0 & \sigma^2\mathbb{1} \end{bmatrix}\right).$$

This is still of exactly the same form as eq. (S4). Hence model (S21) is identical to model (S2), but has hidden covariates orthogonal to the known covariates. □

Note that we can parameterize the model with hidden variables orthogonal to the known confounders, $\mathbf{Z}^T\mathbf{X} = 0$, but only if we allow the covariances of their effects on gene expression, $\mathrm{Cov}(v_i, w_i) = \mathbf{D}$, to be non-zero. Equivalently, we can parameterize the model such that the random effects of hidden variables are statistically independent of the effects of the known confounders, $\mathrm{Cov}(v_i, w_i) = 0$, but only if we allow the hidden variables to overlap with the known confounders, $\mathbf{Z}^T\mathbf{X} \neq 0$. Mathematically, the choice of orthogonal hidden factors will be much more convenient.

Note also that a transformation to orthogonal hidden factors always induces non-zero covariances among the known confounders via the term $\mathbf{M}\mathbf{A}\mathbf{M}^T$. Hence an important difficulty with the model where $\mathbf{B}$ is assumed to be diagonal, as used in [5], comes from the fact that non-orthogonal hidden variables are needed to model off-diagonal covariances between the known confounders. It is much more intuitive to model these directly by assuming a general covariance matrix.

S6.2 Restricted maximum-likelihood solution for the latent variables

Lemma 6. *Without loss of generality, we may assume that the latent variables have unit norm, are linearly independent, and are mutually orthogonal.*

Proof. The proof is identical to the proof of Lemma 3 – it is straightforward to verify that the transformation to orthonormal variables also do not change the form of the off-diagonal term $\mathbf{Z}\mathbf{D}\mathbf{X}^T$ in the covariance matrix $\mathbf{K}$, but merely lead to a reparameterization of the matrix $\mathbf{D}$. □

To solve the full model, we follow an approach similar to the standard restricted maximum-lilelihood method for linear mixed models [6,7]: we write the negative log-likelihood function $\mathcal{L} = \log\det(\mathbf{K}) + \mathrm{tr}(\mathbf{K}^{-1}\mathbf{C})$ as a sum

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2, \tag{S22}$$

where $\mathcal{L}_2$ will be the log-likelihood restricted to the subspace orthogonal to the known confounders $\mathbf{Z}$. We will estimate the latent variables $\mathbf{X}$ and their effect covariances $\mathbf{A}$ by maximizing $\mathcal{L}_2$, and estimate the effect covariances $\mathbf{B}$ and $\mathbf{D}$ involving the known confounders

by maximizing $\mathcal{L}_1$. Solving for the latent variables on a restricted subspace is motivated by the observation that if $y \in \mathbb{R}^n$ is a sample from the model (S2), that is, $p(y) = \mathcal{N}(0, \mathbf{K})$, then

$$\mathbf{U}_2\mathbf{U}_2^T y = \mathbf{U}_2\mathbf{U}_2^T\mathbf{Z}v + \mathbf{U}_2\mathbf{U}_2^T\mathbf{X}w + \mathbf{U}_2\mathbf{U}_2^T\epsilon = \mathbf{X}w + \epsilon'.$$

In other words, restricted to the subspace orthogonal to $\mathbf{Z}$, the general model becomes a probablistic PCA model where all variation in the data is explained by the latent variables.

To obtain the decomposition (S22), we partition $y \in \mathbb{R}^n$ as $y = (y_1, y_2)^T$, where $y_1 = \mathbf{U}_1^T y \in \mathbb{R}^d$ and $y_2 = \mathbf{U}_2^T y \in \mathbb{R}^{n-d}$, and write

$$p(y) = p(y_1, y_2) = p(y_1 \mid y_2)p(y_2),$$

or

$$\log p(y) = \log p(y_1, y_2) = \log p(y_1 \mid y_2) + \log p(y_2).$$

Hence

$$\mathcal{L} = -\frac{2}{m}\sum_{i=1}^{m} \log p(y_i) = \underbrace{-\frac{2}{m}\sum_{i=1}^{m} \log p(y_{i1} \mid y_{i2})}_{\mathcal{L}_1} \underbrace{-\frac{2}{m}\sum_{i=1}^{m} \log p(y_{i2})}_{\mathcal{L}_2}$$

Using standard results for the marginal and conditional distributions of a multivariate Gaussian, we have

$$\begin{aligned} p(y_2) &= \mathcal{N}(0, \mathbf{K}_{22}) \\ p(y_1 \mid y_2) &= \mathcal{N}\big(\mathbf{K}_{12}\mathbf{K}_{22}^{-1}y_2, (\mathbf{K}_{11} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\mathbf{K}_{21})\big), \end{aligned}$$

where we used the partitioned matrix notation of eq. (S8). In particular,

$$\begin{aligned} \mathcal{L}_2 &= \log\det(\mathbf{K}_{22}) + \frac{1}{m}\sum_{i=1}^{m} \langle \mathbf{U}_2^T y_i, \mathbf{K}_{22}^{-1}\mathbf{U}_2^T y_i \rangle \\ &= \log\det(\mathbf{K}_{22}) + \frac{1}{m}\sum_{i=1}^{m} \operatorname{tr}\big(\mathbf{K}_{22}^{-1}\mathbf{U}_2^T y_i y_i^T \mathbf{U}_2\big) \\ &= \log\det(\mathbf{K}_{22}) + \operatorname{tr}\big(\mathbf{K}_{22}^{-1}\mathbf{C}_{22}\big). \end{aligned}$$

Note that $\mathbf{K}_{22} = \mathbf{U}_2^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{U}_2^T + \sigma^2\mathbb{1}$, and hence $\mathcal{L}_2$ depends only on $\mathbf{X}$, $\mathbf{A}$ and σ^2. The restricted maximum likelihood solution for the latent variables follows immediately:

Theorem 3. *Let $\hat{\mathbf{X}} \in \mathbb{R}^{n\times p}$, $\hat{A} \in \mathbb{R}^{d\times d}$, and $\hat{\sigma}^2$ be the solution of*

$$\{\hat{\mathbf{X}}, \hat{\mathbf{A}}, \hat{\sigma}^2\} = \operatorname*{argmin}_{\mathbf{X},\mathbf{A},\sigma^2} \mathcal{L}_2(\mathbf{X}, \mathbf{A}, \sigma^2),$$

where the minimum is taken over all $\mathbf{X}$ with $\mathbf{X}^T\mathbf{Z} = 0$, and all positive semi-definite diagonal matrices $\hat{\mathbf{A}}$. If there exists $j > p$ such that $\lambda_p > \lambda_j$, then

$$\hat{\mathbf{X}} = \mathbf{U}_2\mathbf{W}_p \tag{S23}$$

$$\hat{\mathbf{A}} = \operatorname{diag}(\lambda_1 - \hat{\sigma}^2, \ldots, \lambda_p - \hat{\sigma}^2) \tag{S24}$$

$$\hat{\sigma}^2 = \frac{1}{n-d-p}\sum_{j=p+1}^{n-d} \lambda_j \tag{S25}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-d}$ *are the sorted eigenvalues of* $\mathbf{C}_{22}$ *with corresponding eigenvectors* $w_1, \ldots, w_{n-d} \in \mathbb{R}^{n-d}$*, and* $\mathbf{W}_p = (w_1, \ldots, w_p) \in \mathbb{R}^{(n-d)\times p}$ *is the matrix with the first p eigenvectors of* $\mathbf{C}_{22}$ *as columns.*

Proof. Defining $\tilde{\mathbf{X}} = \mathbf{U}_2^T\mathbf{X} \in \mathbb{R}^{(n-d)\times p}$, we have $\mathbf{K}_{22} = \tilde{\mathbf{X}}\mathbf{A}\tilde{\mathbf{X}}^T + \sigma^2\mathbb{1}$, and $\mathcal{L}_2$ becomes precisely the minus log-likelihood of the model without known covariates (Section S5), as a function of the latent variables $\tilde{\mathbf{X}}$ on the *reduced* $(n-d)$-dimensional space orthogonal to the known confounders $\mathbf{Z}$. Hence by Theorem 2,

$$\begin{aligned}\hat{\tilde{\mathbf{X}}} &= \mathbf{W}_p \\ \hat{\mathbf{A}} &= \operatorname{diag}(\lambda_1 - \sigma^2, \ldots, \lambda_p - \sigma^2),\end{aligned}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-d}$ are the sorted eigenvalues of $\mathbf{C}_{22}$ and $\mathbf{W}_p \in \mathbb{R}^{(n-d)\times p}$ is the matrix having the corresponding first p eigenvectors as columns. Note that $\hat{\mathbf{A}}$ is positive semi-definite by Lemma 4 and the assumption that there exists $j > p$ such that $\lambda_p > \lambda_j$. It remains to 'pull-back' $\tilde{\mathbf{X}}$ to the original n-dimensional space, using the orthogonality condition (S20):

$$\hat{\mathbf{X}} = (\mathbf{U}_1\mathbf{U}_1^T + \mathbf{U}_2\mathbf{U}_2^T)\hat{\mathbf{X}} = \mathbf{U}_2\mathbf{U}_2^T\hat{\mathbf{X}} = \mathbf{U}_2\hat{\tilde{\mathbf{X}}} = \mathbf{U}_2\mathbf{W}_p.$$

This proves eqs. (S23) and (S24). □

S6.3 Solution for the variance parameters given the latent variables

With $\hat{\mathbf{X}}$, $\hat{\mathbf{A}}$ and $\hat{\sigma}^2$ determined by the minimization of $\mathcal{L}_2$ in Theorem 3, $\mathcal{L}_2(\hat{\mathbf{X}}, \hat{\mathbf{A}}, \hat{\sigma}^2)$ is constant in terms of the parameters $\mathbf{B}$ and $\mathbf{D}$ that remain to be optimized. Hence optimizing $\mathcal{L}_1$ with respect to these parameters is the same as optimizing the total negative log-likelihood $\mathcal{L}(\hat{\mathbf{X}}, \hat{\mathbf{A}}, \mathbf{B}, \mathbf{D}, \hat{\sigma}^2)$ w.r.t. $\mathbf{B}$ and $\mathbf{D}$. We have:

