A mixture model to decompose the heritable basis of complex traits

Luis Torada Aguilella

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

A mis padres

Acknowledgements

I would first like to thank those whose direct contributions have made this DPhil possible. I am very grateful to my supervisors Robert Davies, Satu Nahkuri, Simon Myers and Tony Kam-Thong for their feedback and advice, with a special mention to Simon, from whom I have learnt enormously about how excellent research is done and the critical importance of simplicity and detail. I am similarly very grateful to my funding bodies: the Engineering and Physical Sciences Research Council (EPSRC) and the Medical Research Council (MRC) (grant number EP/L016044/1). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z.

I am grateful to all the great people around me that have filled my time in Oxford with countless enriching conversations and wonderful times: the colleagues from my SABS cohort, Antonia Kormpa, Carlos Outeiral, David Garciandía, Deborah Sulem, Fabhiola Oviedo, Georgios Kalantzis, Maria Kiourlappou, Martin Buttenschoen, Miriam Stricker, Natalia García, Solveig van der Vegt, Tatjana Schulze and many others. I would like to specially thank my dear friend Georgios Kalantzis, with whom I have had the honour to share my entire journey in Oxford, the department of Statistics, and Statistical Genetics, and from whom I keep learning everyday. Our scientific discussions, but most importantly his personal integrity, cleverness and kindness, have been an immensely valuable guide throughout our DPhils.

Thank you to my parents, to whom I am dedicating this thesis, for their continuous and unconditional support. They have always heroically put me and my education first, making sure that I never lacked the resources to succeed. This thesis is a direct consequence of their help.

Thank you to those further away, who have understood my absence and supported me from a distance. A special thanks for this to Hannah Hassani, Jesús Cantero and Luis Borja García, and also to the rest of my family and friends.

Finally, thanks to the sources of so many echoes of love and dedication coming from teachers and mentors all the way back from school, who have in a way prepared me for the stage that I am concluding with this thesis. I hope that the closing of this chapter brings me closer to being able to give back all that I have received.

Abstract

Pharmacological drugs aim at modulating traits by targeting their causal mechanisms. However, knowledge of such causal mechanisms is scarce and constitutes the main bottleneck in drug development today. To address the issue, the consensus is to associate as many genetic variants as possible to traits of interest in order to then investigate their function and assign them to different mechanisms on that basis. With that purpose, there is an ongoing large-scale coordinated effort to systematically sequence and phenotype larger cohorts of individuals and to map functional elements in the genomes of human cells across tissues. Despite these advances, existing methods do not take full advantage of these increasingly available resources to associate genetic variants to traits and isolate the different mechanisms through which they operate. We have developed a mixture model that integrates multi-trait Genome-Wide Association Study (GWAS) z-scores and functional annotations of Single Nucleotide Polymorphisms (SNPs) to simultaneously boost GWAS power and group together SNPs that likely operate through a similar mechanism. The parameters of the model can be quickly inferred, and we show with realistic simulations that we can recover substantially more true associations than linear regression or a multi-trait GWAS meta-analysis method (MTAG), while recovering the simulated interpretable mixture model components. We applied our model to Coronary Artery Disease (CAD) and Autism Spectrum Disorder (ASD), finding three components for CAD (which point to the regulation of LDL, risk for smoking, and systolic blood pressure, respectively), and a single component for ASD that implicates the fetal and adult brain. Overall, our mixture model constitutes a powerful new framework to integrate the increasingly available functional annotations of the genome and multi-trait GWAS z-scores to uncover the mechanisms that drive complex traits. We expect our method to be progressively insightful as more and better data becomes available. This is expected to be specially true for complex neurodevelopmental disorders like ASD, whose driving mechanisms may only be detectable when considering developmental stage and cell-type specific functional annotations and traits.

Contents

List of Figures

List of Tables

List of Abbreviations

ADHD . . . . Attention Deficit Hyperactivity Disorder.

ASD . . . . . Autism Spectrum Disorder.

CAD . . . . . Coronary Artery Disease.

EA . . . . . Educational Attainment.

GTEx . . . . . Genotype-Tissue Expression project.

GWAS . . . . . Genome-wide Association Study.

HapMap3 . . . HapMap project, phase 3.

LD . . . . . Linkage Disequilibrium.

LDL . . . . . Low-Density Lipoprotein.

MAF . . . . . Minor Allele Frequency.

MDD . . . . . Major Depressive Disorder.

MM . . . . . Mixture Model.

PPA . . . . . Posterior Probability of Association.

Roadmap . . . Roadmap Epigenomics project.

ROC . . . . . Receiver-Operating Characteristic.

SCZ . . . . . Schizophrenia.

SEG . . . . . Specifically Expressed Genes.

SMK . . . . . Past smoking.

SNP . . . . . Single Nucleotide Polymorphism.

sysBP . . . . . Systolic Blood Pressure.

1000G . . . . . 1000 Genomes Project.

Notation

Data size

J Number of SNPs (indexed by i or j).

K Number of non-null components (we always infer K+1 components: 1 null and K non-null. Components are indexed by k, with $k = 0$ corresponding to the null component).

N_s Number of individuals in the cohort where trait s was measured.

P Number of traits (indexed by s or t).

Q Number of annotations.

Hidden data

e_i $[e_i \in \mathbb{R}^P]$ Non-genetic component of z_i ($z_i = g_{i,k} + e_i$).

$g_{i,k}$ $[g_{i,k} \in \mathbb{R}^P]$. Genetic component of z_i, given that SNP i belongs to the k^{th} component ($z_i = g_{i,k} + e_i$).

h_i $[h_i \in \{0, 1, .., K\}]$ Component to which SNP i belongs.

Observed data

a_i $[a_i \in \mathbb{R}^Q]$ Raw annotations for SNP i. The raw annotations of J SNPs are stored in $A \in \mathbb{R}^{J \times Q}$.

l_i $[l_i \in \mathbb{R}^Q]$ LD annotations for SNP i. The LD annotations of J SNPs are stored in $L \in \mathbb{R}^{J \times Q}$.

X, Y $[X \in \mathbb{R}^{N_s \times J}, Y \in \mathbb{R}^{N_t \times J}]$ Centred and standardised genotypes matrices of N_s and N_t individuals, respectively, at J SNPs. When needed, a second genotypes matrix for N_t individuals is named $Y \in \mathbb{R}^{N_t \times J}$.

y_s $[y_s \in \mathbb{R}^{N_s}]$ centred measurements of trait s in N_s individuals.

z_i $[z_i \in \mathbb{R}^P]$ GWAS z-scores for SNP i and P traits. The z-scores of J SNPs are stored in a matrix $Z \in \mathbb{R}^{J \times P}$.

Parameters

θ $[\theta = \{C, V, w, f\}]$ Set of all the parameters of the model (excluding the likelihood weights, v, which are fixed).

C $[C = \{\Sigma_k : k \in \{0, ..., K\}\}]$ Set of covariance matrices (Σ_0 is for the null component).

f $[f = \{f_k : k \in \{0, ..., K\}\}]$ For each component, f_k is a function of $l_i^T V_k$ that mostly prevents the occurrence of negative values. In summary, it segments the linear predictions into continuous inter-percentile intervals and adjusts the slope within each interval to either improve the local fit (typically not needed) or avoid negative predictions. (See section 2.3.6).

V $[V \in \mathbb{R}^{Q \times (K+1)}]$ Annotation weights for each component.

v $[v \in \mathbb{R}^J]$ Likelihood weights (v_i for SNP i).

w $[w \in \mathbb{R}^{K+1}]$ Weights for the $K + 1$ components. It controls the proportion of SNPs in each component.

Other variables

$\pi_{i,k}$ Prior probability that SNP i belongs to component k (given its LD annotations).

$r_{i,k}$ Posterior probability that SNP i belongs to component k (given its LD annotations and z-scores). This is not to be confused with the correlation of two SNPs, $r_{i,j}$, which we use briefly during the derivation of the model (section 2.2. We deliberately reserve capital letters for either constants, matrices or sets).

1

Introduction and background

Contents

1.1 Towards the pharmacological modulation of complex traits

A trait is a measurable attribute of an individual. Autism, height, body mass index, the concentration of glucose in blood, educational attainment, diabetes, etc. are all human traits; the fact that many disease traits seriously affect our daily lives justifies our effort to control them at will, perhaps by pharmacological means.

Most traits are determined by a combination of genetic and environmental factors, with the genome alone explaining a significant part of their variance [1–3]. Genome-wide association studies (GWAS) aim at associating genetic variants (typically, but not only, single nucleotide polymorphisms, SNPs[1]) to traits, hoping that these will

[1]SNPs are by definition present in more than 1% of the individuals of the studied population.

ultimately point to the mechanisms driving the traits [7], among other applications (e.g. risk prediction [8]). A linear model of association is generally used (or logistic, if traits are binary), with the linear coefficients often being referred to as 'SNP effects' and their estimates as 'summary statistics'. Over the last fifteen years, tens of thousands of SNPs have been associated to thousands of traits [5, 9], and these numbers will continue to grow as larger and more deeply phenotyped cohorts are studied. Consider for example the UK Biobank project [10], which has genotyped and measured thousands of traits in roughly 500,000 individuals from the UK, or recent GWAS surpassing the million of participants [11–14]. However, the great challenge for the present decade is to translate this vast number of associations into an actionable understanding of the mechanisms by which they affect traits [5, 15].

From GWAS to biological insights

Translating GWAS associations (or 'hits') into biological insights has typically consisted of three sequential steps [7]. First, because the associated variants are not always the truly causal variants due to linkage disequilibrium (LD, that is, correlations between variants), the truly causal variants are estimated (or 'fine-mapped' [16]). Second, credible sets of causal variants are linked to the likely affected genes nearby. Finally, the function of the implicated genes and variants is investigated experimentally, which hopefully points to the driving mechanism and to potential pharmacological targets [5, 7]. If multiple GWAS associations implicate several genes, interpreting their functions as a group can further help to delineate the driving mechanism of traits [17–19]. This is typically done by assigning genes to pre-defined functional categories, calculating a score for the strength of

(cont.) Microarrays genotype only a fraction of the SNPs in the genome, allowing for the imputation of many other variants thanks to available haplotype panels like the one offered by the 1000 Genomes Project [4]. Although ultra-rare variants cannot generally be well imputed due to not being well represented in such panels, most GWAS focus on more common variants due to the lower cost of microarrays compared to whole-exome or whole-genome sequencing (WES, WGS), as well as for statistical power considerations [5]. Rarer variants have the potential of having larger and more interpretable effects on traits than more common variants on which selection has not acted that strongly [6] so, to enjoy these benefits, WES and WGS may gain popularity as their prices drop in the future.

their association to a trait, and checking for dependencies between the score and functional category memberships (often referred to as 'gene-set analysis') [17].

There are some examples of at least partial success in using GWAS hits to point to key biology of traits (blood concentration of urate, IGF-1 or testosterone [20], obesity, schizophrenia, osteoporosis, etc. [5, 7]). A widely known example is Coronary Artery Disease (CAD), for which the implication of increased levels of LDL led to drugs that inhibit the endogenous synthesis of LDL ('statins') and have reduced mortality at a large scale [21, 22]. In general, GWAS insights into the biology of traits have retrospectively translated into a higher likelihood of drug approvals [23, 24].

The mechanisms that drive most traits are however not well understood, with a notable example among other psychiatric traits being Autism Spectrum Disorder (ASD), for which currently there is no pharmacological treatment [25]. More than one hundred genetic variants have been associated to ASD with some evidence of functional convergence[2] [25, 26], but actionable insights remain elusive partly due to the complexity of the trait, which involves the brain and early stages of development [25].

Leveraging the polygenic component of complex traits

ASD, like most traits, is likely affected by hundreds or thousands of causal variants scattered throughout the genome, and is said to be a 'complex trait' in contrast to traits driven only by one or few variants ('Mendelian traits') [27, 28]. A minority of the variants that affect complex traits are often rare and play stronger and more interpretable roles in controlling the traits (as they affect 'core' genes), whereas the great majority tend to be more common, have weaker individual effects, and operate through less interpretable regulatory roles [6]. Indeed, most GWAS associations lie in non-coding regions of the genome [29]. A problem is that the weak effects of most variants influencing complex traits means that we generally lack the statistical power to associate them to traits in a GWAS [5]. Furthermore, regulatory elements,

[2]They are connected via cis- and trans-regulation and protein binding, they are mainly expressed in early fetal stages and in certain areas of the brain, and at least some of them relate to three cell-signalling pathways (mTOR, Wnt and MAPK) [25].

where most of the GWAS associations lie, typically have cell-type specific activity. This means that, in order to interpret the function of GWAS hits, the activity of the regulatory elements has to be determined experimentally for each cell-type separately, requiring an enormous investment of resources [30].

Recent methodological advances have to some extent overcome the need for statistically significant variants. Instead, they use genome-wide SNPs to estimate concentrations of genetically explained variance (or 'heritability') in different functional categories of SNPs ('heritability enrichments' [31], for example, in enhancers[3] active in the liver) [33–36]. These methods sometimes even simultaneously use their estimated heritability enrichments to boost GWAS power [35, 36].

As for the cell-type-specific mapping of active regulatory elements, technological advances have enabled the production of genome-wide maps of regulatory activity by the combination of a range of molecular assays with DNA sequencing [37, 38]. Regulatory elements are stretches of DNA that, in combination with binding proteins, control gene expression [39]. Their differential activity is controlled by the so called 'epigenetic marks', and is responsible for regulating human traits and the differentiation of cells within the human body [37, 38]. Specifically, epigenetic marks are slight modifications of the DNA (e.g. cytosine methylations) or of the histones that form the nucleosomes (e.g. methylations or acetylations of one of their aminoacids[4]), and can be detected experimentally at a large scale by methods such as MRE-seq[5] or Chip-seq[6] [37, 38]. The activity of some types of regulatory elements is strongly associated to specific combinations of overlapping epigenetic marks (e.g. active enhancers with H3K27ac [40]), and higher-level regulatory states can also be either defined in an unsupervised way [41] or determined experimentally (e.g. accessible chromatin regions with DNase we footprinting [42], or transcription

[3]An enhancer is a type of regulatory element in the genome that, in cooperation with binding proteins called 'transcription factors', control the expression of one or several genes [32].

[4]These are typically abbreviated by a code of the type of 'HK4me3', where H points to a histone modification, K4 to the fourth lysine of the chain of aminoacids of the histone, and me3 to the type of modification (in this case the addition of three methyl groups).

[5]MRE-seq: methylation-sensitive restriction enzymes and sequencing. It is used to detect DNA methylation.

[6]Chip-seq: chromatin immunoprecipitation and sequencing. It is used to detect histone modifications and transcription factor binding.

factor binding with Chip-seq [42]). Throughout this thesis, we will collectively refer to epigenetic marks and any other local information about the DNA related to gene expression as functional annotations (e.g. SNP-specific chromatin accessibility or transcription factor binding).

Consortia dedicated to systematically generating functional annotations of the genome typically release their results in the form of publicly available catalogues. Examples of widely used catalogues are GTEx [43] (for gene expression across tissues), Roadmap [38] and ENCODE [44] (both for epigenetic marks across tissues[7]). Although existing catalogues are still sparse and mostly for bulk tissue (i.e. they do not distinguish between cell types within the same tissue), these resources are rapidly expanding and increasingly moving toward single cell-type resolution. To give a sense of the speed at which they are growing, two years ago Vierstra *et al.* [42] released about 4.5m transcription binding sites in the genome that are active in different combinations of 243 cell types and tissues, often also predicting the specific transcription factors that bind in each site. A year later NIH announced the extension of the GTEx project to the study of gene expression of about 30 tissues at four different developmental stages[8]. In December 2022, Rood *et al.* published a perspective article in *Nature Medicine* commenting on the potential for the Human Cell Atlas project for medicine [45], which aims at identifying all the cell-types in the human body and indexing them based on their molecular profiles. Of particular relevance for ASD, two months earlier the HCA project announced two new releases for both the adult and first-trimester human brain [46, 47], and simultaneously the PsychEncode project is also systematically mapping functional elements of the brain [48]. Furthermore, newly available datasets are not only the products of new experiments, but also of meaningful transformations of already existing ones. For example, Boix *et al.* [40] imputed thousands of new epigenomic fields for hundreds of tissues combining data from several existing resources (such as GTEx and Roadmap), defined enhancer modules based on cross-tissue enhancer activity

[7] And as part of the broader International Human Epigenome Consortium: `https://ihec-epigenomes.org/research/projects/`

[8] `https://www.genome.gov/news/news-release`

patterns, and linked the enhancers to their candidate regulated genes. Wang *et al.* and the PsychEncode team [49] also derived regulatory networks from their adult brain data collection by linking genes, enhancers and transcription factors.

As discussed above, one way of investigating the biology of traits consists of implicating functional elements and cell-types throughout the genome for a given trait. However, finding patterns of SNP effects across traits can provide further biological insights by telling the extent to which the genetic components of traits 'overlap' or correlate ('genetic correlations' [50]). GWAS hits generally have effects on more than one trait [5, 51], and the overall genetic components of traits are often significantly correlated [52–54]. Like heritability enrichments, genetic correlations can also be used to boost GWAS power [55].

In conclusion, a key idea is that aggregating functional information across many variants, either in the form of their likely affected functional elements in the genome or their patterns of effects across traits, has the potential to shed light on the mechanisms that drive complex traits.

Modelling multiple driving mechanisms

Instead of studying the overall functional properties of a single group of causal variants, a recent novel approach has been to cluster the variants into functionally distinct groups (or 'functional modules') in order to distinguish between multiple causal mechanisms [56, 57]. In the next section, we will compare some recently developed methods that model the existence of multiple groups of SNPs based on different metrics of functional similarity. As we will emphasise, none of them leverages all the newly available functional information about SNPs, and in some cases their clustering ability is overly constrained by design. This thesis fills the gap with a flexible new method that uses functional annotations of SNPs and GWAS summary statistics for multiple traits to cluster SNPs into different functional groups in an unsupervised way. As we will explain, our new method also boosts GWAS power by simultaneously leveraging the functional properties of the identified groups to calculate probabilities of association of SNPs with a trait of interest.

1.2 Previous efforts to identify the functional modules that drive complex traits

There are several features that differentiate recent methods modelling the existence of multiple functional modules affecting a trait (Table 1.1). A major distinguishing feature is whether they find the functional modules in an unsupervised way or whether they instead select the pre-defined modules that matter for a trait. The methods Stratified LD score regression (s-LDSC) [33], fgwas [36] and RSS-NET [35] depend on pre-defined functional modules in the form of SNP annotations (for example, whether they are in an enhancer that is active in the kidney or not), and estimate whether causal SNPs concentrate in those modules. That is to say, they estimate whether having a particular annotation increases the likelihood of SNPs being causal with respect to some baseline or, equivalently, whether an annotation is 'enriched in heritability'. The different methods do it in slightly different ways: s-LDSC estimates only heritability enrichments by regressing squared standardised GWAS summary statistics (z-scores) on a transformed version of SNP annotations (called LD annotations, as we will explain in the next section), whereas fgwas and RSS-NET, at the cost of speed, use a Bayesian framework and are also able to calculate posterior probabilities of association that leverage their estimates of heritability enrichments. RSS-NET, compared to fgwas, additionally leverages knowledge about the connectivity of SNPs in regulatory networks to further boost GWAS power.

Among the three methods, s-LDSC is almost routinely used whenever a new GWAS is carried out, among other reasons for its simplicity and speed, which make it convenient. As an example, Sinnott-Armstrong *et al.* [20] implicate regulatory elements active in the kidney in the modulation of the serum levels of urate. They go on to weight the relative contribution to heritability of the urate synthesis pathway in particular, concluding that it plays only a small role. In a more recent study, s-LDSC is used in combination with more sophisticated annotations that capture cell-type, disease or multi-cell-type specific pathways (based on single-cell expression data) to investigate immune and brain related traits [58].