Theorem 4. *Let* $\hat{\mathbf{B}} \in \mathbb{R}^{d\times d}$ *and* $\hat{D} \in \mathbb{R}^{d\times(n-d)}$ *be the solution of*

$$\{\hat{\mathbf{B}}, \hat{\mathbf{D}}\} = \operatorname*{argmin}_{\mathbf{B},\mathbf{D}} \mathcal{L}_1(\hat{\mathbf{X}}, \hat{\mathbf{A}}, \mathbf{B}, \mathbf{D}, \hat{\sigma}^2) = \operatorname*{argmin}_{\mathbf{B},\mathbf{D}} \mathcal{L}(\hat{\mathbf{X}}, \hat{\mathbf{A}}, \mathbf{B}, \mathbf{D}, \hat{\sigma}^2),$$

subject to the constraint that $\mathbf{B}$ *and* $\mathbf{B} - \mathbf{D}\hat{\mathbf{A}}^{-1}\mathbf{D}^T$ *are positive semi-definite. If*

$$\lambda_{\min}(C_{11}) > \hat{\sigma}^2, \tag{S26}$$

then

$$\hat{\mathbf{B}} = \mathbf{V}\mathbf{\Gamma}_1^{-1}(\mathbf{C}_{11} - \hat{\sigma}^2\mathbb{1})\mathbf{\Gamma}_1^{-1}\mathbf{V}^T \tag{S27}$$

$$\hat{\mathbf{D}} = \mathbf{V}\mathbf{\Gamma}_1^{-1}\mathbf{C}_{12}\mathbf{W}_p \tag{S28}$$

where as before

$$\mathbf{Z} = \begin{pmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{pmatrix}\begin{pmatrix}\mathbf{\Gamma}_1 \\ 0\end{pmatrix}\mathbf{V}^T$$

is the singular value decomposition of $\mathbf{Z}$*, and* $\mathbf{W}_p = (w_1, \ldots, w_p) \in \mathbb{R}^{(n-d)\times p}$ *is the matrix with the first p eigenvectors of* $\mathbf{C}_{22}$ *as columns.*

Proof. Note that the conditions $\mathbf{B}$ and $\mathbf{B} - \mathbf{D}\hat{\mathbf{A}}^{-1}\mathbf{D}^T$ positive semi-definite are to ensure that the matrix $\begin{pmatrix} \mathbf{B} & \mathbf{D} \\ \mathbf{D}^T & \hat{\mathbf{A}} \end{pmatrix}$ is positive semi-definite. Next note that with $\hat{\mathbf{X}}^T$ known, the covariance matrix $\mathbf{K}$ can be written as

$$\mathbf{K} = \begin{pmatrix} \mathbf{Z} & \hat{\mathbf{X}} \end{pmatrix} \begin{pmatrix} \mathbf{B} & \mathbf{D} \\ \mathbf{D}^T & \hat{\mathbf{A}} \end{pmatrix} \begin{pmatrix} \mathbf{Z}^T \\ \hat{\mathbf{X}}^T \end{pmatrix} + \hat{\sigma}^2 \mathbb{1}$$

Hence the total log-likelihood is identical to the model with known covariates $\tilde{\mathbf{Z}} = \begin{pmatrix} \mathbf{Z} & \hat{\mathbf{X}} \end{pmatrix}$ and no latent variables (Section S4). The *unconstrained* maximizing solution (that is, where $\mathbf{A}$ and σ^2 are also optimized) for the model with known covariates $\tilde{\mathbf{Z}}$ is given by Theorem 1. Due to $\hat{\mathbf{X}}^T\mathbf{Z} = 0$ and the definition of $\hat{\mathbf{X}}$, the singular value decomposition of $\tilde{\mathbf{Z}}$ is given by

$$\tilde{\mathbf{Z}} = \begin{pmatrix} \mathbf{U}_1 & \hat{\mathbf{X}} & \mathbf{U}_3 \end{pmatrix} \begin{pmatrix} \mathbf{\Gamma}_1 & 0 \\ 0 & \mathbb{1} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{V}^T & 0 \\ 0 & \mathbb{1} \end{pmatrix},$$

where the columns of $\mathbf{U}_3 \in \mathbb{R}^{n \times (n-d-p)}$ span the space orthogonal to the columns of $\tilde{\mathbf{Z}}$. Hence the unconstrained solution, can be written as (cf. eqs. (S14)–(S15))

$$\begin{pmatrix} \hat{\mathbf{B}} & \hat{\mathbf{D}} \\ \hat{\mathbf{D}}^T & \hat{\mathbf{A}}' \end{pmatrix} = \begin{pmatrix} \mathbf{V} & 0 \\ 0 & \mathbb{1} \end{pmatrix} \begin{pmatrix} \mathbf{\Gamma}_1^{-1} & 0 \\ 0 & \mathbb{1} \end{pmatrix} \begin{pmatrix} \mathbf{U}_1^T \\ \hat{\mathbf{X}}^T \end{pmatrix} (\mathbf{C} - \hat{\sigma}'^2 \mathbb{1}) \begin{pmatrix} \mathbf{U}_1 & \hat{\mathbf{X}} \end{pmatrix} \begin{pmatrix} \mathbf{\Gamma}_1^{-1} & 0 \\ 0 & \mathbb{1} \end{pmatrix} \begin{pmatrix} \mathbf{V}^T & 0 \\ 0 & \mathbb{1} \end{pmatrix}$$
$$\hat{\sigma}'^2 = \frac{\mathrm{tr}(\mathbf{U}_3^T \mathbf{C} \mathbf{U}_3)}{n - d - p}$$

First note that $\hat{\sigma}'^2 = \hat{\sigma}^2$, because we can write $\mathbf{U}_3 = \mathbf{U}_2 \mathbf{W}_{\sim p}$, where $\mathbf{W}_{\sim p} \in \mathbb{R}^{(n-d)\times(n-d-p)}$ is the matrix with the $n - d - p$ last eigenvectors of $\mathbf{C}_{22}$.

Working out the block matrix product results in:

$$\begin{aligned}
\hat{\mathbf{B}} &= \mathbf{V}\mathbf{\Gamma}_1^{-1}\mathbf{U}_1^T(\mathbf{C} - \hat{\sigma}^2\mathbb{1})\mathbf{U}_1\mathbf{\Gamma}_1^{-1}\mathbf{V}^T = \mathbf{V}\mathbf{\Gamma}_1^{-1}(\mathbf{C}_{11} - \hat{\sigma}^2\mathbb{1})\mathbf{\Gamma}_1^{-1}\mathbf{V}^T \\
\hat{\mathbf{D}} &= \mathbf{V}\mathbf{\Gamma}_1^{-1}\mathbf{U}_1^T\mathbf{C}\hat{\mathbf{X}} = \mathbf{V}\mathbf{\Gamma}_1^{-1}\mathbf{U}_1^T\mathbf{C}\mathbf{U}_2\mathbf{W}_p = \mathbf{V}\mathbf{\Gamma}_1^{-1}\mathbf{C}_{12}\mathbf{W}_p \\
\hat{\mathbf{A}}' &= \hat{\mathbf{X}}^T(\mathbf{C} - \hat{\sigma}^2\mathbb{1})\hat{\mathbf{X}} = \mathbf{W}_p^T\mathbf{U}_2^T(\mathbf{C} - \hat{\sigma}^2\mathbb{1})\mathbf{U}_2\mathbf{W}_p = \mathbf{W}_p^T(\mathbf{C}_{22} - \hat{\sigma}^2\mathbb{1})\mathbf{W}_p \\
&= \mathrm{diag}(\lambda_1 - \hat{\sigma}^2, \ldots, \lambda_p - \hat{\sigma}^2)
\end{aligned}$$

Hence, also the estimate $\hat{\mathbf{A}}' = \hat{\mathbf{A}}$. Because the unconstrained optimization of $\mathcal{L}$ given $\hat{\mathbf{X}}$ results in the same estimate for $\mathbf{A}$ and σ^2 as the intial constrained optimization where these parameters were given, it follows that also the estimates of $\mathbf{B}$ and $\mathbf{D}$ must be the same:

$$\{\hat{\mathbf{B}}, \hat{\mathbf{D}}\} = \underset{\mathbf{B},\mathbf{D}}{\mathrm{argmin}}\, \mathcal{L}(\mathbf{B}, \mathbf{D} \mid \hat{\mathbf{X}}, \hat{\mathbf{A}}, \hat{\sigma}^2) = \underset{\mathbf{B},\mathbf{D}}{\mathrm{argmin}} \min_{\mathbf{A},\sigma^2} \mathcal{L}(\mathbf{A}, \mathbf{B}, \mathbf{D}, \sigma^2 \mid \hat{\mathbf{X}}).$$

□

S6.4 LVREML maximizes the variance explained

It is tempting to ask whether the combined solution from Theorems 3 and 4 optimizes the *total* likelihood among all possible p-dimensional sets of latent variables. To address this

problem, let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be an arbitrary matrix of latent variables whose columns are normalized, mutually orthogonal and orthogonal to the columns of $\mathbf{Z}$, $\mathbf{X}^T\mathbf{X} = \mathbb{1}$ and $\mathbf{X}^T\mathbf{Z} = 0$. Because $\mathbf{U}_2$ is only defined upto a rotation, we can always choose

$$\mathbf{U}_2 = \begin{pmatrix} \mathbf{X} & \mathbf{Q} \end{pmatrix}$$

with $\mathbf{Q} \in \mathbb{R}^{n \times (n-d-p)}$ satisfying $\mathbf{Q}^T\mathbf{Q} = \mathbb{1}$, $\mathbf{Q}^T\mathbf{X} = 0$ and $\mathbf{Q}^T\mathbf{Z} = 0$. From the proof of Theorem 4 we immediately obtain:

Proposition 1. *Let* $\mathbf{A}(\mathbf{X}) \in \mathbb{R}^{p \times p}$, $\mathbf{B}(\mathbf{X}) \in \mathbb{R}^{d \times d}$, $\mathbf{D}(\mathbf{X}) \in \mathbb{R}^{d \times (n-d)}$ *and* $\sigma^2(\mathbf{X}) > 0$ *be the solution of*

$$\{\mathbf{A}(\mathbf{X}), \mathbf{B}(\mathbf{X}), \mathbf{D}(\mathbf{X}), \sigma^2(\mathbf{X})\} = \underset{\mathbf{A},\mathbf{B},\mathbf{D},\sigma^2}{\operatorname{argmin}} \mathcal{L}(\mathbf{A}, \mathbf{B}, \mathbf{D}, \sigma^2 \mid \mathbf{X}).$$

Then

$$\begin{aligned} \mathbf{B}(\mathbf{X}) &= \mathbf{V}\mathbf{\Gamma}_1^{-1}(\mathbf{C}_{11} - \hat{\sigma}^2\mathbb{1})\mathbf{\Gamma}_1^{-1}\mathbf{V}^T \\ \mathbf{D}(\mathbf{X}) &= \mathbf{V}\mathbf{\Gamma}_1^{-1}\mathbf{U}_1^T\mathbf{C}\mathbf{X} \\ \mathbf{A}(\mathbf{X}) &= \mathbf{X}^T(\mathbf{C} - \hat{\sigma}^2\mathbb{1})\mathbf{X} \\ \sigma^2(\mathbf{X}) &= \frac{\operatorname{tr}(\mathbf{Q}^T\mathbf{C}\mathbf{Q})}{n-d-p} \end{aligned}$$

□

Plugging these values into the negative log-likelihood function results in a function that depends only on $\mathbf{X}$:

Proposition 2. *Let* $\mathbf{X} \in \mathbb{R}^{n \times p}$ *be an arbitrary choice of latent variables with associated maximum-likelihood estimates for the covariance parameters given by Proposition 1. Then, upto an additive constant*

$$\mathcal{L}_{\mathbf{X}} = \log\det\Big(\mathbf{X}^T\big[\mathbf{C} - \mathbf{C}\mathbf{U}_1(\mathbf{U}_1^T\mathbf{C}\mathbf{U}_1)\mathbf{U}_1^T\mathbf{C}\big]\mathbf{X}\Big) + (n-d-p)\log\big(\hat{\sigma}^2(\mathbf{X})\big) \tag{S29}$$

Proof. Recall from Theorem 2 that the maximum-likelihood estimate for $\mathbf{K}$ given $\mathbf{X}$ and its associated maximum-likelihood parameters estimates is given by

$$\hat{\mathbf{K}}(\mathbf{X}) = \begin{pmatrix} \mathbf{U}_1^T\mathbf{C}\mathbf{U}_1 & \mathbf{U}_1^T\mathbf{C}\mathbf{X} & 0 \\ \mathbf{X}^T\mathbf{C}\mathbf{U}_1 & \mathbf{X}^T\mathbf{C}\mathbf{X} & 0 \\ 0 & 0 & \hat{\sigma}^2\mathbb{1} \end{pmatrix}$$

while the covariance matrix $\mathbf{C}$ can be written as

$$\mathbf{C} = \begin{pmatrix} \mathbf{U}_1^T\mathbf{C}\mathbf{U}_1 & \mathbf{U}_1^T\mathbf{C}\mathbf{X} & \mathbf{U}_1^T\mathbf{C}\mathbf{Q} \\ \mathbf{X}^T\mathbf{C}\mathbf{U}_1 & \mathbf{X}^T\mathbf{C}\mathbf{X} & \mathbf{X}^T\mathbf{C}\mathbf{Q} \\ \mathbf{Q}\mathbf{C}\mathbf{U}_1 & \mathbf{Q}\mathbf{C}\mathbf{X} & \mathbf{Q}^T\mathbf{C}\mathbf{Q} \end{pmatrix}$$

Hence

$$\begin{aligned}\mathcal{L}_{\mathbf{X}} = \mathcal{L}\big(\hat{\mathbf{K}}(\mathbf{X})\big) &= \log\det\big(\hat{\mathbf{K}}(\mathbf{X})\big) + \operatorname{tr}\big(\hat{\mathbf{K}}(\mathbf{X})^{-1}\mathbf{C}\big) \\ &= \log\det\begin{pmatrix}\mathbf{U}_1^T\mathbf{C}\mathbf{U}_1 & \mathbf{U}_1^T\mathbf{C}\mathbf{X} \\ \mathbf{X}^T\mathbf{C}\mathbf{U}_1 & \mathbf{X}^T\mathbf{C}\mathbf{X}\end{pmatrix} + (n-d-p)\log(\hat{\sigma}^2) + (d+p) + \frac{\operatorname{tr}(\mathbf{Q}^T\mathbf{C}\mathbf{Q})}{\hat{\sigma}^2} \\ &= \log\det\begin{pmatrix}\mathbf{U}_1^T\mathbf{C}\mathbf{U}_1 & \mathbf{U}_1^T\mathbf{C}\mathbf{X} \\ \mathbf{X}^T\mathbf{C}\mathbf{U}_1 & \mathbf{X}^T\mathbf{C}\mathbf{X}\end{pmatrix} + (n-d-p)\log(\hat{\sigma}^2) + (d+p) + (n-d-p)\end{aligned}$$

Using equation (S11) for the determinant of a partitioned matrix, we have

$$\begin{aligned}\log\det\begin{pmatrix}\mathbf{U}_1^T\mathbf{C}\mathbf{U}_1 & \mathbf{U}_1^T\mathbf{C}\mathbf{X} \\ \mathbf{X}^T\mathbf{C}\mathbf{U}_1 & \mathbf{X}^T\mathbf{C}\mathbf{X}\end{pmatrix} &= \log\det(\mathbf{U}_1^T\mathbf{C}\mathbf{U}_1) + \log\det(\mathbf{X}^T\mathbf{C}\mathbf{X} - \mathbf{X}^T\mathbf{C}\mathbf{U}_1(\mathbf{U}_1^T\mathbf{C}\mathbf{U}_1)^{-1}\mathbf{U}_1\mathbf{C}\mathbf{X}) \\ &= \log\det(\mathbf{U}_1^T\mathbf{C}\mathbf{U}_1) + \log\det\Big(\mathbf{X}^T\big[\mathbf{C} - \mathbf{C}\mathbf{U}_1(\mathbf{U}_1^T\mathbf{C}\mathbf{U}_1)^{-1}\mathbf{U}_1^T\mathbf{C}\big]\mathbf{X}\Big).\end{aligned}$$

Ignoring the constants $\log\det(\mathbf{U}_1^T\mathbf{C}\mathbf{U}_1)$ and n which do not depend on $\mathbf{X}$, we obtain eq. (S29). □

Due to the determinant term in eq. (S29), it is not clear whether the restricted maximum-likelihood solution $\hat{\mathbf{X}}$ of Theorem 3 (with its associated maximum-likelihood covariance parameters of Theorem 4) is the absolute minimizer of $\mathcal{L}_{\mathbf{X}}$,

$$\hat{\mathbf{X}} = \operatorname*{argmin}_{\mathbf{X}\in\mathbb{R}^{n\times p}, \mathbf{X}^T\mathbf{X}=\mathbb{1}, \mathbf{X}^T\mathbf{Z}=0} \mathcal{L}_{\mathbf{X}} \quad ?$$

However, we do have the following result:

Theorem 5. *The restricted maximum-likelihood solution $\hat{\mathbf{X}}$ of Theorem 3 is the set of p latent variables that minimizes the residual variance among all choices of p latent variables,*

$$\hat{\mathbf{X}} = \operatorname*{argmin}_{\mathbf{X}\in\mathbb{R}^{n\times p}, \mathbf{X}^T\mathbf{X}=\mathbb{1}, \mathbf{X}^T\mathbf{Z}=0} \sigma^2(\mathbf{X})$$

Proof. By Proposition 1 and the arguments leading up to it, we can write

$$\operatorname{tr}(\mathbf{C}_{22}) = \operatorname{tr}(\mathbf{X}^T\mathbf{C}\mathbf{X}) + \operatorname{tr}(\mathbf{Q}^T\mathbf{C}\mathbf{Q}^T) = \operatorname{tr}\Big((\mathbf{U}_2^T\mathbf{X})^T\mathbf{C}_{22}(\mathbf{U}_2^T\mathbf{X})\Big) + \operatorname{tr}\Big((\mathbf{U}_2^T\mathbf{Q})^T\mathbf{C}_{22}(\mathbf{U}_2^T\mathbf{Q})\Big),$$

where as before $\mathbf{C}_{22} = \mathbf{U}_2^T\mathbf{C}\mathbf{U}_2$ is the restriction of $\mathbf{C}$ to the $(n-d)$-dimensional subspace orthogonal to the d known covariates, and the columns of $\mathbf{U}_2^T\mathbf{X}$ and $\mathbf{U}_2^T\mathbf{Q}$ span mutually orthogonal subspaces within this $(n-d)$-dimensional space. Hence $(n-d-p)\sigma^2(\mathbf{X}) = \operatorname{tr}(\mathbf{Q}^T\mathbf{C}\mathbf{Q}^T)$ is the trace of $\mathbf{C}_{22}$ over the residual $(n-d-p)$-dimensional space orthogonal to the latent variables, within the subspace orthogonal to the d known covariates. By the Courant-Fisher min-max theorem for eigenvalues [2], the $(n-d-p)$-dimensional subspace of $\mathbb{R}^{n-d}$ with *smallest* trace is the subspace spanned by the eigenvectors of $\mathbf{C}_{22}$ corresponding to its $(n-d-p)$ smallest eigenvalues. By Theorem 3, this is exactly the subspace obtained by choosing $\mathbf{X}$ equal to the restricted maximum-likelihood solution $\hat{\mathbf{X}}$. □

S7 Selecting covariates and the latent dimension

Two practical problems remain: how to choose the latent variable dimension parameter p and which known covariates to include?