Table 1.1: Methods comparison. For g-SEM, the subscript in K* emphasises that there are constraints in the kind of modules that it can find (as discussed in the main text). Question marks mean that there is insufficient information in the paper to reach a conclusion on the matter.

	s-LDSC	fgwas, RSS-NET	GRPC	g-SEM	PDR	MM
Number of modules	K	K	K	K*	K	K
Unsupervised	x	x	✓	✓	✓	✓
Multiple traits	x	x	✓	✓	✓	✓
Annotations	✓	✓	x	x	x	✓
Genome-wide SNPs	✓	✓	x	✓	✓	✓
Boosts GWAS power	x	✓	x	✓	✓	✓
Fast	✓	?, x	?	x	x	✓

In contrast to s-LDSC, fgwas and RSS-NET, there are three other methods that group SNPs into functional modules in an unsupervised way: GRPC (Genetic Risk Profile Clustering, abbreviating the title of the methods section of Cortes et *al.* 2020 [51] where it is described), g-SEM (genomic Structural Equation Modelling) [59] and PDR (Pleiotropic Decomposition Regression) [60]. GRPC, g-SEM and PDR base their clustering of SNPs on their patterns of GWAS effects across multiple traits. A major difference between them is that GRPC uses only statistically significant summary statistics (and is consequently restricted to the union of GWAS hits across traits), whereas g-SEM and PDR use genome-wide SNPs regardless of their association status. The frameworks that they use are also very different, with GRPC using hierarchical clustering, g-SEM being based on structural equations modelling (SEM), and PDR using a mixture of Gaussians. Within the SEM framework, they first estimate the 'genetic covariance matrix' of a collection of traits (i.e. the covariance matrix of their genetic components) and then model the genetic components of multiple traits as the result of a linear system of hidden genetic components, in a way that is consistent with the estimated genetic covariance matrix. SNPs are then assigned to a hidden genetic component if their multi-trait effects seem to be mediated by the component. A limitation of g-SEM is that the number of

freely estimated parameters in their linear system of hidden genetic variables cannot exceed the number of non-redundant elements in the genetic covariance matrix [61]. In practice, this limitation forces strong constraints about the number and nature of the functional modules to be found: for example, we would be unable to find three functional modules with arbitrary covariance matrices for three different traits.

In the paper where GRPC was introduced [51], they associated 3,000 independent loci to at least one of thousands of traits from the UK Biobank, and then clustered them in that way into 339 groups. Although they did not use SNP annotations for the clustering, they used them afterwards as an additional metric for validating the functional similarity of SNPs within clusters. They found that many clusters affected nearby genes enriched in annotations related to biological processes at different levels of resolution: for example, one cluster had 16 loci affecting a range of metabolic traits, and nearby genes were enriched in functions related to lipoproteins, suggesting a vague functional similarity of the affected genes.

Using g-SEM, Mallard *et al.* 2022 [62] found two somewhat correlated hidden genetic variables that explain most of the heritability of a collection of psychiatric traits. One of the hidden components affects common psychiatric symptoms and disorders, whereas the other component affects rarer and more severe psychiatric disorders. Associating SNPs to each hidden component led to the prioritisation of 11 genes for one component and 16 genes for the other (with two genes of overlap). As with GRPC, SNP annotations were used only *a posteriori* to validate the functional similarity of the SNPs in each cluster and to understand the difference between the clusters. The prioritised genes for both components were enriched in general neurodevelopmental and CNS-related functions, with some differences between the factors (e.g. neuronal axons vs. neuronal dendrites). In order to further validate the factors, they genetically correlated the two factors with morphological features of different regions of the brain. Based on the correlations, they prioritised an additional set of genes for each factor whose spatial expression was consistent with the genetic correlation estimates. They note that the temporal expression of the prioritised genes resembles that of prenatal and postnatal inhibitory neuronal genes.

When PDR was introduced [60], the authors used it to investigate the mechanisms that modulate the genetic risk of coronary artery disease (CAD). They analysed CAD together with triglycerides (TG), HDL cholesterol (HDL), hypertension (HT), body-mass index (BMI) and high total cholesterol (HC), and found three groups of SNPs that differed in their pleiotropic patterns of effects across the six traits. They interpreted the three groups of SNPs as accounting for different fractions of CAD heritability via mechanisms related mostly to BMI, HT and HC, respectively, which are risk factors for CAD. In order to further validate their identified groups of SNPs, they assessed whether functional annotation enrichments of nearby genes supported their previous interpretation of the groups. For the BMI cluster, the annotation enrichments implicated the brain in mediating the SNP effects, whereas the visceral adipose tissue was implicated for the HC cluster, consistent with their previous interpretation of the two clusters.

Overall, none of the methods from above leverages SNP annotations and multi-trait summary statistics simultaneously, and none of them finds arbitrary (without strong constraints) functional modules in an unsupervised way using genome-wide SNPs (Table 1.1). By extension, none of them leverages all that information to increase GWAS power. In this thesis, we address these limitations with the development of a new mixture model (MM) that is also fast to fit (Table 1.1). In the next section, we will outline how the MM works to then describe it in more detail in chapter 2.

1.3 Our approach

We cluster SNPs into K groups in an unsupervised way based on two sources of functional similarity between SNPs: their effects on multiple traits (pleiotropic patterns) and their functional annotations. Importantly, like the methods described above, we rely on a linear relationship between genotypes and traits. For N_s individuals, let y_s be a vector of measurements of trait s and $X \in \mathbb{R}^{N_s \times J}$ be a matrix with their centred and standardised genotypes. Then:

$$y_s = X^T \beta_s + \epsilon_s \tag{1.1}$$

Where β_s are the (mean-zero) true effects of the J SNPs on trait s, and ϵ_s are the residuals of the model. We will now explain how we formalise the two types of functional similarity of a cluster of SNPs.

We formalise group-specific pleiotropic patterns as group-specific covariance matrices of multi-trait SNP effects, Σ'_k (to distinguish it from Σ_k, a scaled version that we will use more frequently later). For group k, made of the set of SNPs $\mathcal{J}_k$, letting $\beta_i \in \mathbb{R}^P$ be the vector of effects of SNP i on P traits:

$$\Sigma'_k = \frac{1}{|\mathcal{J}_k|} \sum_{i \in \mathcal{J}_k} \beta_i \beta_i^T \tag{1.2}$$

We formalise group-specific patterns of functional annotations as associations between combinations of annotations and the probabilities of SNPs belonging to a group (i.e. as group-specific 'annotation enrichments'). Let $a_i \in \mathbb{R}^Q$ be a vector with Q annotations for SNP i, and h_i index the group to which SNP i belongs. Then, for some function g_k:

$$\mathrm{p}(i \in \mathcal{J}_k) = \mathrm{p}(h_i = k) = g_k(a_i) \tag{1.3}$$

A mixture model puts the two types of functional properties of clusters together by modelling the joint distribution of the multi-trait SNP effects and the group memberships of SNPs given the SNP annotations. When marginalising the SNP effects from the joint distribution, their probabilities become convex combinations (weighted by the probabilities of belonging to the different groups, the 'mixing weights') of the group-specific distributions of effects (the 'mixed distributions', with covariance matrices Σ'_{h_i}):

$$\mathrm{p}(\beta_i | a_i) = \sum_k \mathrm{p}(h_i = k | a_i) p(\beta_i | h_i = k) \tag{1.4}$$

If we instead marginalise the group memberships, we get the (posterior) probabilities that SNPs belong to the different groups given both the annotations and

the SNP effects, $\mathrm{p}(h_i = k|\beta_i, a_i)$, which can be used to assign individual SNPs to clusters with a quantified certainty.

In practice, we have the SNP annotations but we obviously do not have the true SNP effects. Instead, we have estimates of the true SNP effects, and these estimates generally do not take into account the correlations between SNPs (LD). For a simple, univariate linear regression GWAS:

$$\hat{\beta}_{i,s} = \frac{X_i^T y_s}{N_s \mathrm{var}(X_i)} \tag{1.5}$$

The variance of this estimator depends on the SNP frequency, so a standardised version that removes that dependency is what we refer to as the GWAS z-score:

$$z_{i,s} = \frac{\hat{\beta}_{i,s}}{\sqrt{\hat{\mathrm{var}}(\hat{\beta}_{i,s})}} \tag{1.6}$$

As a consequence, we adapt equation 1.4 to model the distribution of multi-trait GWAS z-scores instead of the multi-trait true effects. Because of LD, the z-scores of nearby SNPs are correlated, and as we will show in the next chapter, this can be introduced into the model by transforming the raw SNP annotations into the so called 'LD annotations'. Intuitively, the LD annotations of a given SNP summarise the overall annotations of correlated SNPs nearby. Specifically, they are a weighted sum of the annotations of SNPs within a window around a focal SNP, where weights are the squared correlations of SNPs with the focal SNP i, $r_{i,j}$. The LD annotations of SNP i are:

$$l_i = \sum_j r_{i,j}^2 a_j \tag{1.7}$$

If we also model the distribution of the estimation error in the z-scores, with covariance matrix Σ_0, and noting that z_i will generally be approximately multivariate normal (as we will explain in the next chapter), the resulting mixture model is:

$$\mathrm{p}(z_i|l_i) = \mathrm{p}(h_i = 0|l_i)\mathcal{N}(z_i|0, \Sigma_0) + \sum_{k>0} \mathrm{p}(h_i = k|l_i)\mathcal{N}(z_i|0, \Sigma_0 + \Sigma_k) \tag{1.8}$$

Where Σ_k is a scaled version of Σ'_k by some constant.

In the next chapter, we will properly derive and expand equation 1.8, and describe how we infer its parameters.

1.4 Thesis overview

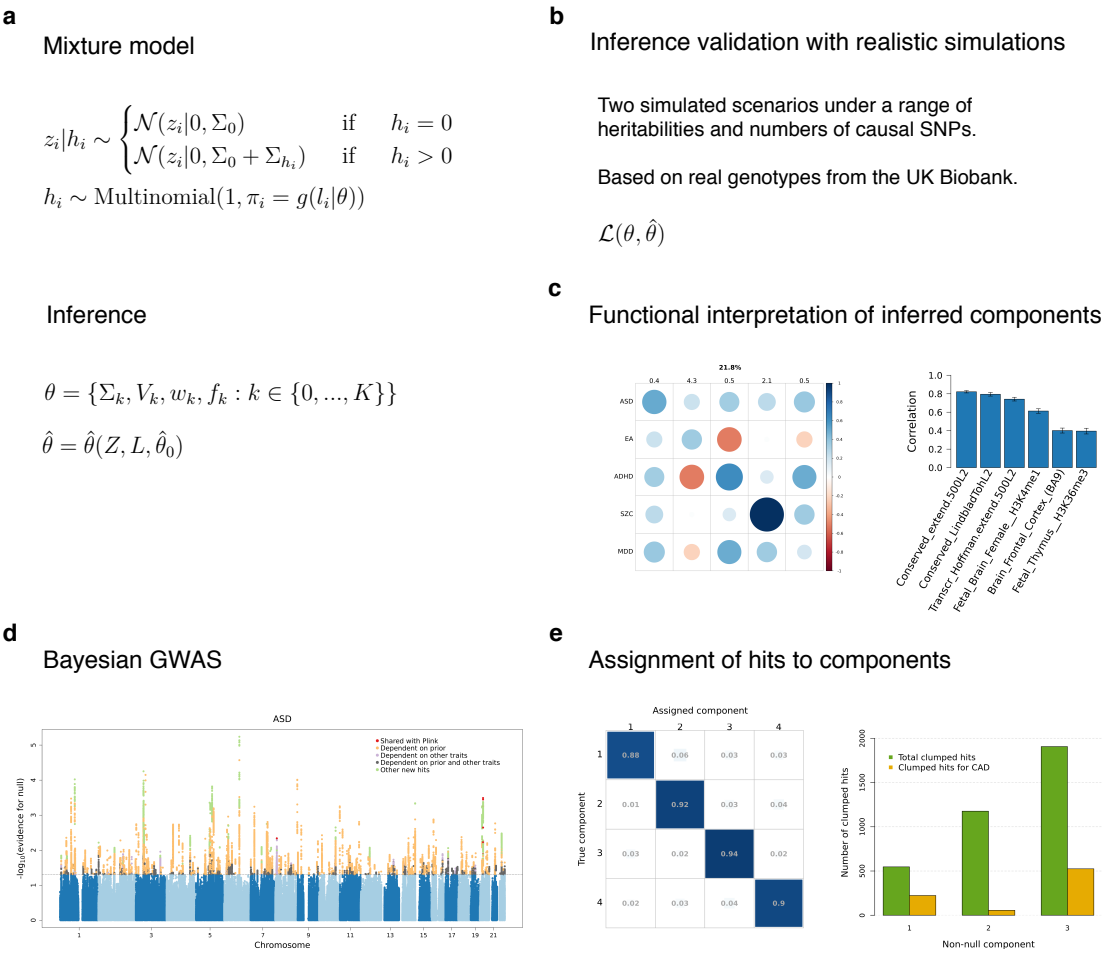


Figure 1.1: Thesis overview. The main content of the thesis can be structured into three blocks: the description of our model and inference procedure (**a**), the validation of our inference procedure (**b**), and the application of our method to study both simulated and real datasets (**c-e**). In **a**, $\hat{\theta}$ refers to our estimator of the true parameters θ, which depends on the z-scores (Z), the LD-annotations (L), and an initial parameter guess ($\hat{\theta}_0$). In **b**, $\mathcal{L}$ is some loss-function that succinctly represents our validation of the inference procedure.

Figure 1.1 summarises the main content of the remaining chapters of this thesis. In chapter 2, we describe the mixture model in detail, derive it from a linear model

of complex traits under some assumptions, and describe our inference procedure. In the same chapter we also provide details about all the analyses that follow inference.

The beginning of chapter 3 is devoted to validating our inference procedure with a range of realistic simulated datasets. As part of the inference validation, we assess our ability to recover the true prior probabilities and genetic correlations of each component. The rest of the chapter is devoted to validating our downstream analysis pipeline, which consists of three steps: (i) interpreting the inferred components functionally, (ii) leveraging the inferred genetic correlations and annotation enrichments to boost GWAS power for a given trait (referred to as 'Bayesian GWAS'), and (iii) assigning GWAS hits to components.

In chapter 4, the same pipeline is used to study two real traits: CAD and ASD. Finally, in chapter 5, we summarise our overall contribution and discuss the limitations and potential improvements of our work.

2 Methods

Contents

2.1 Model

Let $h_i \in \{0, 1, .., K\}$ and θ represent the parameters of the model. Our model for the distribution of multi-trait GWAS z-scores given the LD annotations of SNPs is:

$$p(z_i|l_i,\theta) = \pi_{i,0}\mathcal{N}(z_i|0,\Sigma_0) + \sum_{k=1}^{K}\pi_{i,k}\mathcal{N}(z_i|0,\Sigma_0+\Sigma_k)$$
$$\pi_{i,k} = p(h_i = k|l_i,\theta) = \frac{w_k f_k(l_i^T V_k)}{\sum_{k'=0}^{K} w_{k'} f_{k'}(l_i^T V_{k'})} \tag{2.1}$$

Equations 2.1 describe a mixture of multivariate Gaussian distributions where the 'prior' probability that the multi-trait GWAS z-score of SNP i belongs to the k^{th} component, $\pi_{i,k}$, is a function of the LD annotations of the SNP, l_i. Specifically, $\pi_{i,k}$ is proportional to a linear combination of the LD annotations of SNPs, $l_i^T V_k$; f_k is a non-parametric function that mainly prevents negative values from happening and makes its output have mean 1 (see section 2.3.6 for further details) so that w_k can tune the mean, controlling the expected proportion of SNPs in the component given the LD annotations. Lastly, $w_k f_k(l_i^T V_k)$ needs to be normalised by $\sum_{k'=0}^{K} w_{k'} f_{k'}(l_i^T V_{k'})$ to become a probability.

The multivariate Gaussian distributions of all components have mean zero (due to the approximately random assignment of effect alleles for the GWAS) and share a covariance matrix factor Σ_0 that accounts for estimation error (and, in doing so, for any overlap of GWAS cohorts). If a SNP does not belong to the 'null' component (i.e. if $h_i > 0$), then its z-score is the sum of an error factor, $e_i \sim \mathcal{N}(0,\Sigma_0)$, and an independent genetic factor, $g_{i,k} \sim \mathcal{N}(0,\Sigma_k)$:

$$z_i = g_{i,k} + e_i \;;\; h_i = k \quad \Rightarrow \quad z_i|h_i > 0 \sim \mathcal{N}(0,\Sigma_0+\Sigma_{h_i}) \tag{2.2}$$

To further explain our mixture model, we summarise how we would sample z-scores from it (find a more detailed explanation in section 2.5). For a given number of components K, we first calculate the probability that each SNP belongs to each component, π_i, using their LD annotations; we would then sample component memberships, h_i, from π_i, and use them to sample z-scores from the right distribution (a multivariate normal with a component-specific covariance matrix and mean zero).

2.2 Model derivation

We rely on a linear model for a complex trait:

$$y_s = X^T \beta_s + \epsilon_s \tag{2.3}$$

y_s is a centred random vector of trait s measurements for N_s individuals, and $X \in \mathbb{R}^{N_s \times J}$ is a matrix with the centred and standardised genotypes at J SNPs for the same individuals. β_s are the SNP effects on the trait, and $\epsilon_s \sim \mathcal{N}(0, S)$ is an independent noise term with mean zero and some covariance matrix S.

In practice we do not observe the true effects but rather their GWAS estimates. For a univariate linear regression GWAS (in which the effect of each SNP is estimated independently of the rest), assuming that known confounders have already been regressed out, the estimates have the form:

$$\hat{\beta}_{i,s} = \frac{X_i^T y_s}{N_s \text{var}(X_i)} = \frac{X_i^T y_s}{N_s} \tag{2.4}$$

Equation 2.4 also holds for a univariate logistic regression estimator if the trait measurements are standardised [63]. This is because SNPs are expected to have very small effects, so letting $\tilde{y}_s = y_s / \sqrt{\text{var}(y_s)}$ and v be the proportion of cases, $\hat{\beta}_{i,s}^{\text{logistic}} \approx \frac{\hat{\beta}_{i,s}}{\sqrt{v(1-v)}} = \frac{X_i^T \tilde{y}_s}{N_s \text{var}(X_i)}$. In the case of linear mixed models [64–66], estimates can be regarded as equal to those of linear regression after removing the controlled confounding from the trait measurements, which would at least reduce $\text{var}(y_s)$. The same logic applies to inverse-variance weighted meta-analyses of any of the mentioned GWAS estimates, since the resulting estimate can be interpreted as any of the former but with modified trait variance [55]. We would only need to consider the exact change in variance if we were meta-analysing logistic regression estimates and we wanted to convert the estimated observed-scale heritability to its liability-scale equivalent [63].