To choose p, we will use the following result:

Lemma 7.

$$\operatorname{tr}(\mathbf{C}) = \operatorname{tr}(\hat{\mathbf{K}}) = \operatorname{tr}(\mathbf{Z}\hat{\mathbf{B}}\mathbf{Z}^T) + \operatorname{tr}(\hat{\mathbf{X}}\hat{\mathbf{A}}\hat{\mathbf{X}}^T) + n\hat{\sigma}^2$$

Proof. Use Theorem 4 to compute

$$\begin{aligned}
\operatorname{tr}(\mathbf{Z}\hat{\mathbf{B}}\mathbf{Z}) &= \operatorname{tr}\Big(\mathbf{U}_1\mathbf{\Gamma}_1\mathbf{V}^T\big[\mathbf{V}\mathbf{\Gamma}_1^{-1}(\mathbf{C}_{11} - \hat{\sigma}^2\mathbb{1})\mathbf{\Gamma}_1^{-1}\mathbf{V}^T\big]\mathbf{V}\mathbf{\Gamma}_1\mathbf{U}_1^T\Big) \\
&= \operatorname{tr}\big(\mathbf{U}_1\mathbf{C}_{11}\mathbf{U}_1^T\big) - \hat{\sigma}^2\operatorname{tr}\big(\mathbf{U}_1\mathbf{U}_1^T\big) \\
&= \operatorname{tr}(\mathbf{C}_{11}) - d\hat{\sigma}^2,
\end{aligned}$$

where the last step uses the cyclical property of the trace and the fact that $\mathbf{U}_1^T\mathbf{U}_1 = \mathbb{1}_d$. Likewise, we have

$$\begin{aligned}
\operatorname{tr}\big(\hat{\mathbf{X}}\hat{\mathbf{A}}\hat{\mathbf{X}}\big) &= \operatorname{tr}\Big(\mathbf{U}_2\mathbf{W}_p\operatorname{diag}(\lambda_1,\ldots,\lambda_p)\mathbf{W}_p^T\mathbf{U}_2^T\Big) - \hat{\sigma}^2\operatorname{tr}\big(\mathbf{U}_2\mathbf{W}_p\mathbf{W}_p^T\mathbf{U}_2^T\big) \\
&= \sum_{j=1}^{p}\lambda_j - p\hat{\sigma}^2 \\
&= \sum_{j=1}^{n-d}\lambda_j - (n-d)\hat{\sigma}^2 \\
&= \operatorname{tr}(\mathbf{C}_{22}) - (n-d)\hat{\sigma}^2.
\end{aligned}$$

Hence

$$\operatorname{tr}(\hat{\mathbf{K}}) = \operatorname{tr}(\mathbf{Z}\hat{\mathbf{B}}\mathbf{Z}) + \operatorname{tr}(\hat{\mathbf{X}}\hat{\mathbf{A}}\hat{\mathbf{X}}) + n\hat{\sigma}^2 = \operatorname{tr}(\mathbf{C}_{11}) + \operatorname{tr}(\mathbf{C}_{22}) = \operatorname{tr}(\mathbf{C})$$

□

Because $\mathbf{C} = (\mathbf{Y}\mathbf{Y}^T)/m$, the eigenvalues of $\mathbf{C}$ are (proportional to) the squared singular values of the expression data $\mathbf{Y}$. Hence $\operatorname{tr}(\mathbf{Z}\hat{\mathbf{B}}\mathbf{Z})/\operatorname{tr}(\mathbf{C})$ is the proportion of variation in $\mathbf{Y}$ explained by the known covariates, $\operatorname{tr}(\hat{\mathbf{X}}\hat{\mathbf{A}}\hat{\mathbf{X}})/\operatorname{tr}(\mathbf{C})$ the proportion of variation explained by the latent variables, and $n\hat{\sigma}^2/\operatorname{tr}(\mathbf{C})$ is the residual variance.

Our method for determining the number of latent variables lets the user decide *a priori* the minimum amount of variation ρ in the data that should be explained by the known and latent confounders. It follows that given ρ, a "target" value for σ^2 is

$$\sigma^2(\rho) = \min\Big\{\frac{(1-\rho)\operatorname{tr}(\mathbf{C})}{n}, \lambda_{\min}(\mathbf{C}_{11})\Big\},$$

where the minimum with $\lambda_{\min}(\mathbf{C}_{11})$ is taken to ensure that of condition (S26) remains valid. Because the eigenvalues $\lambda_1, \ldots, \lambda_{n-d}$ are sorted, the function

$$f(p) = \frac{1}{n-d-p} \sum_{j=p+1}^{n-d} \lambda_j$$

increases with decreasing p. Hence given ρ, we define $\hat{p}$ as

$$\hat{p} = \min\{p : 0 \leq p < n-d, \lambda_p > \lambda_{n-d}, f(p) < \sigma^2(\rho)\},$$

that is, we choose $\hat{p}$ to be the *smallest* number of latent variables that explain *at least* a proportion of variation ρ of $\mathbf{Y}$, while guaranteeing that the conditions for *all* mathematical results derived in this document are valid.

Note that unless all eigenvalues of $\mathbf{C}_{22}$ are identical, $\hat{p}$ always exists. Once the desired number of latent variables $\hat{p}$ is defined, the latent factors $\hat{\mathbf{X}}$, the variance parameters $\hat{\mathbf{A}}$, and the residual variance estimate $\hat{\sigma}^2$ (which will be the largest possible value less than or equal to the target value $\sigma^2(\rho)$) are determined by Theorem 3. Once those are determined, the remaining covariance parameters $\hat{\mathbf{B}}$ and $\hat{\mathbf{D}}$ are determined by Theorem 4.

A second practical problem occurs when the rank of $\mathbf{Z}$ exceeds the number of samples, such that any subset of n linearly independent covariates explains *all* of the variation in $\mathbf{Y}$. To select a more relevant subset of covariates, we rapidly screen all candidate covariates using the model with a single known covariate (Section S4) to compute the variance $\hat{\beta}^2$ explained by that covariate alone (eq. (S16)). We then keep only those covariates for which $\hat{\beta}^2 \geq \theta \operatorname{tr}(\mathbf{C})$, where $\theta > 0$ is the second free parameter of the method, namely the minimum amount of variation explained by a known covariate on its own. The selected covariates are ranked according to their value of $\hat{\beta}^2$, and a linearly independent subset is generated, starting from the covariates with highest $\hat{\beta}^2$.

S8 Downstream analyses

The inferred maximum-likelihood hidden factors $\hat{\mathbf{X}}$ and sample covariance matrix $\hat{\mathbf{K}}$ are typically used to create a dataset of residuals corrected for spurious sample correlations, to increase the power for detecting eQTLs, or as data-derived endophenotypes [5,8]. We briefly review these tasks and how they compare between LVREML and PANAMA hidden factors.

S8.1 Correcting data for spurious sample correlations

To remove spurious correlations due to the known and latent variance components from the expression data $\mathbf{Y} \in \mathbb{R}^{n \times m}$ (see Section S2), the residuals $\hat{\mathbf{y}}_i \in \mathbb{R}^n$ for gene i with original data $\mathbf{y}_i$ (a column of $\mathbf{Y}$) are contructed as

$$\hat{\mathbf{y}}_i = \hat{\mathbf{K}} \left(\sigma^2_{c,i} \hat{\mathbf{K}} + \sigma^2_{e,i} \mathbb{1}\right)^{-1} \hat{\mathbf{y}}_i$$

where the variance parameters $\sigma^2_{c,i}$ and $\sigma^2_{e,i}$ are fit separately for each gene i [5]. Hence two solutions for the latent factors that give rise to the same $\hat{\mathbf{K}}$ (as observed in Section 2.3 for LVREML and PANAMA) will result in the same residuals.

S8.2 Adjusting for known and latent covariates in eQTL association analyses

Two approaches for mapping eQTLs are commonly used in this context. The first approach tests for an association between SNP $\mathbf{s}_j$ and gene $\mathbf{y}_i$ using a mixed model, where the SNP is treated as a fixed effect, constructing likelihood ratio statistics as

$$\mathrm{LOD}_{i,j} = \log \frac{\mathcal{N}(\mathbf{y}_i \mid \theta \mathbf{s}_j, \sigma^2_{c,i}\hat{\mathbf{K}} + \sigma^2_{e,i}\mathbb{1})}{\mathcal{N}(\mathbf{y}_i \mid 0, \sigma^2_{c,i}\hat{\mathbf{K}} + \sigma^2_{e,i}\mathbb{1})},$$

where the variance parameters $\sigma^2_{c,i}$ and $\sigma^2_{e,i}$ are fit separately for each gene i [5]. Hence for latent factor solutions that give rise to the same $\hat{\mathbf{K}}$ the association analyses will again be identical.

The second approach performs a linear regression of a gene's expression data, typically using the corrected data $\hat{\mathbf{y}}_i$, on the SNP genotypes $\mathbf{s}_j$, using the known and inferred factors as covariates [8], that is, a linear model is fit where

$$\hat{\mathbf{y}}_i = \beta_{i,j}\mathbf{s}_j + \mathbf{Z}\mathbf{a}_i + \hat{\mathbf{X}}\mathbf{b}_i + \epsilon_i \tag{S30}$$

where $\mathbf{Z}$ and $\hat{\mathbf{X}}$ are the matrices of known and estimated latent factors, respectively, and $\mathbf{a}_i \in \mathbb{R}^d$ and $\mathbf{b}_i \in \mathbb{R}^p$ are their respective regression coefficients.

Since maximum-likelihood solutions for the hidden factors by LVREML and PANAMA differ by a linear combination with the known factors $\mathbf{Z}$ that transforms models with hidden factors orthogonal to $\mathbf{Z}$ to equivalent models with hidden factors overlapping with $\mathbf{Z}$, and vice versa (see Section S6.1), it is clear that the same linear transformation will also result in equivalent linear association models in eq. (S30). Hence this type of analysis will also be equivalent between the hidden factors inferred by both approaches.

S8.3 Mapping the genetic architecture of latent variables

Inferred latent variables are sometimes treated as endophenotypes whose genetic architecture is of interest. In this case SNPs are identified that are strongly associated with the latent variables. Different solutions for the latent variables will then clearly result in different sets of significantly associated SNPs.

Using the maximum-likelihood LVREML inferred latent variables that are orthogonal to known confounders is advantageous in this context, because

- The LVREML latent variables are *uniquely defined*. All other solutions that give rise to the same covariance matrix estimate $\hat{\mathbf{K}}$ can be written as a linear combination of the known covariates and the LVREML covariates (see Section S6.1).
- When interpreting associated SNPs, there is no risk of attributing biological meaning to a latent variable that is due to the signal coming from the overlapping known covariates.

To remove the dependence of genetic association analyses on the choice of equivalent sets of latent variables, we recommend performing a multi-trait GWAS on the joint set of known

and latent confounders. If the standard multivariate association test based on canonical correlation analysis [9] is used, results will again be identical between equivalent choices of latent variables, because together with the known confounders they all span the same linear subspace.

Bibliography

[1] Theodore Wilbur Anderson and Ingram Olkin. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear algebra and its applications*, 70:147–171, 1985.

[2] R A Horn and C R Johnson. *Matrix analysis*. Cambridge University Press, 1985.

[3] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[4] Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6(Nov):1783–1816, 2005.

[5] Nicoló Fusi, Oliver Stegle, and Neil D Lawrence. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computational Biology*, 8(1):e1002330, 2012.

[6] H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[7] FN Gumedze and TT Dunne. Parameter estimation and inference in the linear mixed model. *Linear Algebra and its Applications*, 435(8):1920–1944, 2011.

[8] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, 2012.

[9] Manuel AR Ferreira and Shaun M Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–133, 2009.

Supplementary Information for Article II

High-dimensional multi-trait GWAS by reverse prediction of genotypes using machine learning methods

— Supplementary Information —

Muhammad Ammar Malik*, Adriaan-Alexander Ludl and Tom Michoel

* Corresponding author, email: muhammad.malik@uib.no

Supplementary Methods

S1 Canonical Correlation Analysis

Given two sets of random variables $(X_1, X_2, \ldots, X_p)$ and $(Y_1, Y_2, \ldots, Y_q)$, CCA finds linear coefficients $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^q$ that maximize the correlation

$$\rho(a, b) = \text{corr}\left(\sum_{i=1}^{p} a_i X_i, \sum_{j=1}^{q} b_j Y_j\right)$$

It can be shown [1] that the optimal vector $\mathbf{a}$ is an eigenvector of the matrix $\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$, where Σ_{XX}, Σ_{XY} and Σ_{YY} are the covariance matrices among the X and Y variables. In the special case where $q = 1$ (one SNP), Σ_{YY} is a number and Σ_{XY} a column vector, and this matrix takes the form $\Sigma_{XX}^{-1}\mathbf{v}\mathbf{v}^T$, where $\mathbf{v} = \Sigma_{YY}^{-1/2}\Sigma_{XY}$. The (only) eigenvector of such a matrix is $\mathbf{a} = \Sigma_{XX}^{-1}\mathbf{v}$.

To estimate the coefficients $\mathbf{a}$ from data, assume that we have standardized data $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{y} \in \mathbb{R}^n$, such that

$$\sum_{k=1}^{n} x_{ik} = \sum_{k=1}^{n} y_k = 0 \qquad \frac{1}{n-1}\sum_{k=1}^{n} x_{ik}^2 = \frac{1}{n-1}\sum_{k=1}^{n} y_k^2 = 1.$$

Then the estimates for the covariances are

$$\hat{\Sigma}_{XX} = \frac{\mathbf{X}^T\mathbf{X}}{n-1} \qquad \hat{\Sigma}_{XY} = \frac{\mathbf{X}^T\mathbf{y}}{n-1} \qquad \hat{\Sigma}_{YY} = 1,$$

and hence

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

[1] See for instance
Hardoon DR, Szedmak S and Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods *Neural computation* **16**:2639–2664 (2003).

Supplementary Figures

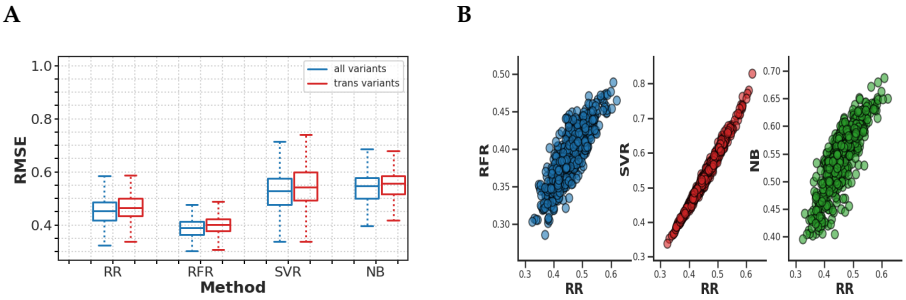


Figure S1: RMSE values for genotype prediction on DREAM5 simulated data. **A.** Boxplots show the distribution of the RMSE values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). **B.** Scatter plots show RMSE values of RFR, SVR, and NB vs RR for all variants. The data shown are for **DREAM Network 2**.

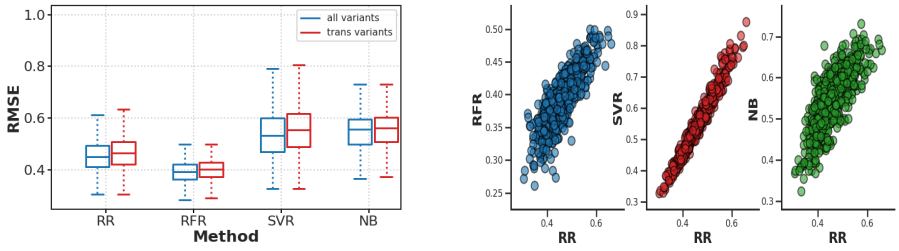


Figure S2: RMSE values for genotype prediction on DREAM5 simulated data. **A.** Boxplots show the distribution of the RMSE values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). **B.** Scatter plots show RMSE values of RFR, SVR, and NB vs RR for all variants. The data shown are for **DREAM Network 3**.

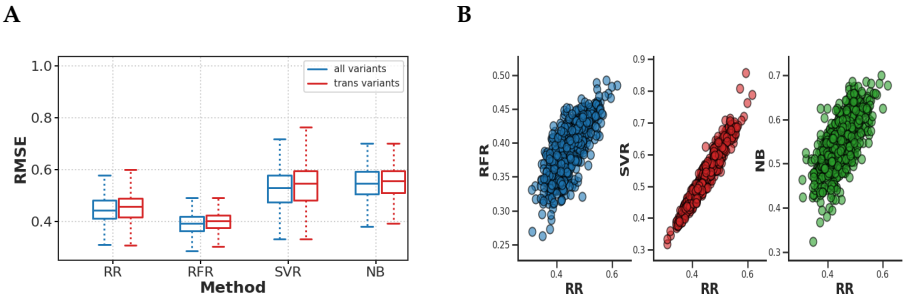


Figure S3: RMSE values for genotype prediction on DREAM5 simulated data. **A.** Boxplots show the distribution of the RMSE values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). **B.** Scatter plots show RMSE values of RFR, SVR, and NB vs RR for all variants. The data shown are for **DREAM Network 4**.

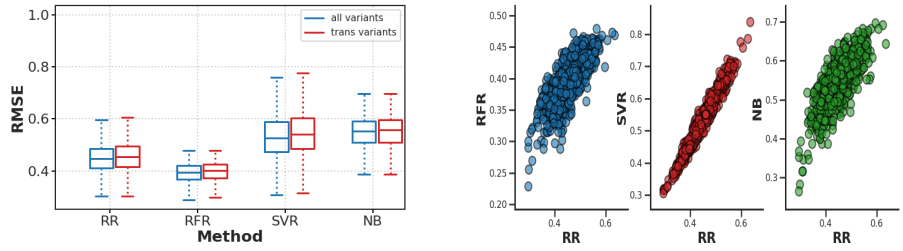


Figure S4: RMSE values for genotype prediction on DREAM5 simulated data. **A.** Boxplots show the distribution of the RMSE values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). **B.** Scatter plots show RMSE values of RFR, SVR, and NB vs RR for all variants. The data shown are for **DREAM Network 5**.

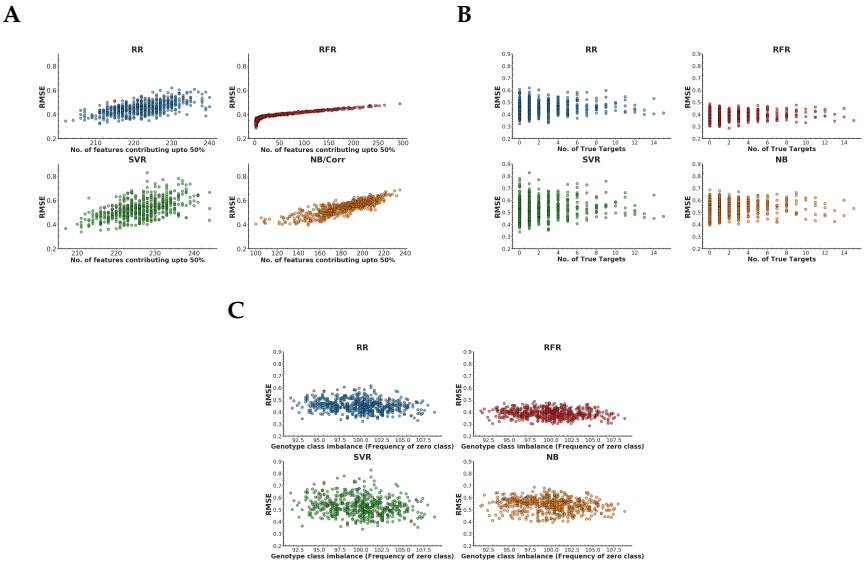


Figure S5: Scatter plots of genotype RMSE values on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). The data shown are for **DREAM Network 2**.

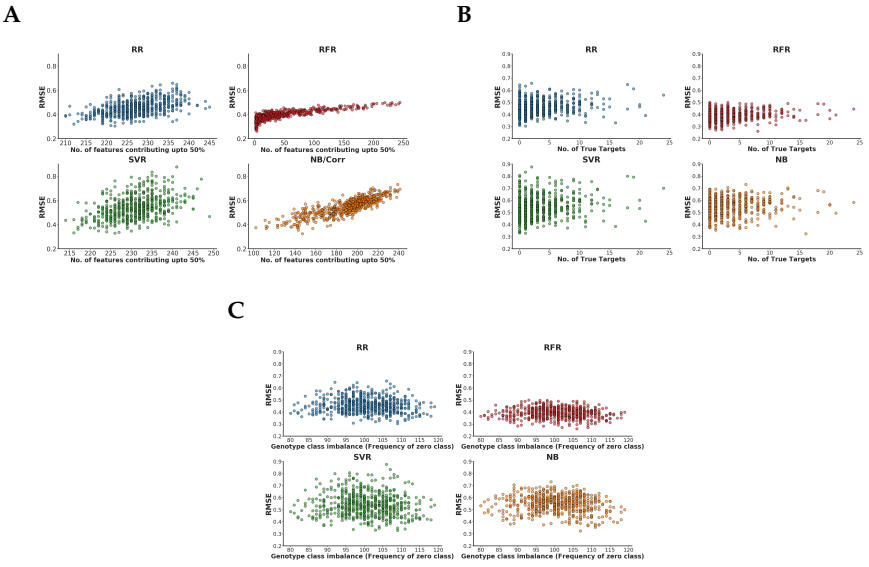


Figure S6: Scatter plots of genotype RMSE values on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). The data shown are for **DREAM Network 3**.