For large N_s (and not too rare variants or binary traits[1]), $\hat{\beta}_{i,s}$ will be approximately normally distributed because of the central limit theorem. We also know that, if the variance explained by a single SNP and by confounders is negligible (so that $\text{cov}(X_i, y_s) \approx 0$), then $\hat{\text{var}}(\hat{\beta}_{i,s}) \approx \frac{\text{var}(y_s)}{N_s \text{var}(X_i)} = \frac{\text{var}(y_s)}{N_s}$. In case genotypes were not standardised before the GWAS, we can standardise the GWAS estimates to make their variance independent of SNP frequencies, becoming GWAS z-scores:

$$z_{i,s} = \frac{\hat{\beta}_{i,s}}{\sqrt{\hat{\text{var}}(\hat{\beta}_{i,s})}} \tag{2.5}$$

Expanding equation 2.5:

$$z_{i,s} \approx \frac{\sqrt{N_s}}{\sqrt{\text{var}(y_s)}} \frac{X_i^T (X\beta_s + \epsilon_s)}{N_s} = \frac{\sqrt{N_s}}{\sqrt{\text{var}(y_s)}} \left(\sum_j r_{i,j} \beta_{i,s} + e_{i,s} \right) \tag{2.6}$$

Where $r_{i,j} = \frac{X_i^T X_j}{N_s}$ and $e_{s,i} = \frac{X_i^T \epsilon_s}{N_s}$. The distribution of $z_{i,s}$ remains normal with mean zero, and consequently the joint distribution of $\{z_{i,s} \in \{1, ..., P\}\}$ is multivariate normal with a covariance matrix to be defined. In order to derive the covariance of $z_{i,s}$ and $z_{i,t}$, we assume the following:

$$\begin{cases} \text{cov}(\beta_{i,s}, e_{i,s}) = 0 \\ \frac{1}{N_s} X^T X = \frac{1}{N_t} Y^T Y \\ \text{cov}(\beta_{i,s}, \beta_{i,t}) = \sum_k p(h_i = k | a_i) \Sigma'_{k,s,t} \approx \sum_k a_i^T V_k \Sigma'_{k,s,t} \end{cases}$$

Where $Y \in \mathbb{R}^{N_t \times J}$ is a second genotypes matrix for N_t individuals. The first two assumptions say that the true linear effects are uncorrelated with the estimation errors, and that LD is the same across GWAS cohorts. The second assumption has two components: first, each SNP belongs to one of several groups and its multi-trait effects has a group-specific covariance matrix; second, the probability that a SNP belongs to a group is a function of its functional annotations. Because for a

[1]BOLT-LMM [64] requires variants to have frequencies greater than 0.1% and binary traits to have prevalence greater than 10% for its association test to be well calibrated [64, 65]. In chapter 5, we discuss a necessary extension of the simulations work presented in chapter 3 in order to assess the behaviour of our MM when variant frequencies or trait prevalence go below these thresholds. The suggested extension will help us to better interpret the validity of the new hits found by our Bayesian GWAS, which we will introduce later.

reasonably small number of traits we expect most SNPs to have no effects on any trait ($p(h_i = 0) \gg \sum_{k>0} p(h_i = k)$), the function relating functional annotations and prior probabilities should be approximately linear. Letting V_k be the linear effects of the annotations on the prior probabilities for belonging to the k^{th} component:

$$p(h_i = k|a_i) \approx a_i^T V_k \tag{2.7}$$

Using the assumptions from above:

$$\begin{aligned}
\mathbb{E}[z_{i,s}z_{i,t}] &= \sqrt{\frac{N_s N_t}{\text{var}(y_s)\text{var}(y_t)}} \left(\sum_j r_{i,j}^2 \mathbb{E}[\beta_{j,s}\beta_{j,t}] + \mathbb{E}[e_{i,s}e_{i,t}] \right) \\
&= \sqrt{\frac{N_s N_t}{\text{var}(y_s)\text{var}(y_t)}} \sum_j r_{i,j}^2 \sum_k a_j^T V_k \Sigma'_{s,t,k} + \frac{N_o \delta_{s,t}}{\sqrt{N_s N_t}} \\
&= \sqrt{\frac{N_s N_t}{\text{var}(y_s)\text{var}(y_t)}} \sum_k l_i^T V_k \Sigma'_{s,t,k} + \Sigma_{s,t,0}
\end{aligned} \tag{2.8}$$

Where $l_i = \sum_j r_{i,j}^2 a_j$ are the 'LD annotations' of SNP i, N_o is the sample overlap between X and Y, and $\delta_{s,t}$ is the phenotypic correlation of traits y_s and y_t. If we exclude the SNPs with the most extreme values of LD (for which $\sum_k l_i^T V_k$ will be large), we can approximate equation 2.8 as a mixture of covariances: $\mathbb{E}[z_{i,s}z_{i,t}] \approx \pi_{i,0}\Sigma_{s,t,0} + \sum_k \pi_{i,k}\Sigma_{s,t,k}$, where $\Sigma_{s,t,k} = c\,\Sigma'_{s,t,k}$ for some constant c. That is, with probability $\pi_{i,k}$ SNP i belongs to the k^{th} component and, if so, its multi-trait GWAS z-scores follow a multivariate Gaussian distribution with covariance Σ_k. For P traits:

$$p(z_i|l_i) = \pi_{i,0}\mathcal{N}(0, \Sigma_0) + \sum_{k=1}^{K} \pi_{i,k}\mathcal{N}(0, \Sigma_k + \Sigma_0) \tag{2.9}$$

Note that we are incorporating LD into the model via component memberships and not via the z-score distributions. This may lead in practice to the introduction of additional components to capture large z-scores in cases when the prior probabilities are not proportionally as large. For example, SNPs may be in weak LD with a strong causal SNP, or multiple causal SNPs may be located in the same locus without a proportional enrichment in the relevant functional annotations. Alternatively, such large z-scores may be absorbed by fewer components and no extra components may be strongly favoured by the likelihood of the model.

For later, it will be useful to express z_i as the sum of two independent terms: one that depends on the true SNP effects, $g_{i,k} \sim \mathcal{N}(0, \Sigma_k)$, and another being estimation error $e_i \sim \mathcal{N}(0, \Sigma_0)$. Concisely:

$$\begin{cases} \text{if} \quad \mathrm{h}_i = 0: \quad z_i = e_i & \Rightarrow \quad z_i | h_i = 0 \sim \mathcal{N}(0, \Sigma_0) \\ \text{if} \quad \mathrm{h}_i > 0: \quad z_i = g_{i,h_i} + e_i & \Rightarrow \quad z_i | h_i > 0 \sim \mathcal{N}(0, \Sigma_0 + \Sigma_{h_i}) \end{cases} \tag{2.10}$$

2.3 Inference

2.3.1 Overview

We want to infer $\theta = \{\Sigma_k, V_k, w_k, f_k : k \in \{0, ..., K\}\}$ ($\theta = \{C, V, w, f\}$ for brevity), given a number of components K, using the observed data $\mathcal{D} = \{Z, L\}$, where $Z_{i,s} = z_{i,s}$ and $L_{i,} = l_i$ (so $Z \in \mathbb{R}^{J \times P}$ and $L \in \mathbb{R}^{J \times Q}$). We do that by iteratively optimising an approximation to the log-likelihood function of the model given an initial parameter guess. As summarised in Algorithm 1, after (1) pre-processing the data, (2) choosing the number of components and (3) initialising the parameters, we (4-16) sequentially optimise the objective function with respect to each of the parameters until either we reach N iterations or converge (defined as a marginal improvement of the likelihood for the last 30 iterations). Both at iteration 100 and after the last iteration, a special step is carried out that consists mostly of removing redundant annotations (those that are only associated with prior probabilities because they are correlated with another annotation with a stronger association) and stabilising the annotation coefficient estimates.

I will now describe each of the mentioned steps in more detail, starting with the definition of the approximate likelihood function that we optimise.

2.3.2 Approximate weighted log-likelihood

If the z-scores where independent, the log-likelihood function would be:

$$l^*(\theta|\mathcal{D}) = \sum_i \log \sum_k p(h_i = k|l_i, \theta) p(z_i|h_i = k, \theta)] \tag{2.11}$$

Algorithm 1: Inference

```
Pre-process
Choose K
Initialise parameters
for i ∈ 1 : N do
    if (i < 100 and i=0 (mod 5)) or i=0 (mod 20) then
        Update V and f
    Update w
    Update C
    Update posterior
    if converged = T or i ∈ {100, N} then
        Forward regression
        Update annotations
        Update V
        Component-specific masking
        if converged = T then
            break
```

That would be the true likelihood if we chose independent SNPs for inference, but doing that would also significantly reduce our coverage of relevant annotations unless performed very carefully. Instead, we use non-independent SNPs and weight the contribution of each SNP to the likelihood from equation 2.11 based on the extent to which they are correlated with other SNPs from the sample. Specifically, we weight the log-likelihood of a SNP by the inverse of the LD score of the SNP calculated using HapMap3 SNPs (the ones used for inference as an heuristic for selecting SNPs with well imputed genotypes). This is the same approach used by LD score regression and seems to work well in practice [67]. The approximate weighted log-likelihood (from now on we will refer to it simply as the likelihood, for convenience) is:

$$l(\theta|\mathcal{D}, v) = \sum_i v_i \log \sum_k p(h_i = k|l_i, \theta) p(z_i|h_i = k, \theta)] \tag{2.12}$$

Where $v_i = 1/l_{i,base\_hm3}$. The intuition is that correlated SNPs can be thought of as 'copies' of each other to the extent to which they are in LD with each other (e.g. they would be exact copies if they were perfectly correlated), so we can count how many copies a SNP has in the sample (with its LD score) and then sample

one of them with uniform probability. If we let $s_i \sim \text{Ber}(v_i)$ represent the sampling outcome for SNP i (so we sample SNPs independently of what we have already sampled), we can see that our weighted log-likelihood is the expected log-likelihood for independent SNP with respect to the distribution of all possible samples:

$$l(\theta|\mathcal{D}, s) = \mathbb{E}_s[\sum_i \log(\sum_k p(h_i = k|l_i, \theta)p(z_i|h_i = k, \theta))^{s_i}] \tag{2.13}$$

2.3.3 Pre-processing

The pre-processing of the input data $\mathcal{D}$ consists of several independent steps:

- SNP filtering: discard SNPs with base LD scores above their 0.99 percentile or with missing values (for either the z-scores, LD annotations or weights).
- SNP alignment: align Z and L based on SNP rsids.
- Annotation screening: in practice we may have hundreds or thousands of potentially informative annotations, even if only a small subset may be needed and not redundant (not very correlated). To avoid overfitting and to reduce computation and memory cost, although not strictly needed, we screen the annotations via what we call 'forward stratified LD score regression'. This consists of first running stratified LD score regression [33] for each of the traits and discarding the annotations with p-values larger than 0.05, keeping only annotations with some evidence of heritability enrichment. The second step consists of, for each trait, iteratively selecting among the remaining annotations the one that best linearly predicts the squared z-scores while controlling for the already accepted ones, and accept it if improves the prediction by more than 1% (this is often called forward regression, see Algorithm 2). The result is, for each trait, a set of not very redundant annotations (thanks to the forward regression) with some evidence for being enriched in heritability. We take the union and carry on with these for inference.

Algorithm 2: Forward regression

$candidates \leftarrow 1 : Q$
$accepted, r^2_{\text{previous}}, r^2_{\text{current}} \leftarrow \emptyset$
while $|candidates| > 0$ **do**
  for $i \in candidates$ **do**
    $r_i^2 \leftarrow \text{OLS}(z_s^2 \sim L_{,(accepted,i)}|weights)$
    $r^2_{\text{current}} \leftarrow (r^2_{\text{current}}, r_i^2)$
  if $\max r^2_{current} > 1.01 \times \max r^2_{previous}$ **then**
    $accepted \leftarrow (accepted, \underset{i \in \text{candidates}}{\text{argmax}}\ r^2_{\text{current},i})$
    $r^2_{\text{previous}} \leftarrow r^2_{\text{current}}$
    $candidates \leftarrow 1 : Q \notin accepted$
  else
    $candidates \leftarrow \emptyset$

2.3.4 Choosing the number of components

Before final inference, we choose the number of components (K) based on two criteria:

1. 5-fold cross-validation (CV): we partition all the sorted SNPs from the training dataset into five contiguous blocks of equal size and, for different numbers of components (e.g. one to six), we do inference using four different blocks each time and calculate the mean per-SNP likelihood of the left-out block using the inferred parameters. We then average the five per-SNP likelihoods obtained for each number of components and plot the results.

2. Component redundancy: in practice we see that cross-validation can favour 'extra' (not explicitly simulated) components that are almost perfect combinations (meaning correlation > 0.95) with positive weights of other components and have large z-score variances. The extra components are most likely explained by SNPs that are in LD with several causal SNPs.

Based on the two criteria from above, we choose the greatest number of components that does not include redundant ones and that precedes the saturation of the CV per-SNP likelihood curve. That is, we first take the greatest number

of components that significantly increases the per-SNP likelihood with respect to having one fewer component, say five components. We then check whether any of the five components can be explained by a linear combination (with positive weights) of the other four. If no component is redundant, then we choose five components for inference; if one component is redundant but fitting the model with four components does not result in any redundant component, then we choose four components for inference.

2.3.5 Parameter initialisation

- V and f: we initialise them indirectly by setting $f_k(l_i^T V_k) = 1$ for every SNP and every component.

- w: it is initialised as 0.95 for the null component, and $0.05/K$ for the rest. Together with $f_k(l_i^T V_k) = 1$, it means that our prior belief is that 95% of the SNPs are null.

- C: for the null component, each element is initialised with the intercepts of bivariate LD-score regressions [52] (as many as different pairs of traits). For the non-null components, we cluster the z-scores in a transformed space that allows us to separate components based on their means (with a variant of k-means), and calculate the covariance matrix of the SNPs in each cluster. We then add to them the null covariance matrix and use the result to initialise the covariance matrices of the non-null components. In further detail, in order to initialise Σ_k for $k > 0$, we first transform the z-scores so that each new dimension is the product of the z-scores for a pair of traits (for example, for traits A and B, the transformed z-scores would have three dimensions: z_A^2, z_B^2 and $z_A z_B$. More generally, for P traits, the transformed z-scores would have $P(P+1)/2$ elements). This is to decrease the overlap of components by making them have different means (as opposed to all having mean zero). We then standardise (i.e. make their variance equal to 1) the transformed z-scores and keep only the 10% that are the most far away from the origin in order

to decrease the overlap of different components. Then, we sub-sample 2000 SNPs (for computational ease, since this number seems sufficient in practice), project them to the surface of a unit-radius sphere (still in $\mathbb{R}^{P(P+1)/2}$) to remove radial variation, and cluster them with a more robust variant of k-means called 'partitioning around medoids' (k-means is confounded by the remaining overlap of z-scores from different components). Finally, we remove cluster members that are very far away from the cluster medoid (which tend to be from a different component) and use the remaining SNPs to calculate the covariance matrix of each cluster using their original z-scores (in $\mathbb{R}^P$).

2.3.6 Parameter updates

I refer to the current parameters (to update) as θ^{old}, and recall from section 2.3.2 that v_i are the likelihood weights. All the updates are guaranteed to increase or leave intact the likelihood of the model.

- w: we optimise the negative log-likelihood with respect to w_k via the Newton-Raphson algorithm [68], with a diagonal approximation of the Hessian:

$$w_k^{new} = w_k^{old} - \frac{\frac{\partial}{\partial w_k}[-l(w_k|\mathcal{D}, v, \theta^{old})]}{\frac{\partial^2}{\partial w_k^2}[-l(w_k|\mathcal{D}, v, \theta^{old})]} \tag{2.14}$$

Differentiating once or twice with respect to w_k gives:

$$\begin{cases} \dfrac{\partial}{\partial w_k}[-l(w_k|\mathcal{D}, v, \theta^{old})] = \sum_i v_i \dfrac{1}{w_k}(\pi_{i,k} - r_{i,k}) \\ \dfrac{\partial^2}{\partial w_k^2}[-l(w_k|\mathcal{D}, v, \theta^{old})] = \sum_i v_i \dfrac{1}{w_k^2}(\pi_{i,k}^2 - r_{i,k}^2) \end{cases} \tag{2.15}$$

And plugging the above into equation 2.14 leads to the following iterative update for w_k within the Newton-Raphson algorithm:

$$w_k = w_k^{old}\left(1 + \frac{\sum_i v_i(\pi_{i,k} - r_{i,k})}{\sum_i v_i(\pi_{i,k}^2 - r_{i,k}^2)}\right) \tag{2.16}$$

Because $\frac{w_k}{\sum_{k'} w_{k'}} = \frac{w_k c}{\sum_{k'} w_{k'} c}$, we implicitly constrain the length of w by forcing $w_0 = 1$ so that the solution to the optimisation problem is unique. The likelihood function with respect to w is not convex, so we reject any update that decreases the likelihood.

- V and f: we propose an update for both V and f for each component and only accept the combination that results in the greatest increase of the log-likelihood, not updating them at all if the proposal decreases the likelihood. The proposal for V_k is the ordinary least-squares solution to the following weighted linear regression problem:

$$\sqrt{v_i} p(h_i = k | z_i, l_i, \theta^{old}) \sum_{k'=0}^{K} w_{k'}^{old} f_{k'}^{old}(l_i^T V_{k'}^{old}) = \sqrt{v_i} l_i^T \beta_k + \epsilon \tag{2.17}$$

This is because $\mathbb{E}_{z_i|l_i}[p(h_i = k|z_i, l_i, \theta^{old})] = \pi_{i,k}$, whose unnormalised version should be approximately linear on l_i (see the model derivation in section 2.2).

f_k refines the linear predictions $l_i^T V_k$ by modelling the residuals ϵ (mainly to remove negative values) and then mean-normalising the result (so that the proportion of SNPs can be tuned by w), and forms part of the same update proposal as V. In more detail, f_k starts by dividing SNPs into bins based on the linear predictions $\sqrt{v_i} l_i^T V_k$, and then maps the mean prediction of each bin to the mean target. It then linearly interpolates the predictions (extrapolating for the smallest and largest prediction values, and mapping negative values to half the value of the mean of the nearest bin), so $f_k(l_i^T V_k)$ can be thought of as a piece-wise linear regression. Based on the model derivation, f_k should be approximately the identity function for most SNPs, and only refine the predictions when needed to prevent negative values.

V and f are not updated in every iteration to speed up the algorithm, and instead they are updated every five iterations until we reach iteration 100, after which they are updated every 20 iterations.

- C: in order to update Σ_k, we indirectly optimise $l(C|\mathcal{D}, v, \theta^{old})$ by optimising its following lower bound:

$$\begin{aligned} Q(C|\mathcal{D}_c, v, \theta^{old}) &= \mathbb{E}[l_c(C|\mathcal{D}_c, v, \theta^{old})] \\ &= \mathbb{E}[\sum_i v_i \sum_k \log[p(g_i, e_i, h_i|C, \theta^{old})]^{\mathbb{I}(h_i=k)}] \end{aligned} \tag{2.18}$$

$l_c(C|\mathcal{D}_c, v, \theta^{old})$ is the negative log-likelihood function of the 'complete data', $\mathcal{D}_c = \{g_i, e_i, h_i : i \in \{1, ..., J\}\}$ (i.e. assuming that we observed all the otherwise hidden variables), and equation 2.18 is the so called 'auxiliary function' in the context of the EM algorithm, so our update of Σ_k can be regarded as the 'M step' for Σ_k within an EM. This function coincides with the log-likelihood when using the parameters of the previous iteration and therefore, because it is a lower bound of the latter, optimising it guarantees an increase in the log-likelihood unless we are already at a maximum. [68].

Expanding $l_c(C|\mathcal{D}_c, v, \theta^{old})$, we can note that only part of it depends on Σ_k, which we call $l_c^*(C|\mathcal{D}_c, v, \theta^{old})$:

$$\begin{aligned} l_c(C|\mathcal{D}_c, v, \theta^{old}) &= \sum_i v_i \sum_k \mathbb{I}(h_i = k)[\log[p(g_i, e_i|h_i = k, C, \theta^{old})] + \log[p(h_i = k|\theta^{old})]] \\ &= \sum_i v_i \sum_k \mathbb{I}(h_i = k)\log[p(g_i, e_i|h_i = k, C, \theta^{old})] + \text{cnt.} \\ &= \frac{1}{2}\sum_i v_i(\mathbb{I}(h_i = 0)[\log|\Sigma_0|^{-1} - e_i^T\Sigma_0^{-1}e_i] \\ &\quad + \sum_{k>0} \mathbb{I}(h_i = k)[\log|\Sigma_k|^{-1} - (g_{i,k} + e_i)^T\Sigma_k^{-1}(g_{i,k} + e_i))] + \text{cnt.}' \\ &= l_c^*(C|\mathcal{D}_c, v, \theta^{old}) + \text{cnt.}' \end{aligned} \tag{2.19}$$

Combining equations 2.19 and 2.18, and differentiating with respect to Σ_0^{-1} and Σ_k^{-1}:

$$\begin{cases} \dfrac{\delta Q(C|\mathcal{D}_c, v, \theta^{old})}{\delta \Sigma_k^{-1}} = \mathbb{E}[-\dfrac{1}{2} \sum_i v_i \mathbb{I}(h_i = k)(\Sigma_k + g_{i,k} g_{i,k}^T)] \\ \\ \dfrac{\delta Q(C|\mathcal{D}_c, v, \theta^{old})}{\delta \Sigma_0^{-1}} = \mathbb{E}[-\dfrac{1}{2} \sum_i v_i \sum_k \mathbb{I}(h_i = k)(\Sigma_0 + e_i e_i^T)] \end{cases} \tag{2.20}$$

Setting equations 2.20 to zero, we get:

$$\begin{cases} \Sigma_k = \dfrac{\sum_i v_i r_{i,k} \mathbb{E}[g_i g_i^T]}{\sum_i v_i r_{i,k}} \\ \\ \Sigma_0 = \dfrac{\sum_i v_i \sum_k r_{i,k} \mathbb{E}[e_i e_i^T]}{\sum_i v_i \sum_k r_{i,k}} \end{cases} \tag{2.21}$$

We can calculate the expectations in equation 2.21 after marginalising g_i and e_i from the distributions of $(g_i + e_i, g_i)$ and $(g_i + e_i, e_i)$, respectively. For $k > 0$, the mean and covariance of $g_i | z_i, h_i = k$ are:

$$\begin{cases} \mu_{i,k} = \Sigma_k (\Sigma_k + \Sigma_0)^{-1} z_i \\ \Omega_k = \Sigma_k - \Sigma_k (\Sigma_k + \Sigma_0)^{-1} \Sigma_k \end{cases} \tag{2.22}$$

And the mean and covariance of $e_i | z_i, h_i = k$ are:

$$\begin{cases} \mu'_{i,k} = \Sigma_0 (\Sigma_k + \Sigma_0)^{-1} z_i \\ \Omega'_k = \Sigma_0 - \Sigma_0 (\Sigma_k + \Sigma_0)^{-1} \Sigma_0 \end{cases} \tag{2.23}$$

For $k = 0$, $g_i = 0$ and the mean and covariance of $e_i | z_i, h_i = 0$ are:

$$\begin{cases} \mu'_{i,k} = z_i \\ \Omega'_k = 0 \end{cases} \tag{2.24}$$

Using the above results and the law of total variance, we can rewrite the updates from equation 2.21 as:

$$\begin{cases} \Sigma_k = \dfrac{\sum_i v_i r_{i,k}(\Omega_k + \mu_i \mu_i^T)}{\sum_i v_i r_{i,k}} \\ \Sigma_0 = \dfrac{\sum_i v_i r_{i,0} z_i z_i^T + \sum_{k>0} v_i r_{i,k}(\Omega'_k + \mu'_i(\mu'_i)^T)}{\sum_i v_i \sum_k r_{i,k}} \end{cases} \tag{2.25}$$

2.3.7 Special steps

A special step takes place at both iteration 100 and the final iteration of the inference algorithm (Algorithm 1).

First, forward regression (Algorithm 2) is used to identify redundant annotations in a component-specific manner. The difference between this forward regression step and the forward regression used to screen annotations (as part of pre-processing) is that here we regress scaled component-specific posterior probabilities on the LD annotations (as in Equation 2.17) as opposed to regressing squared z-scores (during pre-processing we obviously did not have posterior probabilities yet).

We then permanently discard annotations that are redundant for every component and update V and f again. If an annotation is redundant for only some of the components, we remove its effects in a component-specific way by setting its annotation coefficient to zero in those components. This is especially important in the last iteration in order to get stable annotation coefficients, which will then allow us to calculate prior probabilities outside of the training sample (e.g. for a Bayesian GWAS, see section 2.6.3).

2.4 Computational and memory requirements

For the current, not yet optimised implementation, each iteration of the EM takes $\mathcal{O}(JK(Q+P)+KP^3)$ time and requires $\mathcal{O}(J(K+Q+P(P+1)/2))$ memory (it mostly has to store five $J \times K$ matrices, two $J \times Q$ matrices, and two other matrices that together amount to $J \times P(P+1)/2$). Note that for the current intended use of the method (which is approximately $K, P < 20$ and $Q < 300$), J (the number of SNPs) is the largest term and both memory and computational time grow linearly on it.

We can take the analyses from chapter 4 as representative examples: inference for the ASD dataset (with $P = 5$, $K = 2$ and $J = 907,970$) took 15 minutes plus 45 extra minutes spent screening the 500 annotation candidates from which 49 were selected for inference; inference for the CAD dataset (this time $P = 3$, $K = 4$ and $J = 1,085,237$) took 90 minutes, with 50 extra minutes again spent screening the same 500 annotations, from which 57 were kept for inference. In terms of memory, the first scenario required approximately 0.8Gb (after screening the annotations, which would otherwise require more memory if a big matrix of candidate annotations was loaded at once), and the second scenario 1.2Gb.

2.5 Simulations

In Figure 2.1 we summarise how we simulate multi-trait z-scores and build LD annotations. The first step of the simulation process consists of downloading ∼500 raw annotations for ∼10m 1000G SNPs [4] from [34] (the annotations of SNP i are represented by $\mathbf{a}_i$ in the diagram). These include 396 Roadmap annotations, 53 annotations based on GTEx (called 'specifically expressed genes' (SEG) annotations in the paper), and 66 'baseline' annotations (meaning not tissue or cell-type specific). The SEG annotations cover 53 tissues (Table A.2) and tell whether SNPs are within 100kb of a gene that is specifically expressed in each of the tissues, which they defined as being one of the top 10% genes with a greater expression in a tissue than in tissues that belong to other tissue categories (e.g. brain frontal cortex vs. tissues outside the brain). The Roadmap annotations cover 87 tissues (Table A.2) and tell whether SNPs overlap any of the six epigenetic marks measured by the Roadmap project (different tissues have different numbers of available marks, but the six possibilities are: DNase, H3K9ac, H3K27ac, H3K4me1, H3K4me3 and H3K36me3). The baseline annotations (summarised in Table Tables A.1) are based on both SEG and Roadmap annotations (by taking either unions or sums across tissues[2]), and on

[2]One of the baseline annotations is the union of all the SEG annotations (i.e. it annotates whether SNPs are within a 200kb window centered any gene specifically expressed in any tissue), and 12 are based on Roadmap (by taking either the union or sum of the same marks across tissues).)

additional annotations as well (e.g. broad HMM-based annotations (hidden states) based on lower-level baseline epigenetic annotations [41]).

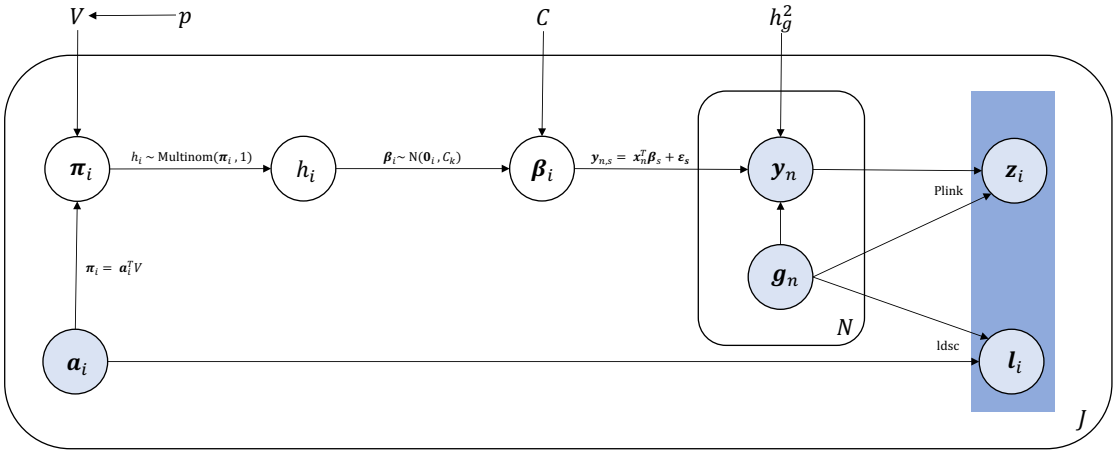


Figure 2.1: Summary of the simulation procedure. Letters represent either parameters (if not encircled) or variables that are either downloaded or simulated (encircled). Coloured circles denote that the corresponding variables can be observed in a real-life setting (e.g. genotypes). The darker band in the right emphasises that only the z-scores and the LD annotations need to be observed for our method to work (i.e. they alone are 'the simulated dataset'). Arrows represent dependencies during the simulation process (e.g. we need V and $\boldsymbol{a}_i$ to calculate $\boldsymbol{\pi}_i$). To ease the interpretation of the diagram, vector variables are written in bold text, and repetitions of a variable for either multiple SNPs or UK Biobank individuals are represented as plates around variables with repetition numbers on the bottom right corner (e.g. a vector of LD annotations is calculated J times, once for each SNP). i indexes SNPs, n indexes individuals, and $\boldsymbol{\beta_s} \in \mathbb{R}^J$ is a vector with the effects of J (standardised) SNPs on trait s. h_g^2 denotes additive heritability, and it is not a vector because all the traits are simulated with the same heritability.

The downloaded raw SNP annotations are then used to get LD annotations ($\mathbf{l}_i$) and to simulate prior probabilities for SNPs belonging to each of the non-null components ($\boldsymbol{\pi}_i$ in the diagram). To get the LD annotations, the raw annotations are transformed by the ldsc software[3] using the recommended settings (mainly a 1 centiMorgan window and 1000G SNPs). The simulated prior probabilities for a SNP are simply a linear function of the raw annotations (which takes into account the proportion of causal SNPs in each component, p. See the text around equation 2.7).

The simulated prior probabilities are then used to sample component memberships (h_i), and the memberships are then used to sample effect sizes for P traits ($\boldsymbol{\beta}_i$) from the corresponding multivariate normal distribution (with mean

[3] `https://github.com/bulik/ldsc`

zero and covariance C_k). The simulated effect sizes are then used to simulate the genetic component of traits as linear functions of the genotypes of 343,969 White British individuals from the UK Biobank [10] ($\boldsymbol{x}_n$ represents the genotype of the n^{th} individual). Noise is then added to match a desired heritability h_g^2, getting the simulated traits $\boldsymbol{y}_n$.

Finally, a univariate linear regression GWAS is done for each SNP using Plink [69]. The standardised linear regression estimates are the simulated z-scores $(\boldsymbol{z}_i)$[4].

2.6 Downstream analyses

With downstream analyses we mean all the analyses that follow inference. These include:

- The interpretation of the inferred parameters as the functional properties of functional modules (2.6.1).
- The assignment of SNPs to the inferred components (2.6.2).
- Bayesian GWAS (2.6.3).
- The identification of the source of evidence that drives each hit (2.6.4).

2.6.1 Functional interpretation of components

The inferred parameters of the model capture the functional signatures of the functional modules, that is, the average functional properties of 'the SNPs in them'. The inverted commas come from the fact that, in practice, generally SNP cannot be unambiguously assigned to one of the modules, so group compositions are rather expected compositions based on posterior probabilities (which capture our confidence about the different alternatives). The point is that, independently of whether we want to use the posterior probabilities to commit to unambiguous ('hard') SNP

[4]Lino Ferreira, a member of our group, did the simulation steps that involve dealing with real genotypes, namely simulating the traits using the simulated SNP effect sizes and the linear regressions with Plink.

assignments when possible (as we will also be doing), it is still possible to characterise the functional properties of the modules using the inferred parameters of the model.

Interpreting the inferred parameters for the mentioned purpose reduces to transforming and visualising them in a meaningful way. To find the annotations that are associated with the probability of belonging to a given component (i.e. the 'annotation enrichments' of a given component), we first discard redundant (correlated) annotations by running a component-specific forward linear regression[5] (Algorithm 2) and then plot the Spearman[6] correlation of the selected LD annotations with the prior probabilities of the component. We add standard errors to the plot based on 100 bootstrap samples[7].

To get the average pattern of SNP effects across traits (and invariant to the heritability of the different traits) we can transform the genetic component of the inferred covariances (Σ_k) into genetic correlation matrices. We can get the component-specific heritability per SNP, up to an unknown constant, by normalising the diagonals of Σ_k by the trait-specific GWAS sample sizes. We can then calculate the relative heritability explained by each component by multiplying the per-SNP heritabilities by the expected number of SNPs in each component (the mean posterior probability for belonging to the component, $\sum_i r_{i,k}/J$), and then normalising by the maximal resulting value. In chapters 3 and 4, we will be plotting genetic correlation matrices with the diagonals replaced with relative per-SNP heritabilities.

2.6.2 Assignment of SNPs to components

We can use the posterior probabilities over component memberships to assign SNPs to the components. We consider that we are confident enough to make an assignment if the posterior probability for a SNP belonging to one of the components is greater than 0.8. That is:

[5]Regressing the scaled posterior probabilities for the component on the LD annotations, as we do when updating the annotation coefficients as part of the EM.

[6]I use the Spearman correlation because we do not expect the relationship to be exactly linear.

[7]Each bootstrap sample is the result of partitioning the (location-based sorted) SNPs from the training dataset into 200 blocks of equal size and then sampling blocks with replacement 200 times.

$$\hat{h}_i = \underset{k}{\text{argmax}}\ \mathbf{1}\{r_{i,k} > 0.8\} \tag{2.26}$$

2.6.3 Bayesian GWAS

Posterior probability of association

Bayesian GWAS, as mentioned previously, is how we refer to the calculation of the probabilities that SNPs (both within and outside the training dataset) are associated with a given trait, given the observed data (i.e. multi-trait z-scores and LD annotations) [70].

If we knew that a SNP belonged to the null component, then we would be sure that the SNP has no effect on any of the traits included in our dataset. If instead we knew that a SNP belonged to one of the non-null components, then we would be confident that the SNP has an effect on at least one of the traits, but not necessarily on a specific focal trait. To quantify the evidence that a SNP is associated with a focal trait conditional on belonging to a specific non-null component, we calculate the probability that the sign of the z-score for the trait has been correctly inferred (e.g. that our z-score estimate for the trait is positive when the true effect of the SNP is also positive). To do that, recall from section 2.3.6 that we are modelling a (multi-trait) z-score as the sum of two random components, given that the SNP belongs to the k^{th} component: a genetic one, $g_{i,k} \sim \mathcal{N}(0, \Sigma_k)$ if $k > 0$ and $g_i = 0$ otherwise, and a non-genetic one, $e_i \sim \mathcal{N}(0, \Sigma_0)$,

$$z_i = g_{i,k} + e_i \tag{2.27}$$

We can first work out the conditional distribution of the genetic component given the observed z-score thanks to having inferred Σ_k and Σ_0 (as explained in the inference section from this chapter):

$$g_{i,k}|z_i \sim \begin{cases} \mathcal{N}(\mu_{i,k}, \Omega_k) & \text{if} \quad k > 0 \\ \delta(0) & \text{if} \quad k = 0 \end{cases} \tag{2.28}$$

Where $\mu_{i,k} = \Sigma_k(\Sigma_k + \Sigma_0)^{-1} z_i$ and $\Omega_k = \Sigma_k - \Sigma_k(\Sigma_k + \Sigma_0)^{-1}\Sigma_k$.

Then we can calculate the probability that $g_{i,k,s}|z_i$ (where s indexes traits and therefore elements in the $g_{i,k}$ vector) has a different sign than the z-score for trait s, $z_{i,s}$. For example, if the observed z-score for a given trait was positive but $\mu_{i,k} = 0$, then $g_{i,k,s}|z_i \sim \mathcal{N}(0, \Omega_{s,s})$ and there would be a 50% chance that the genetic component is positive and therefore that the sign of the z-score is correct. More generally, the probability that the sign of a z-score is correct is the probability of its genetic component having the same sign as the z-score (all this for a specific trait). Note that because $\mu_{i,k,s}$ (the mean of the genetic component given the z-scores) always has the same sign as $z_{i,s}$, we can take its absolute value so that any positive value of the genetic component imply that the z-score has the correct sign.

Putting everything together (that is, weighting the evidence that SNPs belong to each component and, conditioning on that, that the SNPs are associated with a focal trait of interest), the probability of association of SNP i with trait s (that model M_1 is correct as opposed to the null one, M_0), is:

$$p(m_{i,s} = M_1|z_i, l_i, \theta) = \sum_{k>0} r_{i,k}\mathcal{N}(g_{i,k,s} > 0;\ |\mu_{i,s,k}|, \Omega_{k,s,s}) \tag{2.29}$$

To calculate these probabilities for SNPs outside of the training sample (e.g. 1000G SNPs), the only thing to take into account is that the SNPs cannot have extreme base LD scores (so, when analysing the 1000G SNPs, we discard the top 1% with the greatest base LD scores) and that f_k may have to be slightly adjusted in the extremes (if the new $l_i^T V$ are out of the previously seen range during inference).

New independent hits

To get an estimate of the number of independent new loci associated to a trait, we clump hits with Plink (which iteratively takes the top remaining hit and removes other hits within 500kbp that have an LD of $r^2 > 0.1$ with the top hit[8]) and then check whether there is any hit found by Plink within 500kb of the hits that result from the clumping. New independent hits are then defined as independent hits in loci that are associated to a trait by our mixture model but not by Plink.

[8]Specifically, we used version 1.9 and the following flags: –clump –clump-p1 0.05 –clump-kb 500e3 –clump-r2 0.1.

Standard and stratified ROC curves

To assess our ability to correctly classify SNPs as null or causal based on our posterior probabilities of association, the first step consists of defining sets of truly causal and truly null SNPs. To do that, we first take the ∼10m 1000G SNPs for which we originally simulated effect sizes and discard rare variants and the 1% with the greatest LD scores, resulting in ∼6m SNPs (including SNPs both inside and outside the training dataset). We then define causal SNPs for a trait of interest as SNPs with non-zero effect sizes for the trait, and null SNPs as being at least 0.5Mb far from any causal SNP for the trait. We then create two balanced (equally sized) classes of SNPs, one of truly causal SNPs and one of truly null SNPs.

We assess classification performance with a combination of the true positive rate (TPR, the probability that a causal SNP is classified as causal) and the false positive rate (FPR, the probability that a null SNP is classified as causal). A 'Receiver Operating Characteristic' (ROC) curve is a visual way of assessing how the TPR and FPR change as a function of the threshold for significance (because loosening the threshold increases both the TPR and the FPR), and the area under the curve (AUC) is a convenient summary of that relationship. We can investigate the relation between true SNP effects and classification performance by splitting causal SNPs into bins with increasingly large true effects and calculating a ROC curve (or the AUC) for each bin, which is sometimes referred to as 'stratified' ROC curve (or AUC).

2.6.4 Identification of the driving source of evidence for hits

We can easily adapt equation 2.29 to cases where we lack part of the data, like not having z-scores for some of the non-focal traits or not having LD annotations. In fact, we intentionally 'hide' different parts of the data after inference to see how that affects the probabilities of association, ultimately allowing us to identify the sources of evidence that drive each hit. Specifically, we hide either the z-scores from non-focal traits or the LD annotations.