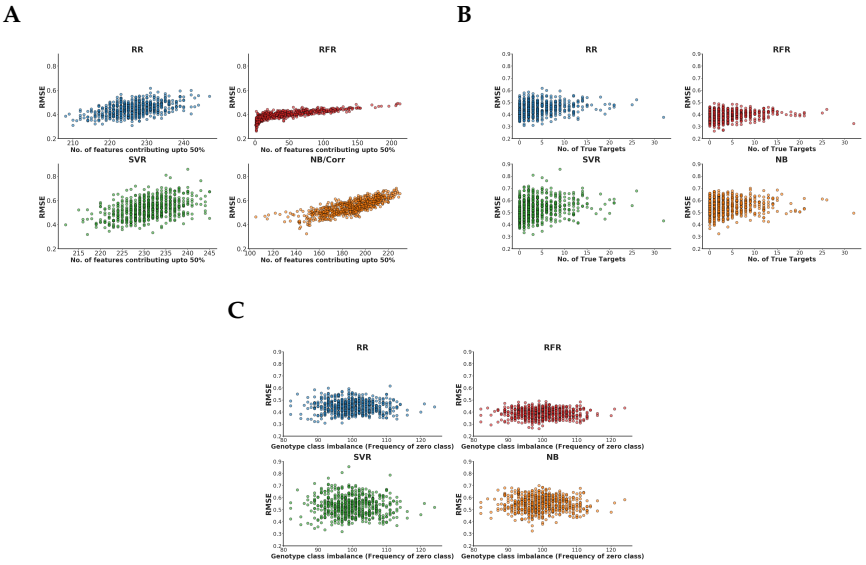


Figure S7: Scatter plots of genotype RMSE values on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). The data shown are for **DREAM Network 4**.

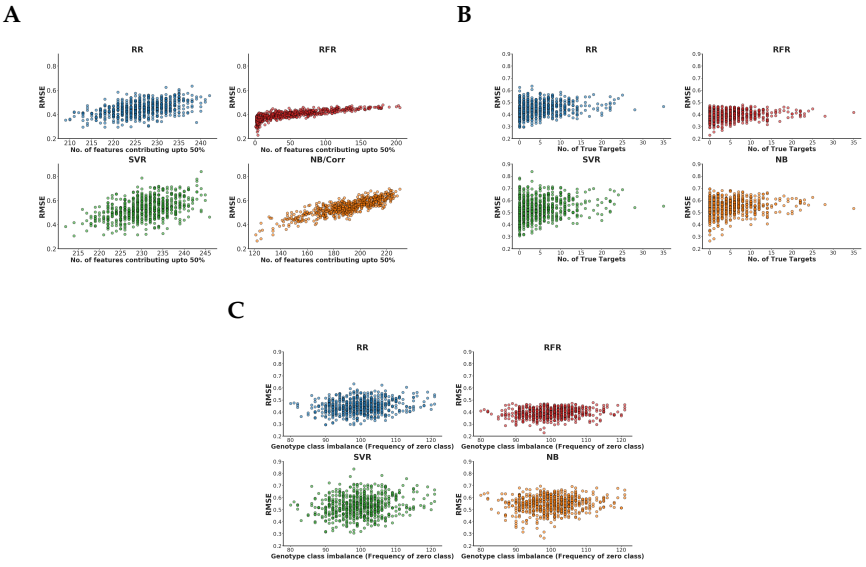


Figure S8: Scatter plots of genotype RMSE values on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). The data shown are for **DREAM Network 5**.

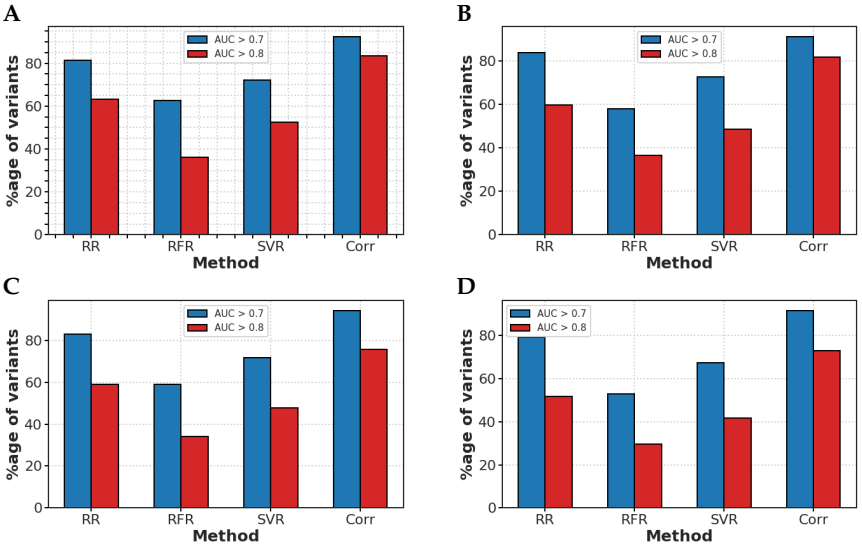


Figure S9: Bar plots show the proportion of variants with trans-eQTL target prediction AUROC > 0.7 (blue) and > 0.8 (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). (**A**) DREAM Network 2, (**B**) DREAM Network 3, (**C**) DREAM Network 4, (**D**) DREAM Network 5.

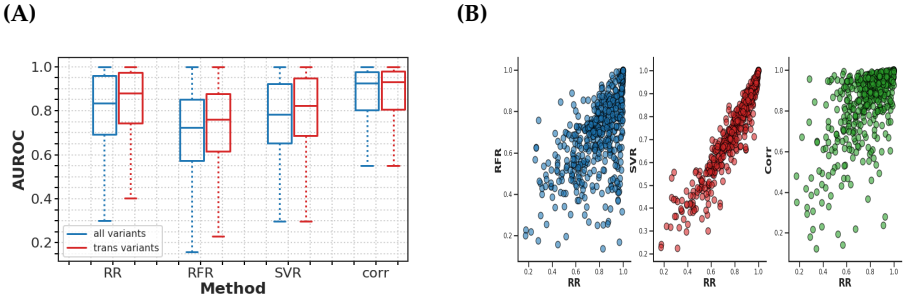


Figure S10: Trans-eQTL target prediction performance on DREAM5 simulated data. **(A)** Boxplots show the distribution of AUROC values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). **(B)** Scatter plots show AUROC values of classification methods RFR, SVR, and Corr vs RR for all variants. The data shown are for **DREAM Network 2**.

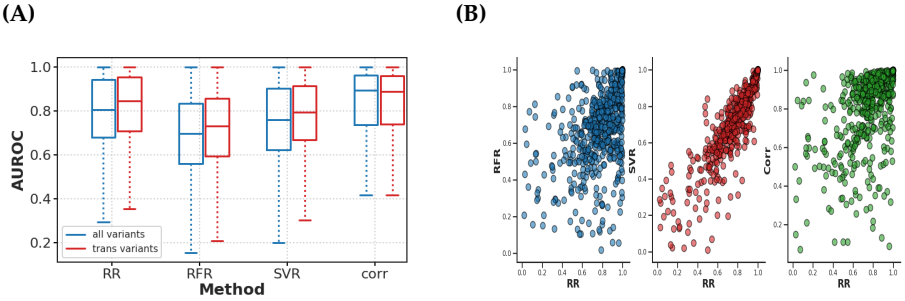


Figure S11: Trans-eQTL target prediction performance on DREAM5 simulated data. **(A)** Boxplots show the distribution of AUROC values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). **(B)** Scatter plots show AUROC values of classification methods RFR, SVR, and Corr vs RR for all variants. The data shown are for **DREAM Network 3**.

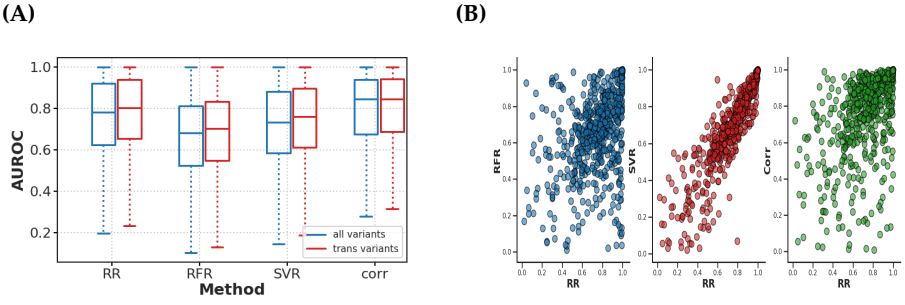


Figure S12: Trans-eQTL target prediction performance on DREAM5 simulated data. **(A)** Boxplots show the distribution of AUROC values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). **(B)** Scatter plots show AUROC values of classification methods RFR, SVR, and Corr vs RR for all variants. The data shown are for **DREAM Network 4**.

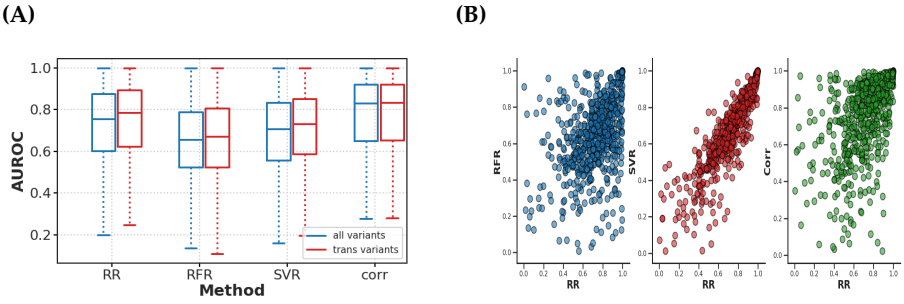


Figure S13: Trans-eQTL target prediction performance on DREAM5 simulated data. **(A)** Boxplots show the distribution of AUROC values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). **(B)** Scatter plots show AUROC values of classification methods RFR, SVR, and Corr vs RR for all variants. The data shown are for **DREAM Network 5**.