Hiding the z-scores of non-focal traits means that we do not use them to calculate the likelihood of the z-scores under different components. In this way, they do not

contribute to the posterior probabilities over component memberships and therefore cannot be used to allocate SNPs into the different components. We also do not use them to calculate the mean and covariance of $g_{i,k}|z_i$ (which is similar to setting all genetic correlations to zero). Hiding the LD annotations means that we average the prior across SNPs and set the prior of every SNP to the result (so, the prior probability for belonging to a given component becomes constant and equal to the expected proportion of SNPs in the component given the LD annotations that we used for inference: $\pi_i = \pi_{i' \neq i} = \sum_j \pi_j / J$).

If a SNP were a hit but stopped being a hit after hiding only the z-scores from other traits, it would mean that the other traits are driving the hit. Similarly, if a hit no longer remained a hit after hiding only its LD annotations, then we would know that the annotations were the main source of evidence for pushing its probability of association above the threshold. If a hit were immune to us hiding either the z-scores or the LD annotations, then there would be two possibilities: either the SNP has strong evidence coming from both sources (so that even if we remove one, the other source is sufficient), or it has a strong z-score for the focal trait (in which case a simple linear regression would detect the hit too, e.g. using Plink). Finally, if a hit depends on both sources of evidence, it means that none of the sources suffices in isolation and it is instead the combination of them that are driving the hit.

We will use these ideas in the simulations chapter to emphasise the source of boosted GWAS power in relation to other GWAS methods (section 3.5), and in in chapter 4 to classify new hits for CAD and ASD.

3 Simulations

Contents

Before modelling datasets of interest, we need to be confident that we can infer the true parameters well enough and that we can use them to recover the true functional modules. Here we do that by simulating a range of realistic datasets and using them to test our inference procedure and our downstream analysis pipeline. we described how we do the simulations in detail in the methods chapter above (section 2.5).

3.1 Description of the simulated datasets

3.1.1 Overview

Each simulated dataset consists of two matrices: a matrix of LD annotations and a matrix of multi-trait GWAS z-scores, both for about 10m 1000G SNPs [4].

What makes the simulated datasets realistic is a combination of the chosen trait heritabilities and numbers of causal SNPs [27], the odds-ratios of the annotations when predicting SNP memberships to causal groups, and the use of real genotypes (from the UK Biobank) and real SNP annotations.

I simulated a total of 40 datasets: 20 for a scenario with three traits and two groups of causal SNPs ('scenario 1'), and 20 for another scenario with five traits and four groups of causal SNPs ('scenario 2'). For each of the two scenarios, the 20 datasets are the result of combining four different heritability values (0.05, 0.1, 0.2 and 0.4) with five different numbers of causal SNPs (5000, 1000, 10000, 20000 and 40000). Each causal group of SNPs from each scenario has a specific genetic correlation pattern and is enriched in specific annotations (Figures 3.1 and 3.2).

Scenario 1 (Figure 3.1) can be thought of as capturing two mechanisms that confer risk for a simplified coronary artery disease (CAD): one that regulates the accumulation of LDL in arteries (involving the liver, small intestine, arteries and lymphocytes), and another one that regulates blood pressure (involving the kidney, adrenal gland, lung, heart and arteries again) [21]. Scenario 2 (Figure 3.2) can be thought of as capturing four mechanisms that confer risk for a simplified standing height: one controls the secretion of growth hormone (involving the hypothalamus and the pituitary gland), which is in turn regulated by sex hormones secreted by the testis, adrenal gland and ovary (the second mechanism); the liver then processes the growth hormone and produces IGF-1 (third mechanism), which is then processed by target tissues such as skeletal muscle and blood cells (fourth mechanism) [71].

There are three similarities between the two scenarios (see section 2.5 for details about the simulation procedure):

- Every trait has the same heritability (so having fewer causal SNPs will result in greater average effects for all SNPs).
- They both have a trait with uniformly distributed heritability across components, with the rest of the traits having most or all of their heritability concentrated into a single component (making them effectively less polygenic).

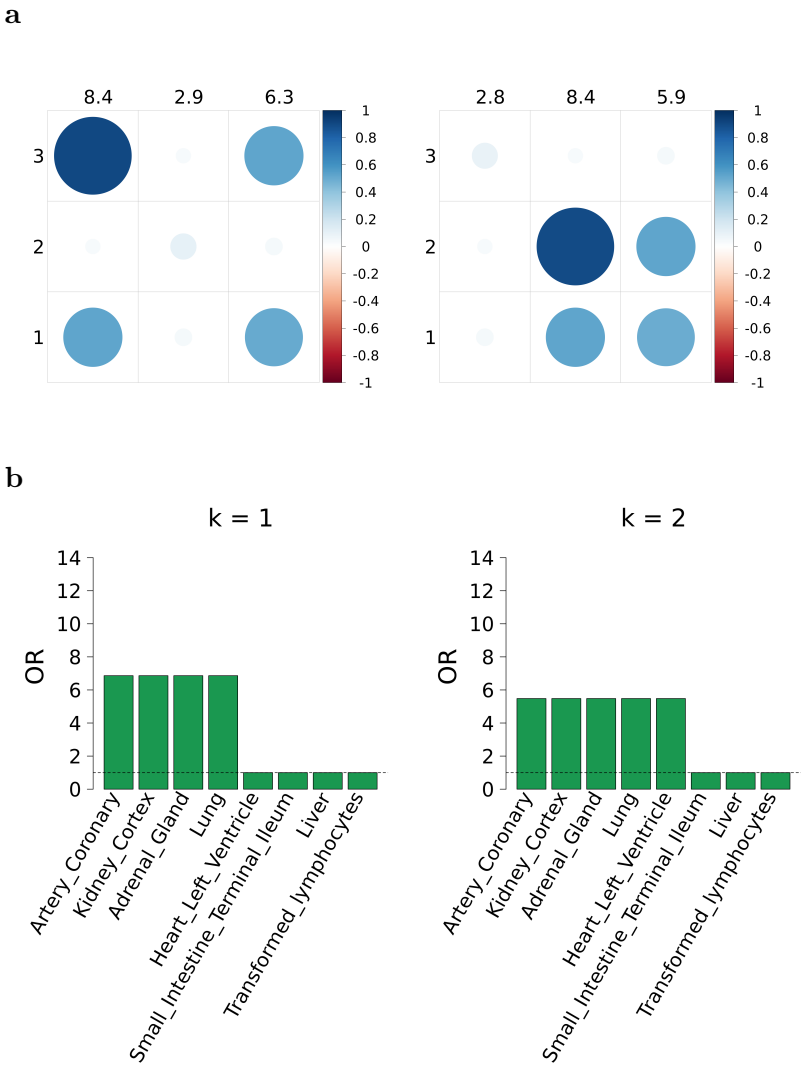


Figure 3.1: Simulated truth for scenario 1. **a** Shows the genetic correlation and heritability pattern of each component. The off-diagonals display genetic correlations, and the diagonals the normalised per-SNP heritability of each component (normalised by dividing by the maximal value across components and traits, which depends on the heritabilities of traits). Row names are the trait names (three in this case), and column names are the variances of the z-scores that truly belong to each component. **b** Shows the odds ratios (OR) of the raw annotations in relation to the prior probabilities (an odds ratio of 1 for a component means that the prior probability for belonging to the component does not depend on that annotation).

- Every non-null component has the same number of causal SNPs (e.g. for two non-null components and a total of 20,000 causal SNPs, each component would have 10,000 causal SNPs).

Throughout this chapter, we will often use two specific datasets (out of the 40), one from each scenario, as examples of simulation properties or results. These are

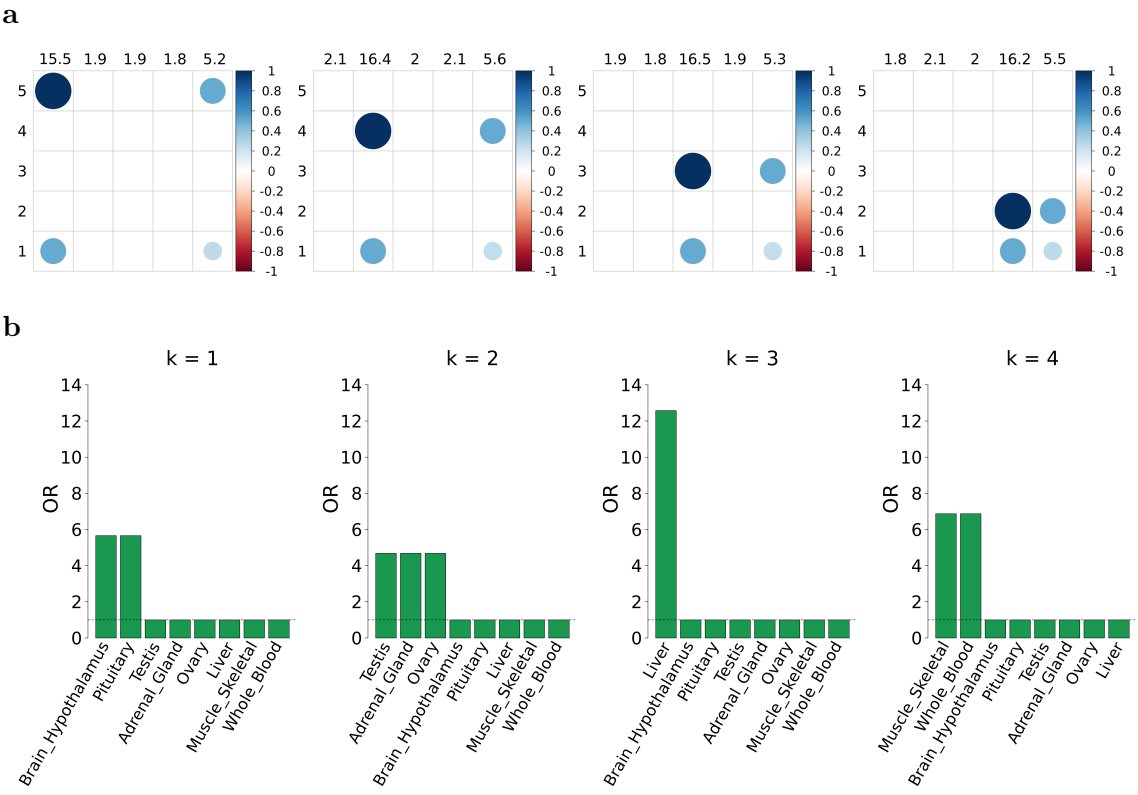


Figure 3.2: Simulated truth for scenario 1. **a** Genetic correlation and heritability pattern of each component. **b** Odds ratios (OR) of the raw annotations in relation to the prior probabilities. (See Figure 3.1 for further details).

the datasets with $h_g^2 = 0.2$ and $20,000$ causal SNPs in total (so $20,000/K$ SNPs per non-null component). we will refer to these datasets as the 'selected datasets'.

3.1.2 Quality control

In this section we show the output of key steps of the simulation process (described in detail in Chapter 2) in order to verify that the simulations are as expected and to provide some additional intuition about the simulated datasets.

The first step in the simulation process is simulating prior probabilities that SNPs belong to the different non-null components, which we show in Figure 3.3 (for the two selected scenarios). Most SNPs have prior probabilities almost equal to zero (not exactly zero because of the small non-zero intercept in the generating linear function) and the range of probabilities is very small (because the simulated 20,000 causal SNPs are only a small fraction of the total number of SNPs). The prior

probabilities take as many different values as combinations of relevant annotations in the 1000G SNPs (for example, in Figure 3.3 b, the prior probabilities for the 4^{th} component (purple bars) take only one of two possible values because only liver has an $OR > 1$ for that component (see Figure 3.2): so either the SNP has the annotation or it does not).

After sampling component memberships from the simulated priors, we can get a sense of the distribution of SNPs throughout the genome by calculating the distances between consecutive causal SNPs. Figure 3.4 shows the cumulative mass function (CMF) of such distances for datasets with different numbers of causal SNPs. The greater the number of causal SNPs, the smaller the range of distances. Most causal SNPs are close enough to each other to violate the assumption of one or few causal SNP per LD block, which makes the simulated datasets challenging in that sense. The selected datasets, which have 20,000 causal SNPs, have only about 10% of their causal SNPs further away than 500kb, and about 40% further away than 100kb.

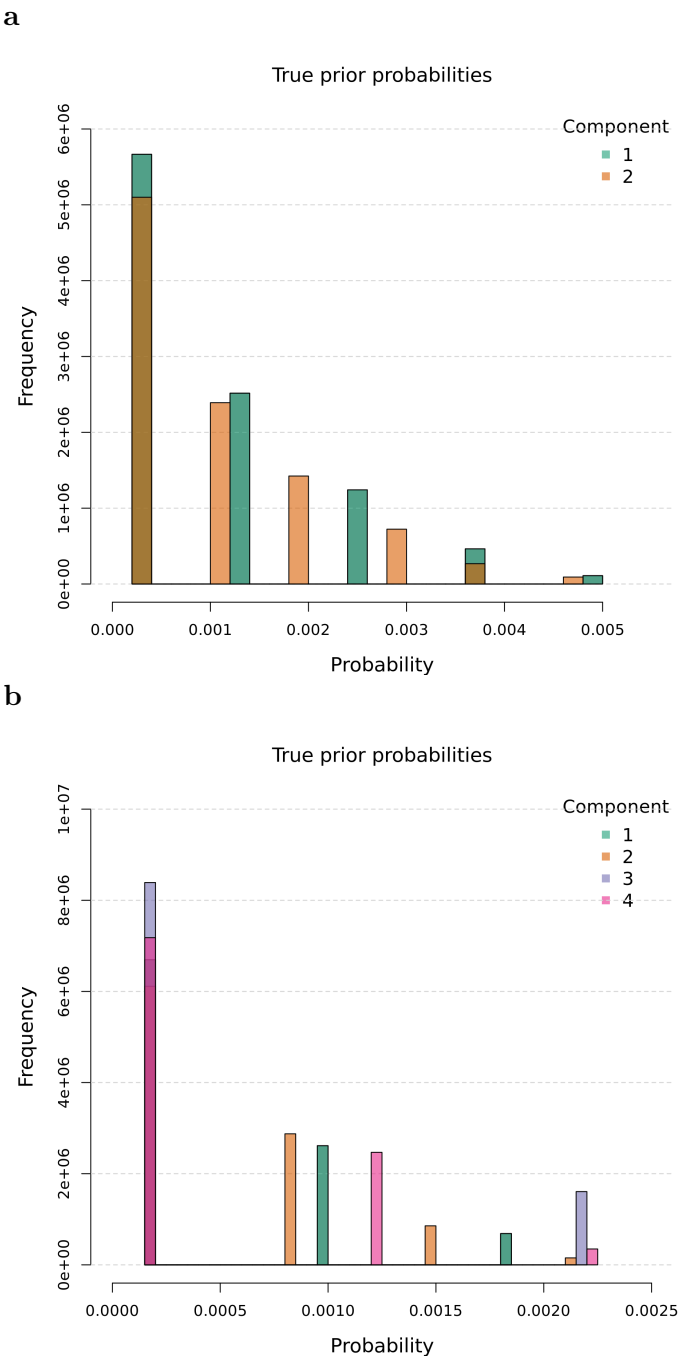
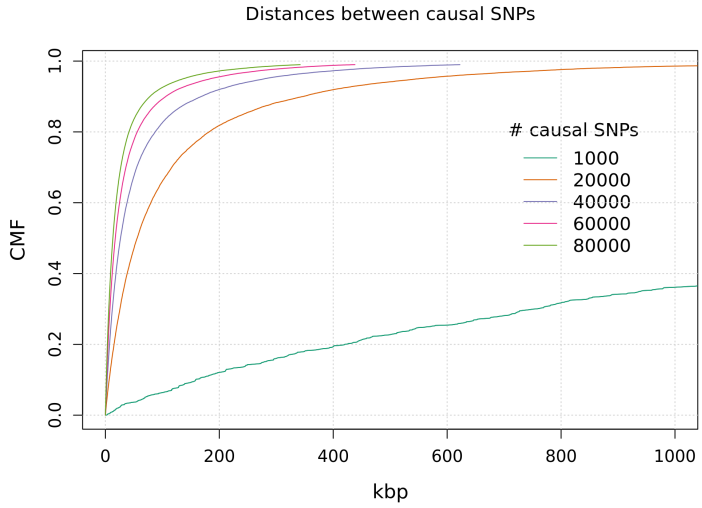


Figure 3.3: Histogram of the simulated priors for the selected datasets from simulated scenarios 1 (**a**) and 2 (**b**).

a



b

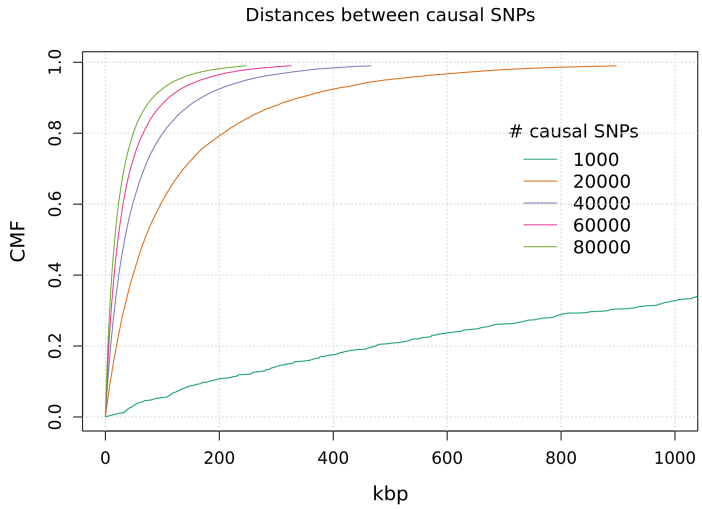


Figure 3.4: Cumulative mass function (CMF) of distances between consecutive causal SNPs (in kilobase pairs, kbp) as a function of the number of simulated causal SNPs. **a** Is for the simulated scenario 1, and **b** is for the simulated scenario 2.

Once we know which SNPs are causal and the component they belong to, we simulate their true effects on the traits. Figures 3.5 and 3.6 show, for selected scenarios 1 and 2, the simulated SNP effects for pairs of traits. In scenario 1, SNP effects have positive variance for every trait and every component (i.e. the heritability of every trait comes from every component), and the first two traits are almost genetically uncorrelated (see the cross shape of the upper plot in Figure 3.5, in accordance with Figure 3.1).