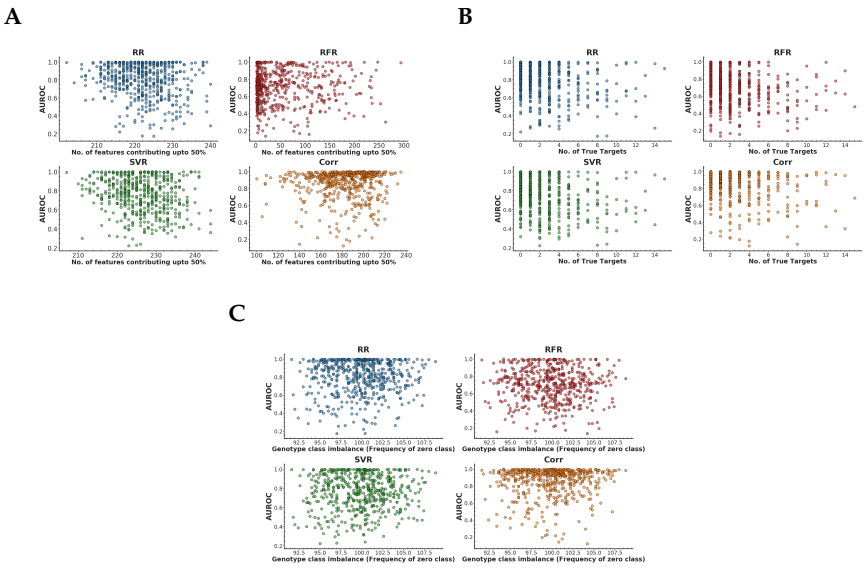


Figure S14: Scatter plots of trans-eQTL target prediction performance (AUROC) on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB). The data shown are for **DREAM Network 2**.

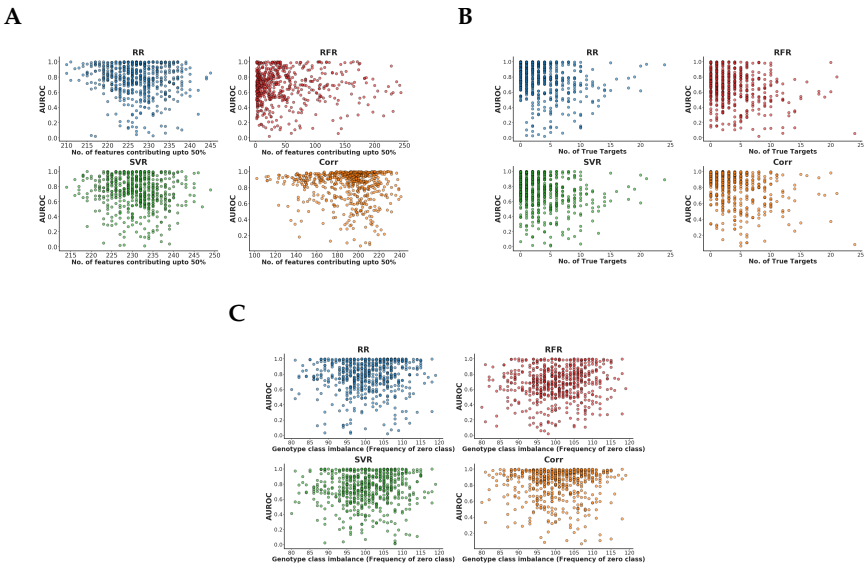


Figure S15: Scatter plots of trans-eQTL target prediction performance (AUROC) on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB). The data shown are for **DREAM Network 3**.

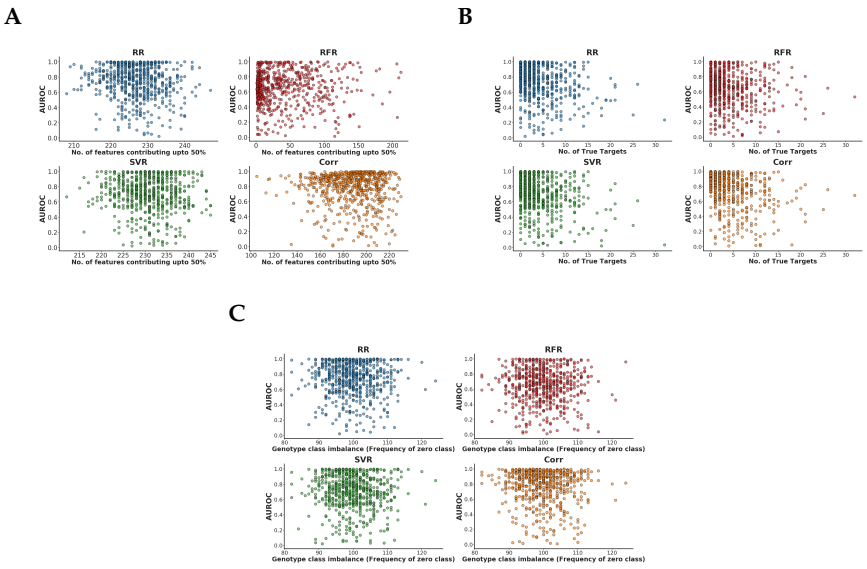


Figure S16: Scatter plots of trans-eQTL target prediction performance (AUROC) on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB). The data shown are for **DREAM Network 4**.

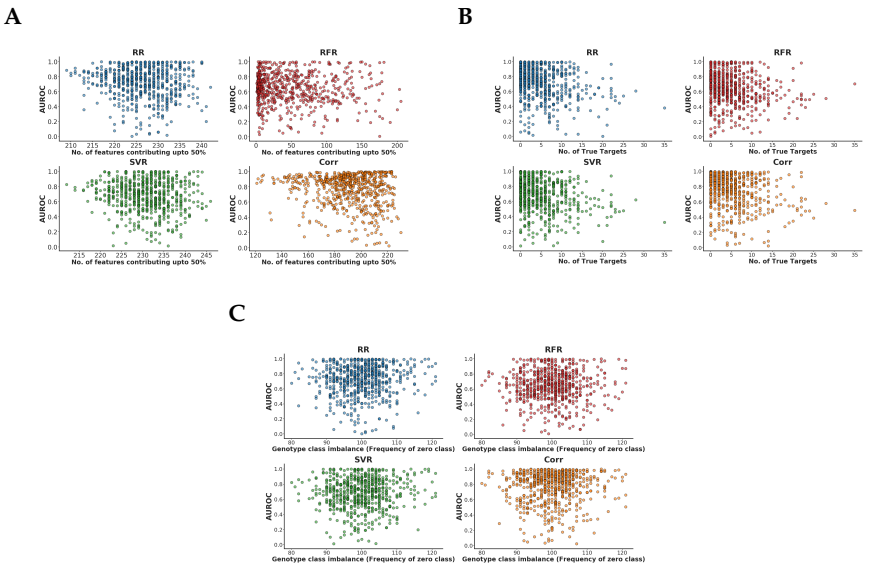


Figure S17: Scatter plots of trans-eQTL target prediction performance (AUROC) on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB). The data shown are for **DREAM Network 5**.

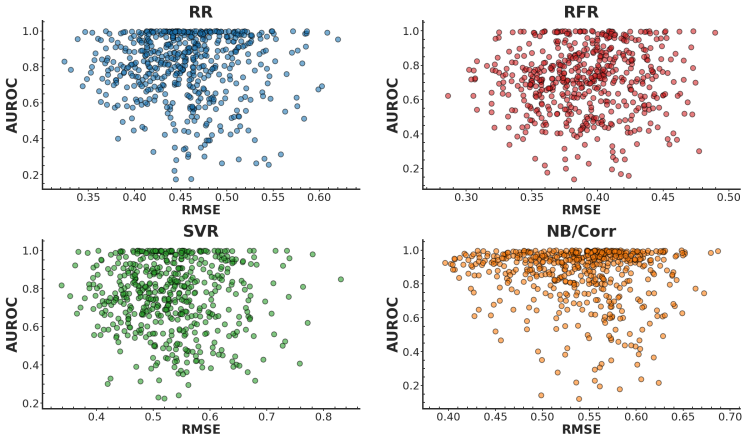


Figure S18: Scatter plots show trans-eQTL target prediction performance (AUROC) vs genotype prediction performance (RMSE) on DREAM5 simulated data for all genetic variants for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB/Corr). The data shown are for **DREAM Network 2**.

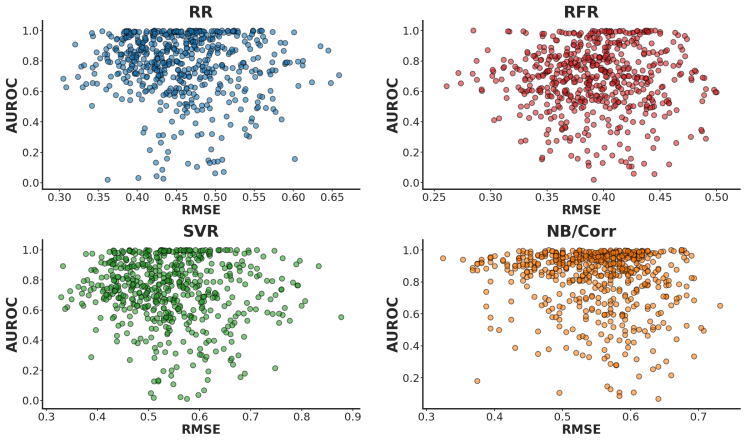


Figure S19: Scatter plots show trans-eQTL target prediction performance (AUROC) vs genotype prediction performance (RMSE) on DREAM5 simulated data for all genetic variants for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB/Corr). The data shown are for **DREAM Network 3**.