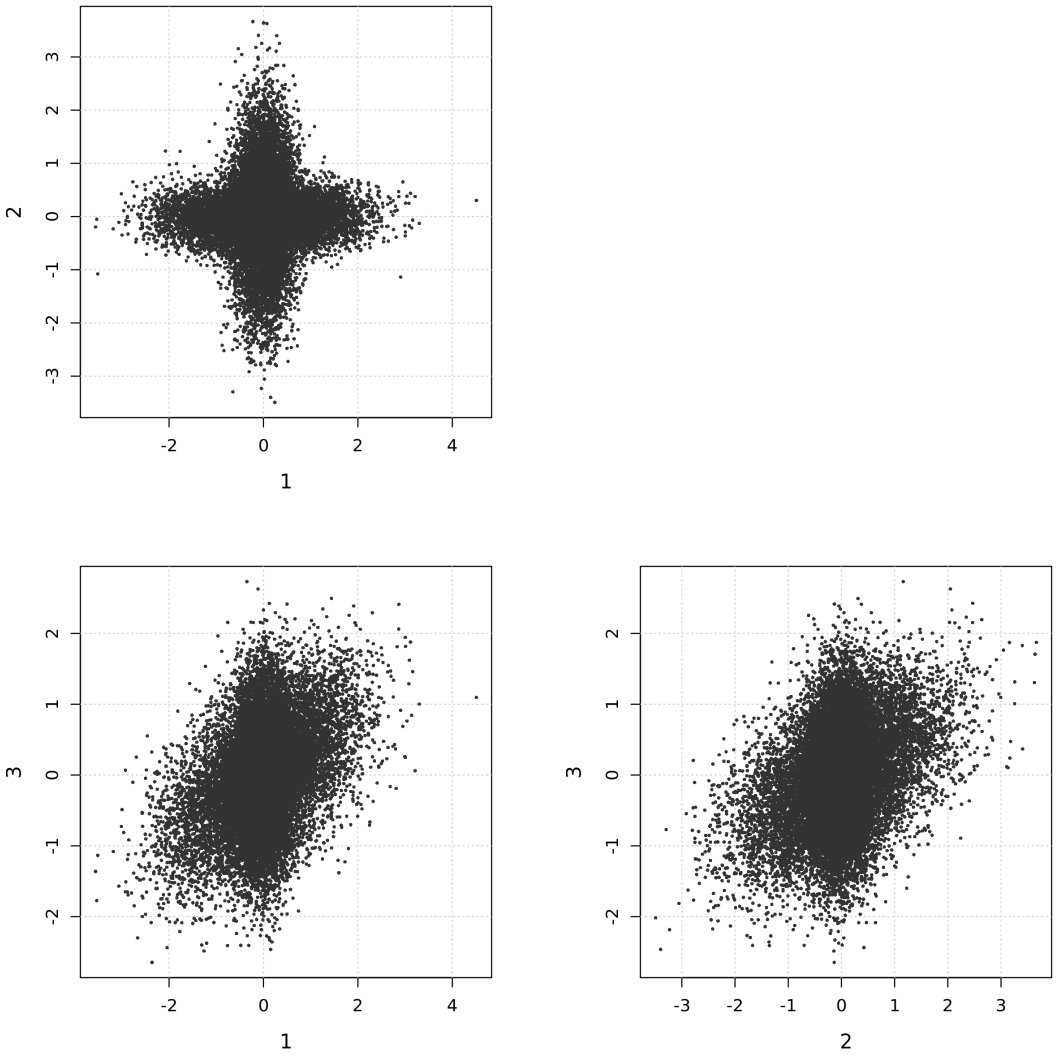


Figure 3.5: Simulated SNP effects for pairs of traits, from the selected dataset from scenario 1 ($h_g^2 = 0.2$ and 20,000 causal SNPs).

In scenario 2, only trait 5 has some heritability stemming from every component,

with the rest of the traits having all their genetic variance concentrated in a single component (Figure 3.6). Trait 5 is genetically correlated with the rest of the traits, but traits 1 to 4 are genetically uncorrelated between them.

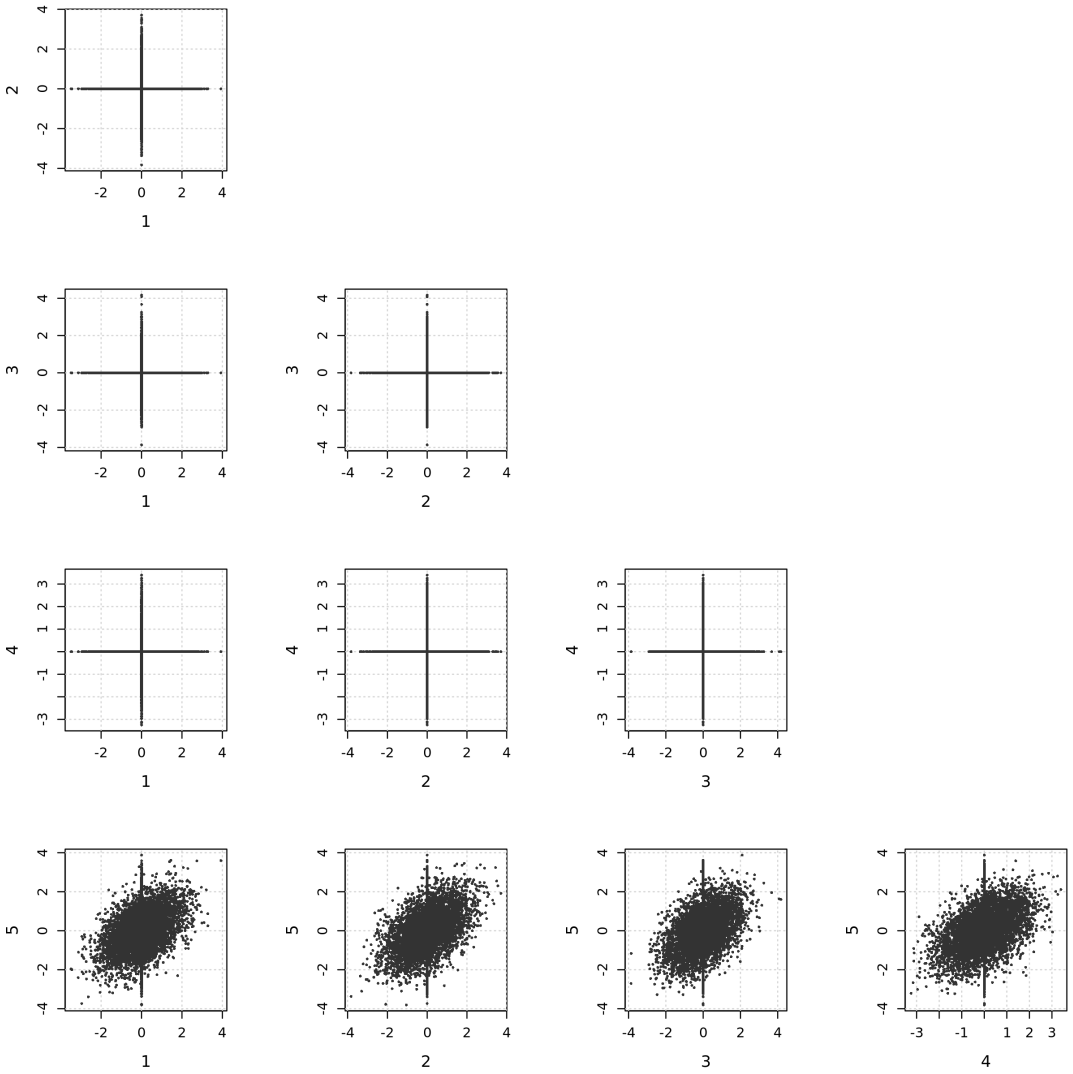


Figure 3.6: Simulated SNP effects for pairs of traits, from the selected dataset from scenario 2 ($h^2_g = 0.2$ and 20,000 causal SNPs).

The simulated SNP effects are combined with 343,969 genotypes of White British individuals from the UK Biobank to simulate traits with a desired heritability (0.2 for the selected datasets). We then do a linear regression GWAS with Plink and standardise the estimates to get z-scores. Figures 3.7 and 3.8 show, for the two selected datasets, the GWAS z-scores for pairs of traits. Components that previously

had zero variance for SNP effects on some trait, now have positive variance for the z-score equivalent due to estimation error (compare for example the upper plots from Figures 3.6 and 3.8). Genetic correlations are still perceivable in the z-score plots.

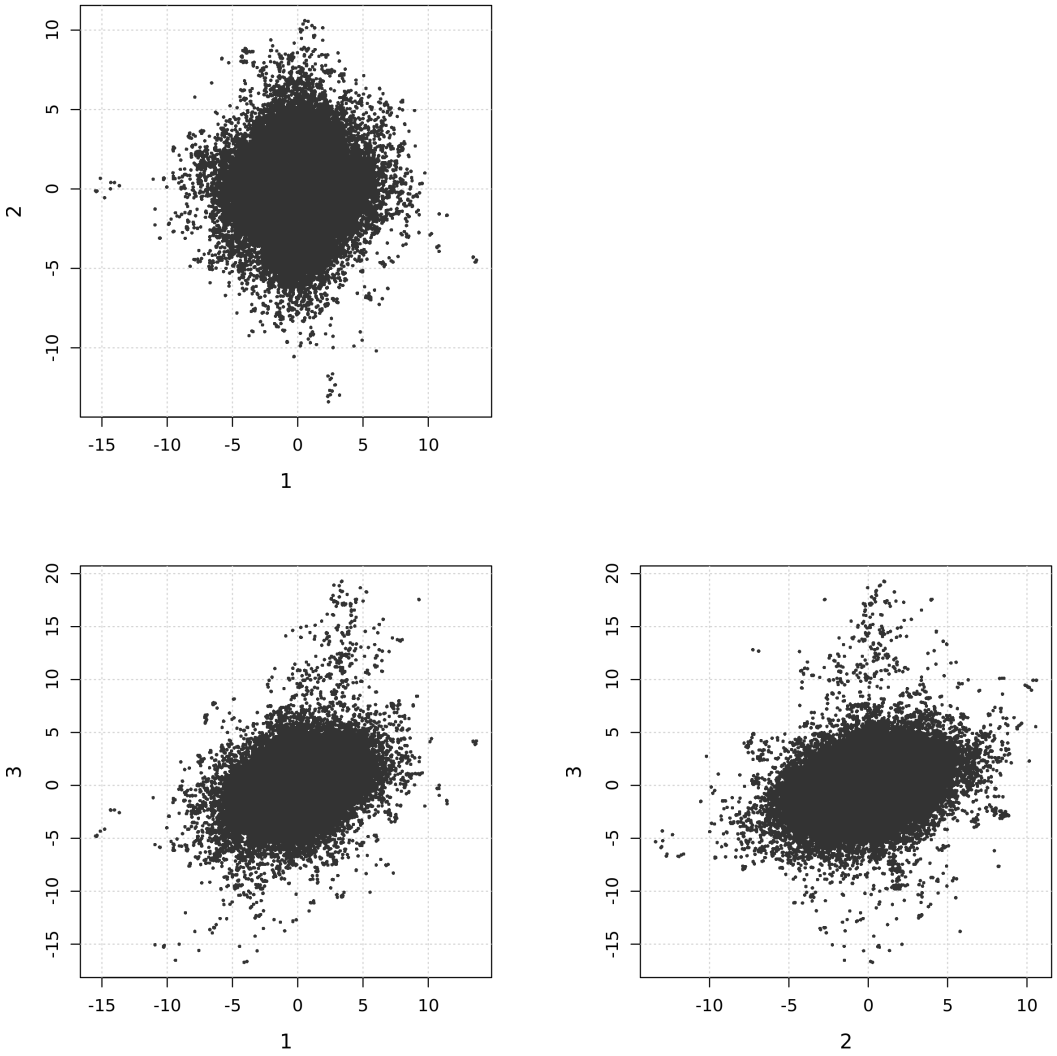


Figure 3.7: Simulated GWAS z-scores for pairs of traits, from the selected dataset from scenario 1 ($h_g^2 = 0.2$ and 20,000 causal SNPs).

A way of summarising the amount of genetic signal in the dataset (the deviation of the distribution of the z-scores from their distribution under the null hypothesis used for the GWAS) is comparing the distribution of the p-values with a uniform distribution. Figure 3.9 shows, for the selected datasets, the quantile-quantile (QQ) plot for the $-\log_{10}$ transformed p-values returned by Plink (based on a t-test). We

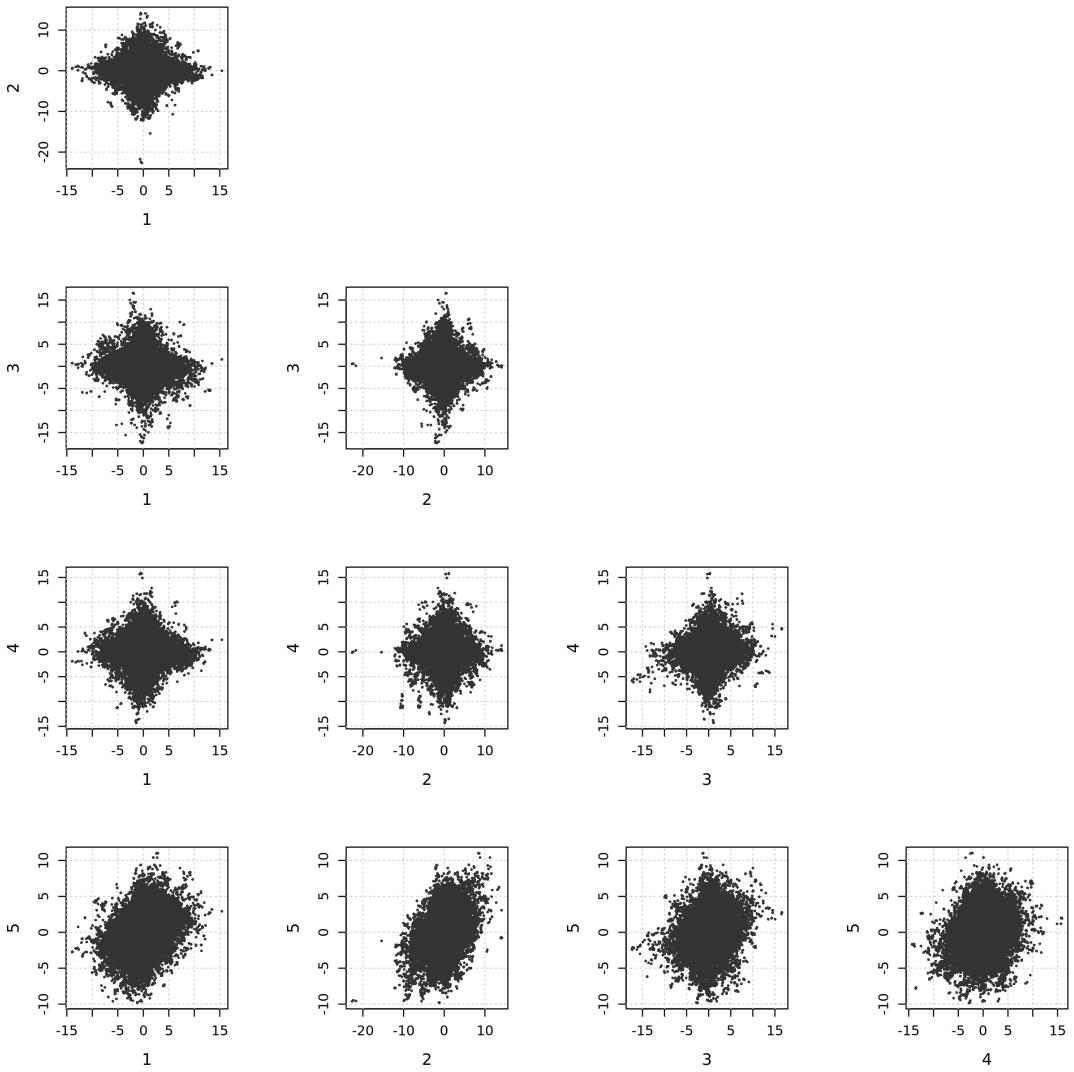


Figure 3.8: Simulated GWAS z-scores for pairs of traits, from the selected dataset from scenario 2 ($h_g^2 = 0.2$ and 20,000 causal SNPs).

later inferred the variance of the z-scores in the null component (equivalent to the intercepts of LD score regression) to be all <1.1, suggesting that the observed genetic signal comes from true genetic effects and not from strong confounding.

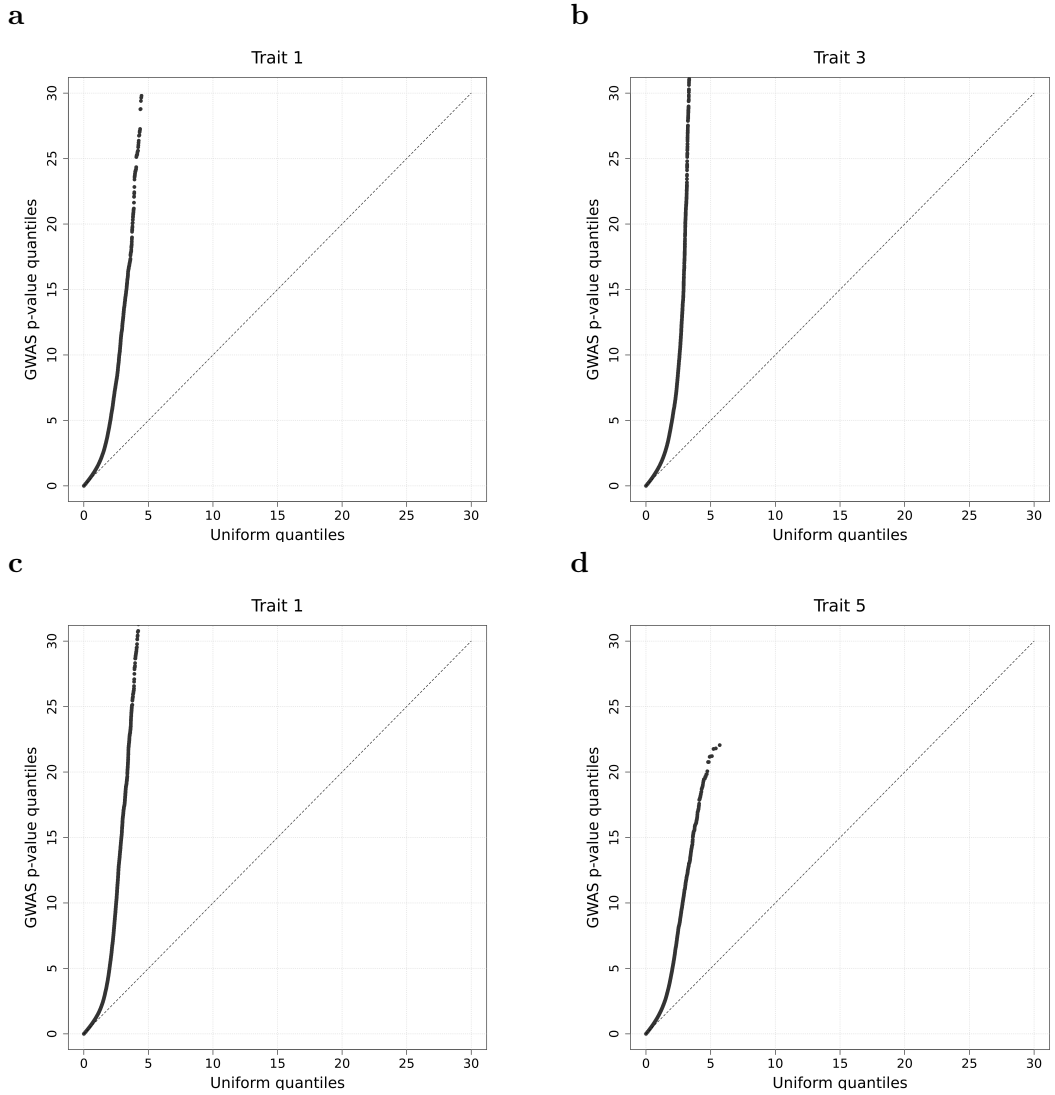


Figure 3.9: QQ-plots for the $-\log_{10}$ p-values of the selected datasets from scenarios 1 (**a-b**) and 2 **(c-d**). We compare the observed distribution of p-values (from the GWAS) with a uniform distribution.

3.2 Validation of the inference procedure

To test our inference procedure, here we investigate our ability to infer (i) the genetic correlation and heritability pattern of each component, and (ii) the prior distribution over component memberships of SNPs both inside and outside the training sample (all together spanning the 1000G SNPs). In the next section, in the context of functionally interpreting the inferred components, we also assess our ability to identify the relevant annotations for each component and we provide specific examples of inferred genetic correlations and (relative) heritabilities.

The first step when attempting to assess inference consists of matching the true and inferred components (by permuting their indexes). We do that based on the correlation of their priors, matching each true component to the inferred component with the most correlated prior probabilities (we use as true prior the LD annotations times the true annotation coefficients, LV, and we use the Spearman correlation to compare it to the inferred priors). For example, if true component 2 has prior probabilities that correlate the most with the inferred component 3, then we match the former with the latter.