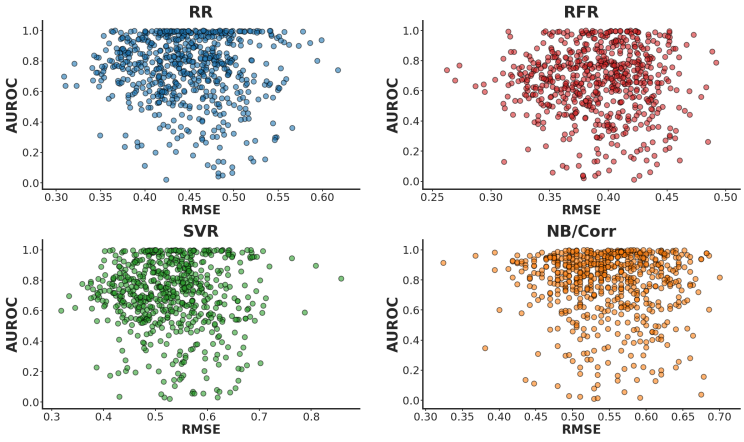


Figure S20: Scatter plots show trans-eQTL target prediction performance (AUROC) vs genotype prediction performance (RMSE) on DREAM5 simulated data for all genetic variants for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB/Corr). The data shown are for **DREAM Network 4**.

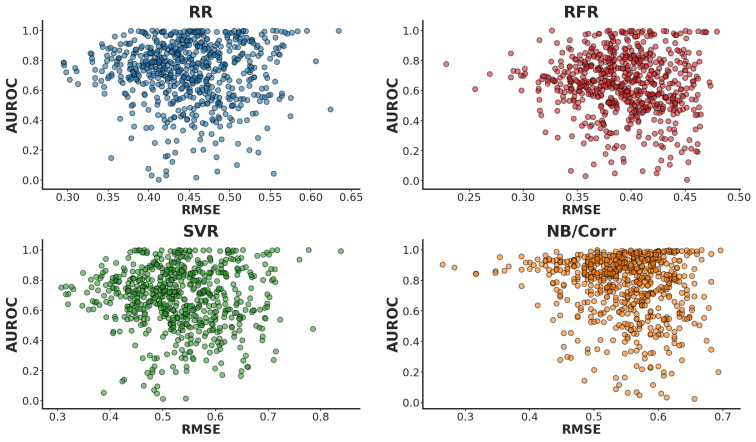


Figure S21: Scatter plots show trans-eQTL target prediction performance (AUROC) vs genotype prediction performance (RMSE) on DREAM5 simulated data for all genetic variants for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB/Corr). The data shown are for **DREAM Network 5**.

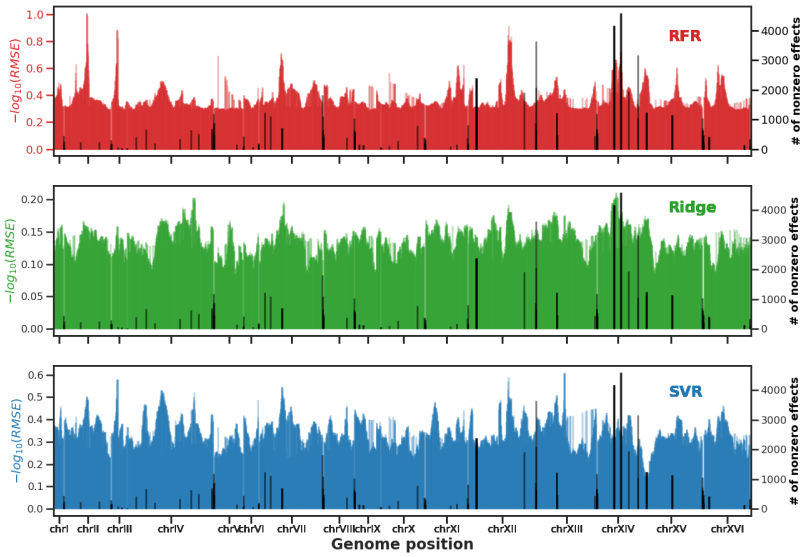


Figure S22: Expression hotspot maps showing the negative log transformed RMSE values vs genome position for 2884 SNPs in the yeast genome, for random forest (RF, top), ridge regression (Ridge, middle), and support vector regression (SVR, bottom). Genes on the same chromosome were excluded as predictors for each SNP. Secondary axis on right shows number of non-zero effects of trans-regulatory hotspot variants from Albert et al. (2018)[2].

[2]Albert, F. W. et al. (2018). Genetics of trans-regulatory variation in gene expression. Elife, 7, e35471

Supplementary Information for Article III

rfPhen2Gen: A machine learning based association study of brain imaging phenotypes to genotypes

— Supplementary Information —

Muhammad Ammar Malik*, Alexander S. Lundervold and Tom Michoel

* Corresponding author, email: muhammad.malik@uib.no

Supplementary Figures

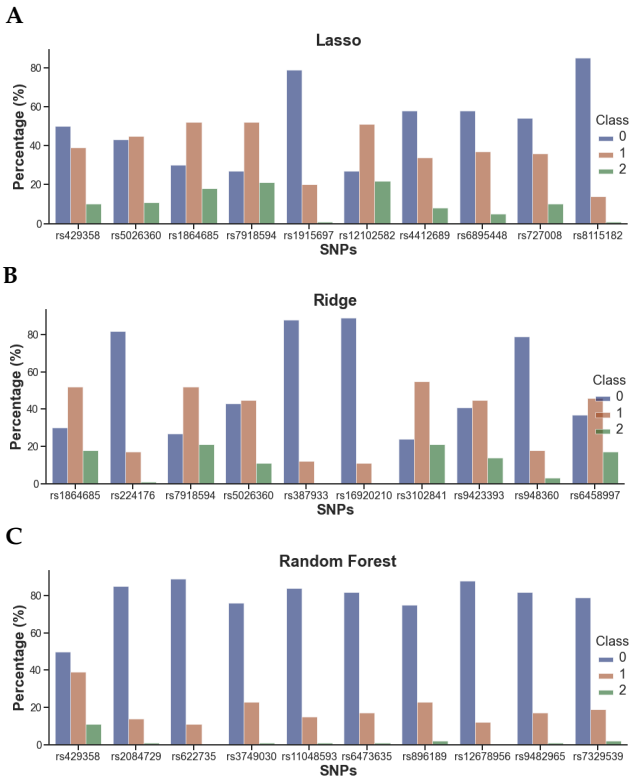


Figure S1: Class distribution for the top 10 SNPs identified by Random Forest, Lasso and Ridge regression.

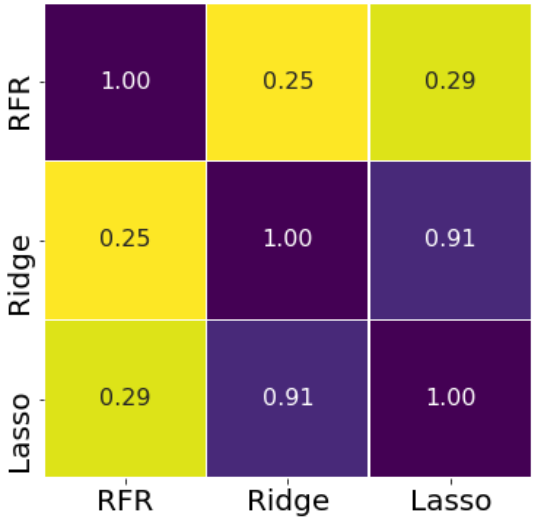


Figure S2: Spearman correlation coefficients for RMSE values across the methods

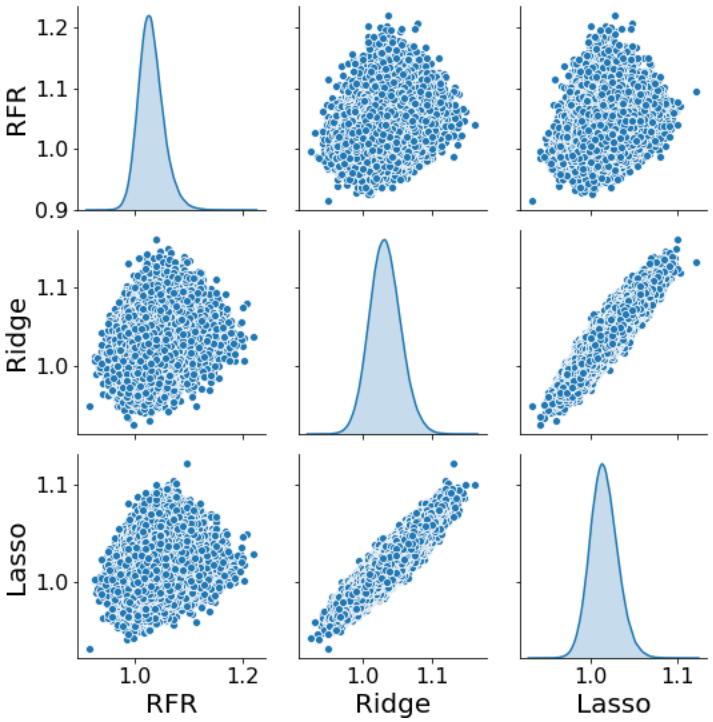


Figure S3: Pairplot showing the RMSE distribution and scatter plots between the methods, Random Forest Regression (RFR), Ridge Regression (RR), Lasso Regression (LR)

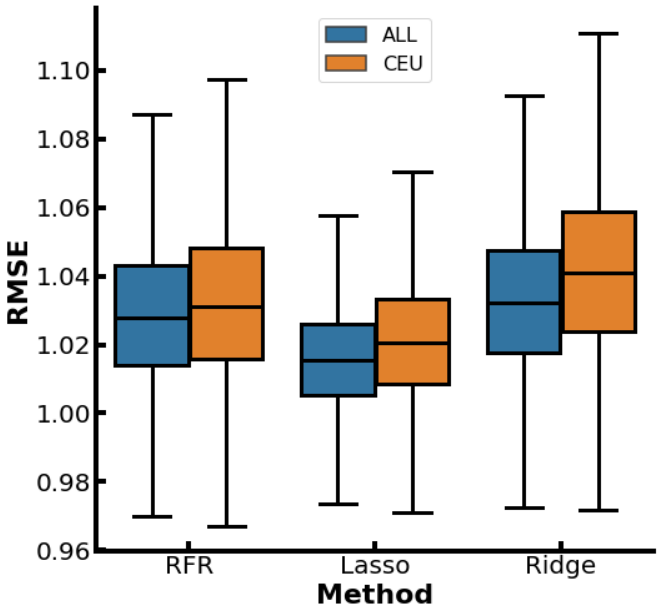


Figure S4: Boxplots showing RMSE distribution across all SNPs for Random Forest Regression (RFR), Lasso Regression and Ridge Regression, for all samples vs only Caucasian samples

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230853634 (print)
9788230850626 (PDF)