Genetic covariances are inferable up to an unknown constant, so we instead assess our ability to infer the genetic correlation matrices and the normalised per-SNP heritability of each component (which are invariant to the total heritability of the traits). For each component, we vectorise the inferred genetic correlations and normalised heritabilities, and then calculate the L_2 error of the vector and take the mean across vector elements. We then average the mean error across components and across 100 bootstrap samples (obtained by partitioning the location-sorted HapMap3 SNPs into 200 equally-sized blocks and re-sampling blocks with replacement 100 times) to get a mean error for the dataset. In Figure 3.10 we show the mean error for each of the 40 simulated datasets. The error bars are the standard deviation of the mean errors calculated with the 100 bootstrap samples. The mean error is small in all cases, and increases with the number of causal SNPs, as it is expected due to greater violations of the assumption of one or few causal SNP per LD block. Also, the greater the heritability, the smaller the error due to less ambiguous separations of SNPs into different components. The mean errors are slightly lower on average for the simulated scenario 2 (Figure 3.10 **b**) probably because the components are slightly less overlapping (most genetic correlations are exactly zero, which is not the case for scenario 1 as shown in Figures 3.1 and 3.2).

To assess our ability to recover the true prior probabilities, we used the Spearman correlation of the inferred priors with the true unnormalised priors (like for matching true and inferred components). For downstream applications (mainly the Bayesian GWAS), we will need prior probabilities for SNPs beyond the training sample, so

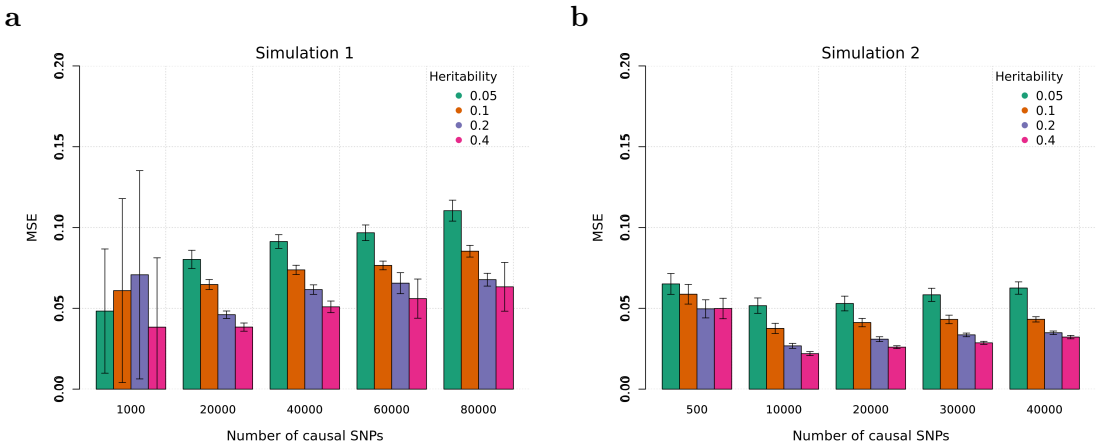


Figure 3.10: Mean error of the genetic correlation and normalised of heritability estimates. **a** is for simulated scenario 1, and **b** for simulated scenario 2. The error bars are the standard deviations of the mean errors based on 100 bootstrap samples.

we calculate and assess prior probabilities for all the 1000G SNPs (∼10m SNPs) excluding the 1% with the greatest LD scores. As before, we average the correlations across components and 100 bootstrap samples to get a point correlation for a dataset. we show the results for each of the 40 simulated datasets in Figure 3.11. The least polygenic datasets (with only 1000 causal SNPs) are the most challenging ones and the only ones for which increasing heritability helps, and even in those cases we are able to get correlations above 0.45 and up to 0.65. Prior probabilities are inferred well in the rest of the cases, and the greater correlations for scenario 1 are probably due to the greater number of relevant annotations per component as compared to scenario 2 (all the annotations having approximately the same odds ratios in both scenarios, see Figures 3.1 and 3.2).

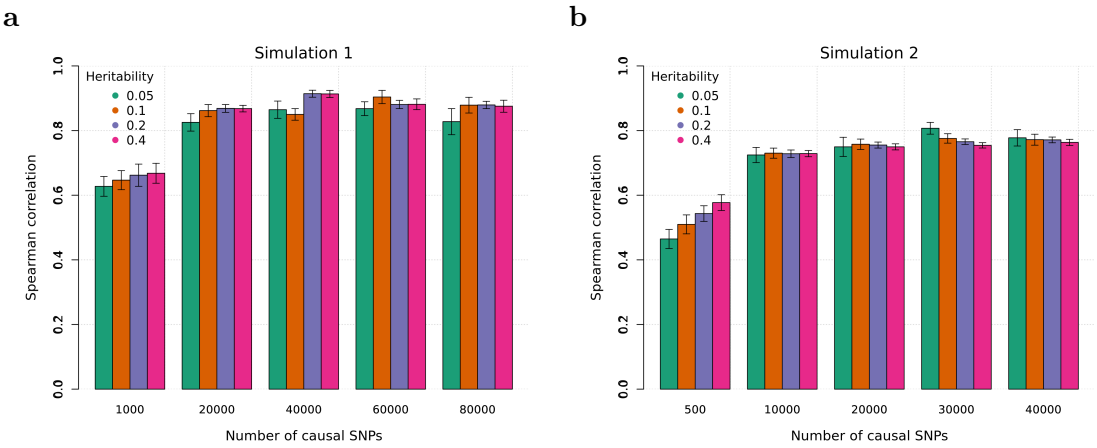


Figure 3.11: Mean correlation of true and inferred prior probabilities. The error bars are standard deviations based on 100 bootstrap samples.

In the next section we will show specific examples of inferred genetic correlations and normalised heritabilities, as well as of component-specific selections of annotations. we will therefore focus now on providing specific examples of the inferred prior probabilities and of the inferred functions f_k, specifically for the two selected datasets from scenarios 1 and 2.

Figures 3.12 and 3.13 show that, as expected, the inferred functions f_k prevent negative values from happening but otherwise leave the linear predictions (LV_k) largely unmodified.

Figures 3.14 and 3.15 compare the true unnormalised priors (LV_k for the k^{th} component) with the inferred unnormalised priors ($f_k(LV_k)$) of non-null components. Every row corresponds to a different inferred component, and every column to a different true component. We matched the components so that the diagonals compare the unnormalised priors of true components with those of matched inferred components. As previously captured by the Spearman correlation, in both cases we are able to learn the true prior probabilities, with little distortions introduced by the inferred f_k.

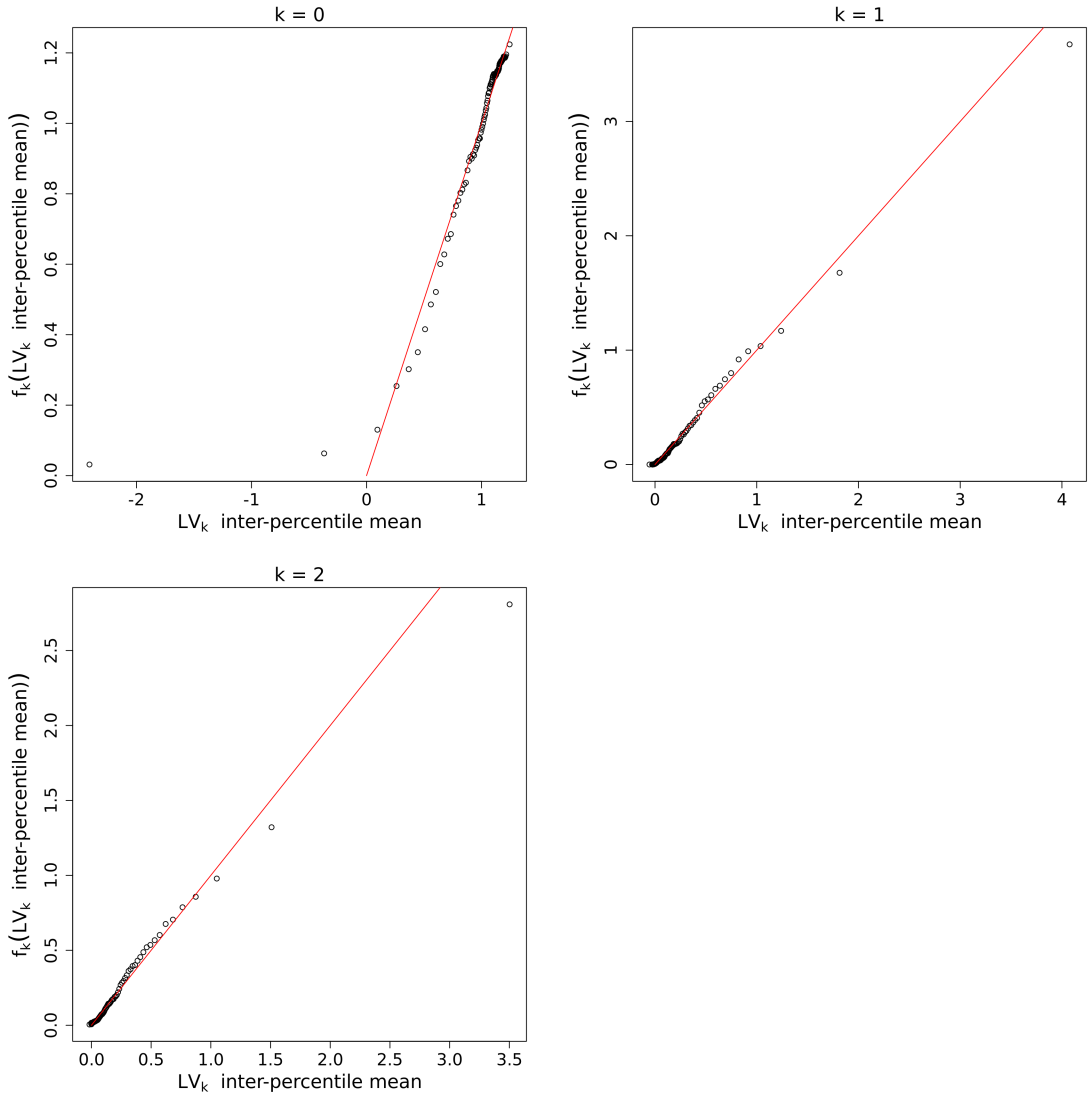


Figure 3.12: Inferred f_k for the selected dataset from simulated scenario 1. The x-axis is the mean linear prediction in each bin, and the y-axis is the mean scaled posterior probability in the same bins.

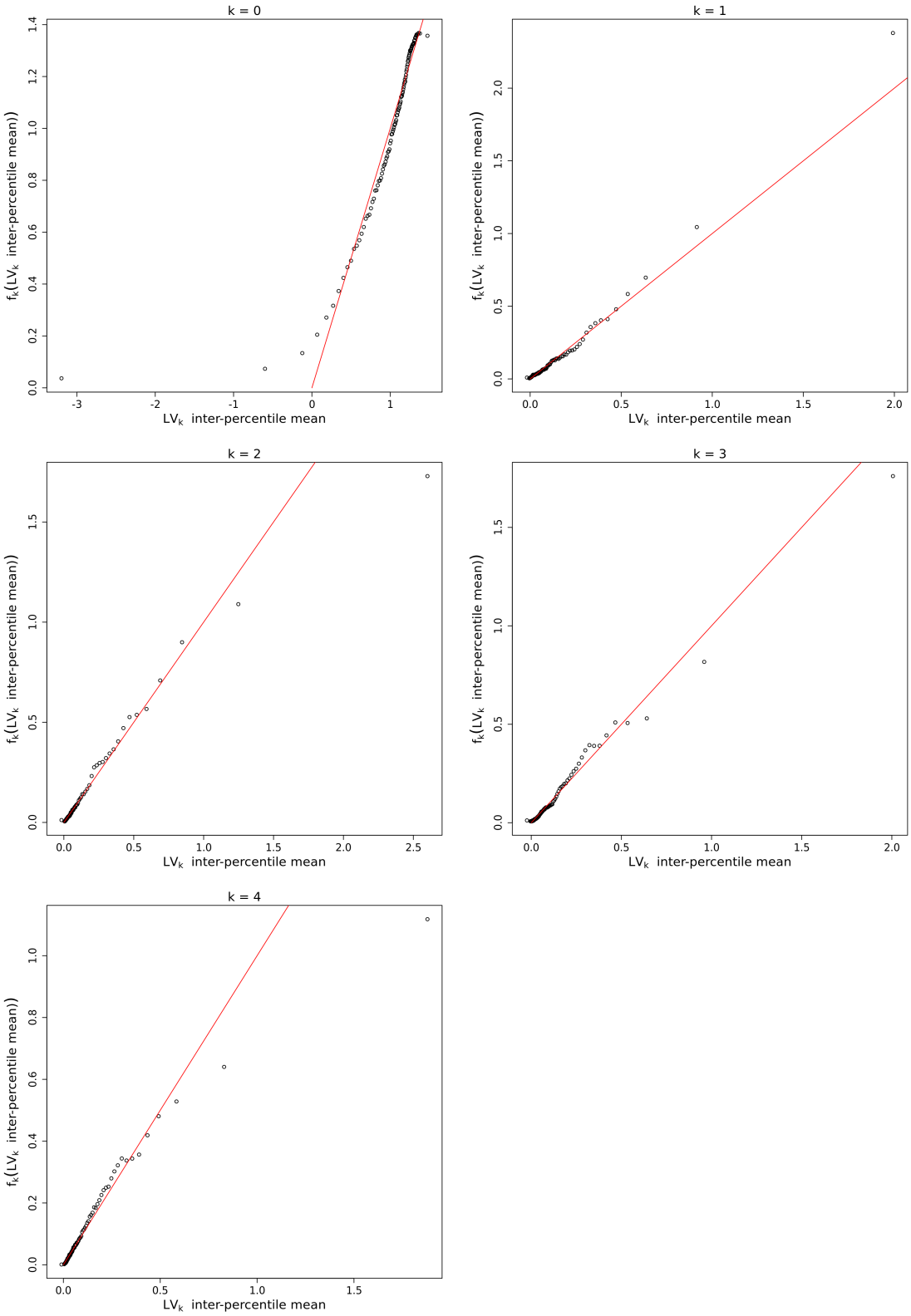


Figure 3.13: Inferred f_k for the selected dataset from simulated scenario 2. The x axis is the mean linear prediction in each bin, and the y axis is the mean scaled posterior probability in the same bins.

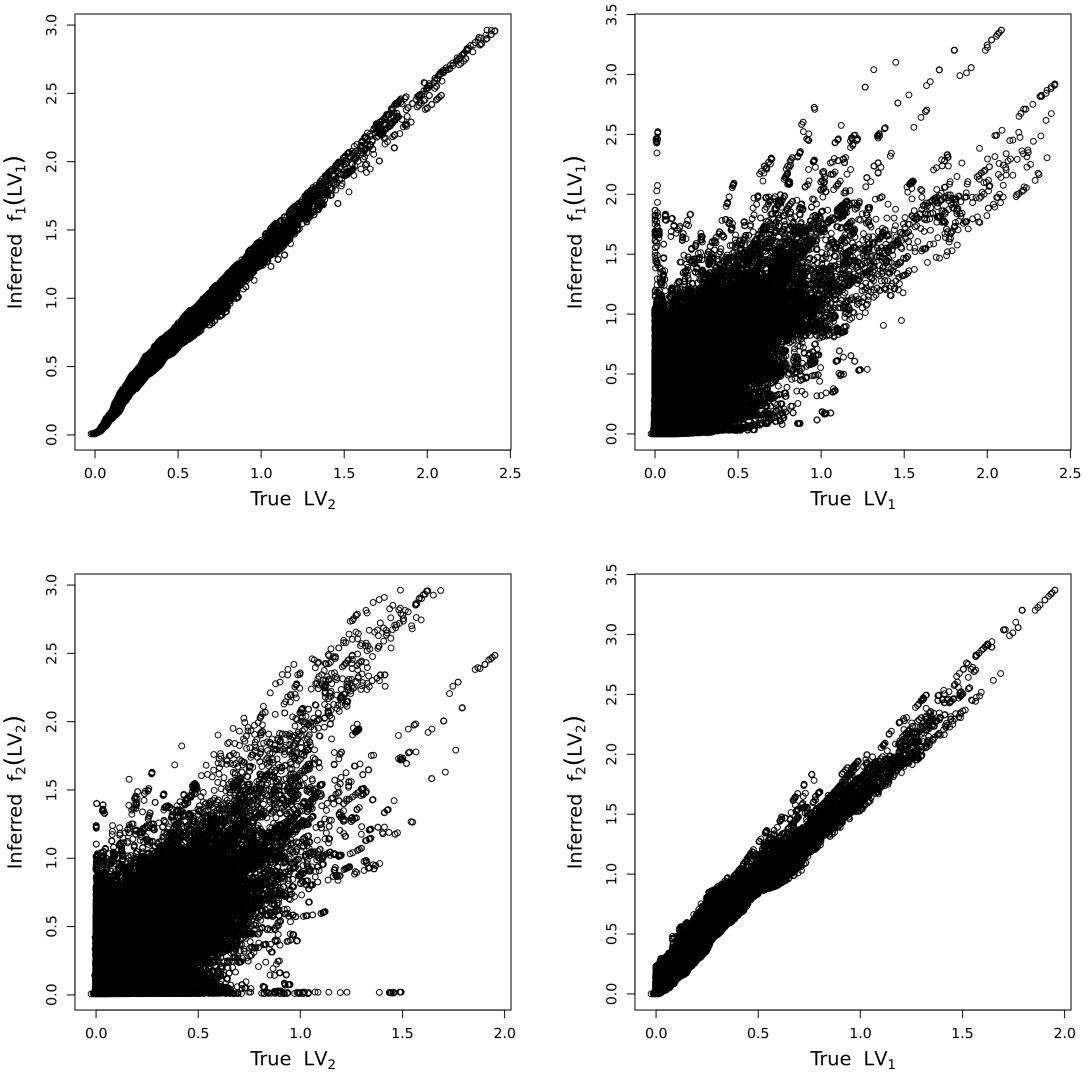


Figure 3.14: True and inferred unnormalised priors for selected datasets from simulated scenario 1. Rows are inferred components and columns true components. The x-axis of each plot is the true unnormalised prior of the corresponding true component (the linear predictions, LV_k), and the y-axis is the unnormalised prior of the corresponding inferred component ($f_k(LV_k)$).

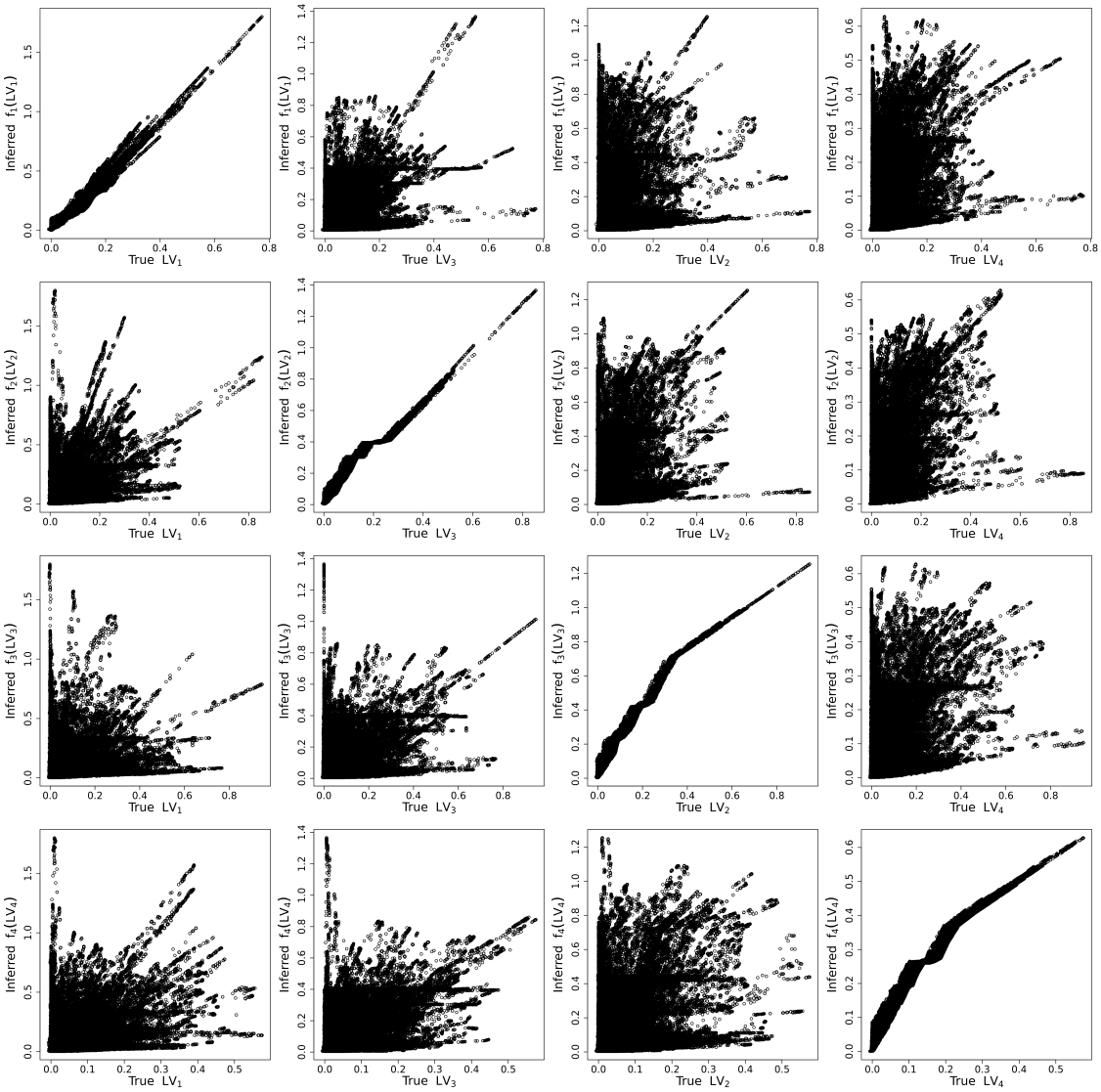


Figure 3.15: True and inferred unnormalised priors for selected datasets from simulated scenario 2. Rows are inferred components and columns true components. The x-axis of each plot is the true unnormalised prior of the corresponding true component (the linear predictions, LV_k), and the y-axis is the unnormalised prior of the corresponding inferred component ($f_k(LV_k)$).

3.3 Selection of the number of components

As explained in the methods chapter (section 2.3.4), we chose the number of components based on 5-fold cross validation (CV) and a measure of component redundancy. Here we show the results for the two selected datasets from simulated scenarios 1 and 2.

In both scenarios, cross-validation favours a redundant, extra component with a large variance for the z-scores of every trait (Figure 3.16). A greater CV likelihood means a better modelling of the distribution of z-scores, but redundant components harden the interpretability of the functional modules (by SNPs moving from the simulated components to the mixed components). we therefore use the number of components favoured by CV for the Bayesian GWAS and one fewer component for the functional interpretation of the inferred components.

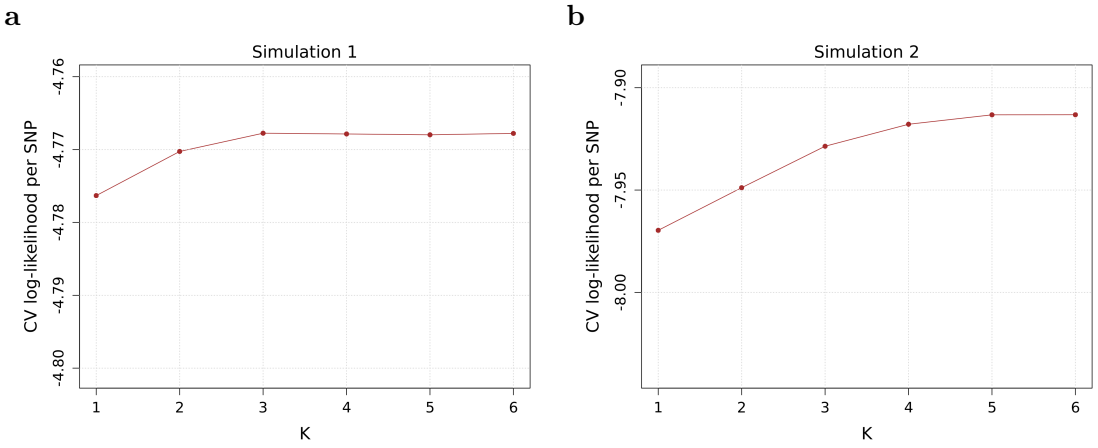


Figure 3.16: 5-fold cross-validation (CV) per-SNP likelihoods for the two selected datasets from simulated scenarios 1 and 2.

3.4 Functional interpretation of inferred components

Here we show how we can interpret the inferred components functionally, continuing with the selected datasets from the two simulated scenarios as examples.

For the selected dataset from simulated scenario 1 (as a reminder: $h_g^2 = 0.2$ and 20,000 causal SNPs), we recovered two equally sized non-null components, as shown

in Figure 3.17 a. The heritability of the third trait is practically evenly distributed across the two components, whereas the other two traits have their heritabilities concentrated mostly in one of the two components. Traits 2 and 3 share an important fraction of their genetic factors, and the same happens with traits 1 and 3. Because the shared genetic factors also account for most of the heritability of traits 1 and 2, we could interpret traits 1 and 2 as the results of two different mechanisms that account for equal fractions of the heritability of trait 3. Recall that traits 1, 2 and 3 were simulated as if they were simplified LDL, blood pressure and CAD, so the two identified mechanisms regulate LDL and blood pressure, respectively.

Looking at the annotation enrichments of each component gives us further information about the two driving mechanisms (Figure 3.17 b): the first component is enriched in SNPs that are near 'specifically expressed genes' (see methods) in tissues that regulate blood pressure, whereas SNPs from the second component implicate tissue related to the lipids metabolism, immune system and arteries. It could be that the implicated tissues are not the truly causal ones but are instead simply the best approximation given our dataset. In this case they are relatively interpretable, so we would conclude that the simulated CAD seems to be driven by two equally important mechanisms: one that regulates blood pressure and one that regulates LDL, each of which is mediated by the corresponding implicated tissues. This is exactly what we simulated (Figure 3.1).

Note that LD explains that the expected numbers of SNPs in the components are greater than the simulated numbers of causal SNPs, and also that the variances of the z-scores are smaller than those of the causal SNPs in each component. Additionally, base LD is implicated for the first component (which was expected due to the small fraction of randomly allocated causal SNPs) but not for the second one, probably because the randomly allocated causal SNPs for the second component tended to overlap redundant annotations that were removed during the forward regression steps.

Similarly, we were able to recover the four driving mechanisms of trait 5 from simulated scenario 2 (Figure 3.18). Note that the annotation that marks 500kbp

windows around transcribed regions is very broad and probably captures the randomly allocated causal SNPs correctly (as base LD does too).

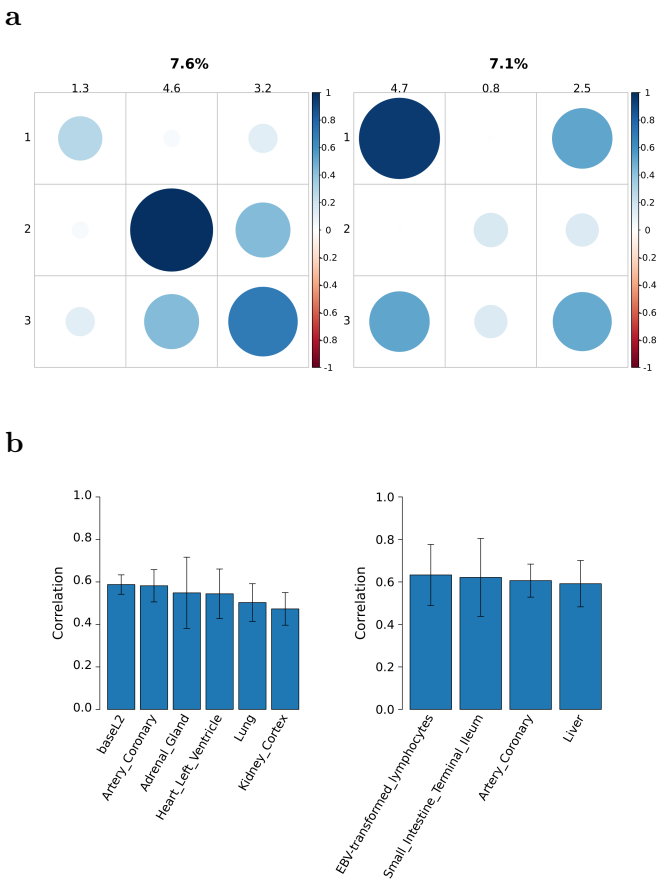


Figure 3.17: Functional interpretation of the inferred components for the selected dataset from simulated scenario 1. **a.** The percentages above the column names are the estimated percentage of SNPs from the training dataset in each component (i.e. the mean posterior probability of belonging to each component). Column names are the variances of the z-scores for each trait, and row names trait indexes. The diagonals are per-SNP heritabilities relative to their maximal value across traits and components. Off-diagonals are genetic correlations. **b.** The y-axis is the Spearman correlation of the LD annotations with the prior probabilities for belonging to the non-null component, and the error bars are standard deviations based on 100 bootstrap samples.

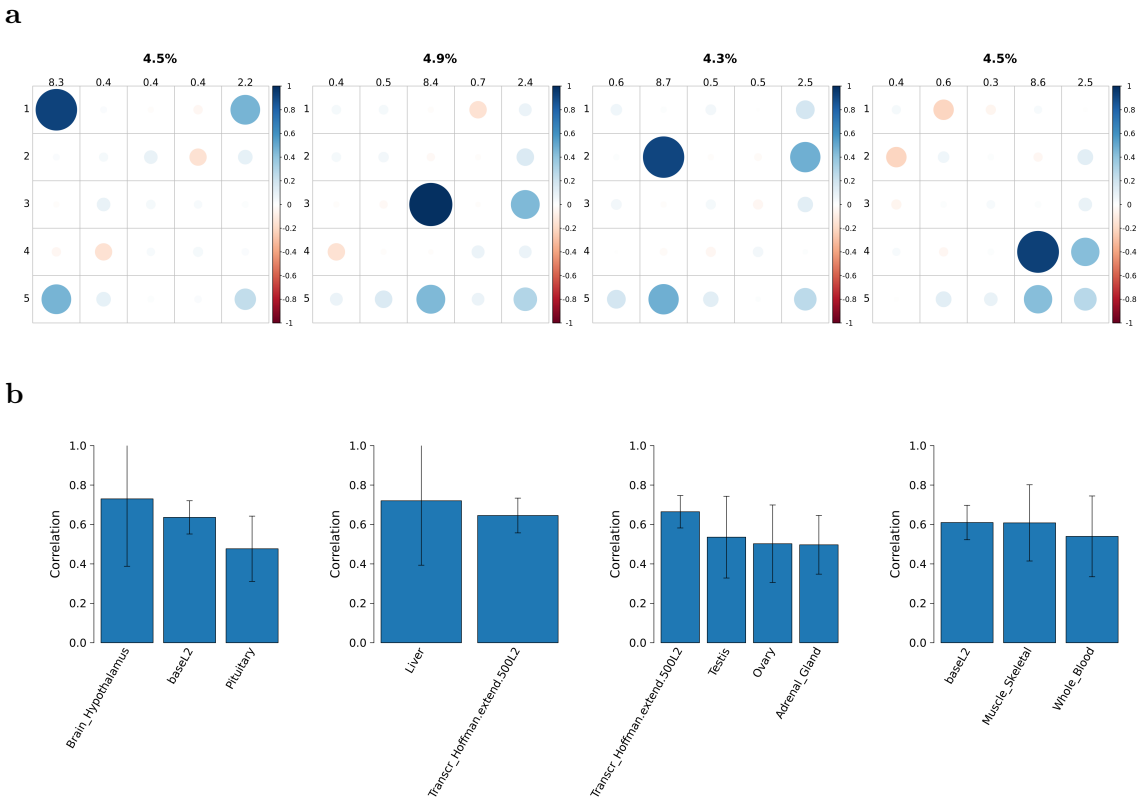


Figure 3.18: Functional interpretation of the inferred components for the selected dataset from simulated scenario 2. **a.** The percentages above the column names are the expected percentage of SNPs from the training dataset in each component (i.e. the mean posterior probability of belonging to each component). Column names are the variances of the z-scores for each trait, and row names trait indexes. The diagonals are per-SNP heritabilities relative to their maximal value across traits and components. Off-diagonals are genetic correlations. **b.** The y-axis is the Spearman correlation of the LD annotations with the prior probabilities for belonging to the non-null component, and the error bars are standard deviations based on 100 bootstrap samples.

3.5 Bayesian GWAS

Here we assess our ability to correctly classify SNPs as either null or causal using our posterior probabilities of association. we continue with the two selected datasets as case studies.

Starting with trait 3 from the selected dataset from scenario 1, Figure 3.19 a compares the ROC curves (see section 2.6.3) of linear regression (implemented by Plink) [69], a multi-trait GWAS meta-analysis (implemented by MTAG [55]), and three versions of our mixture model. MTAG uses GWAS summary statistics to

first estimate the overall genetic covariance of pairs of traits with bivariate LD score regression [52] (i.e. not allowing for different groups of SNPs to have different genetic covariances), and then uses the estimates to boost GWAS power while controlling for any overlap of the GWAS cohorts [55]. For the same number of false positives, our mixture model finds more truly causal SNPs than Plink or MTAG. This was expected because our model uses both SNP annotations and multi-trait z-scores to assign SNPs to non-null components. In fact, when we 'turn off' the use of these features before calculating the probabilities of association (see section 2.6.4), the classification performance of our model drops to the level of Plink, which naturally uses only single-trait z-scores. When we turn off the use of only one of the features (either the annotations or the z-scores for non-focal traits), classification performance drops but remains better than that of Plink.

The mixture model also performs better than MTAG, which uses multiple traits but averages genetic correlations across components. This is the case even when we turn off the use of SNP annotations or multiple traits, suggesting that using more than one component (when appropriate) is beneficial as long as we use either SNP annotations and/or multi-trait z-scores to allocate SNPs to the right component.

As expected, the performance of all the compared methods increases and converges as the true SNP effects for standardised genotypes increase, and the relative performance of our method with respect to the others increases as the SNP effects get smaller (due to the gained power; Figure 3.19 b). The strongest hits are found both by the mixture model and by Plink (Figure 3.19 c), whereas weaker hits are found exclusively by the mixture model (shown in green). Among the hits found exclusively by the mixture model, 3,705 are independent (meaning that there is no other clumped hit within 500kbp).

Results for trait 1 (from the same dataset, scenario 1) are similar (Figure 3.21). For the second simulated scenario, results for trait 5 are also similar except that not using multiple traits has a greater impact in classification performance than before, which was expected due to the stronger non-focal traits and slightly weaker priors (Figure 3.20). However, in this case all the heritability of trait 1 concentrates

in only 5,000 SNPs, which leads to large z-scores and to a similar performance by all methods, as expected (Figure 3.22).

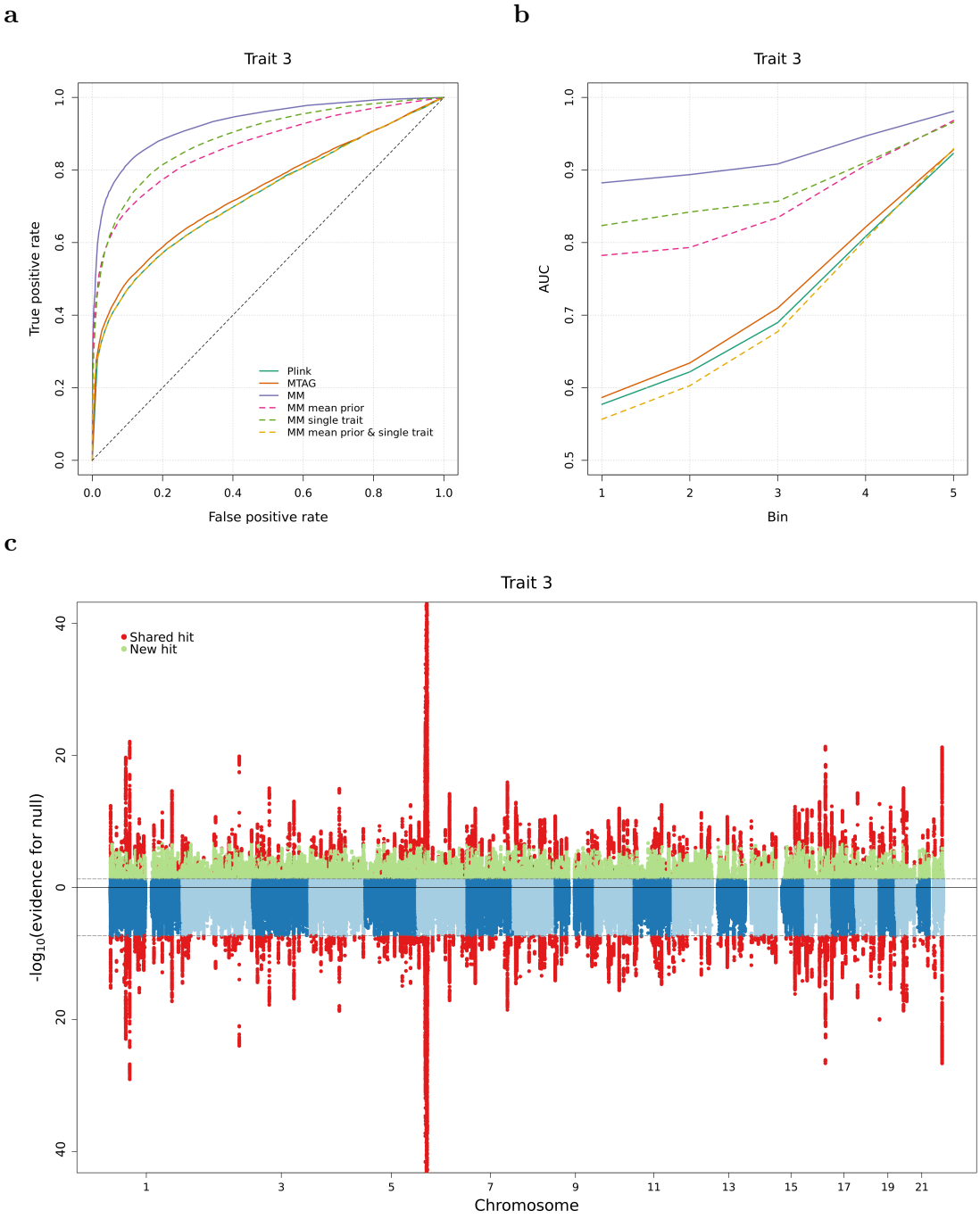


Figure 3.19: Bayesian GWAS for trait 3 from the selected dataset from scenario 1. **a** and **b** share the same legend. In **b**, AUC is the area under the ROC curve and bins are of true SNP effects (see section 2.6.3). In **c**, the upper part is based on the mixture model, and the lower part based on Plink. Evidence for null is a p-value for Plink, and a posterior probability of non-association for the mixture model. Shared hits are SNPs that are significant with Plink and our mixture model, whereas new hits are SNPs that are only significant with the mixture model. The horizontal dashed lines are the thresholds for significance, which are $-\log_{10}(5 \times 10^{-8})$ for Plink and $-\log_{10}(0.05)$ for the mixture model).

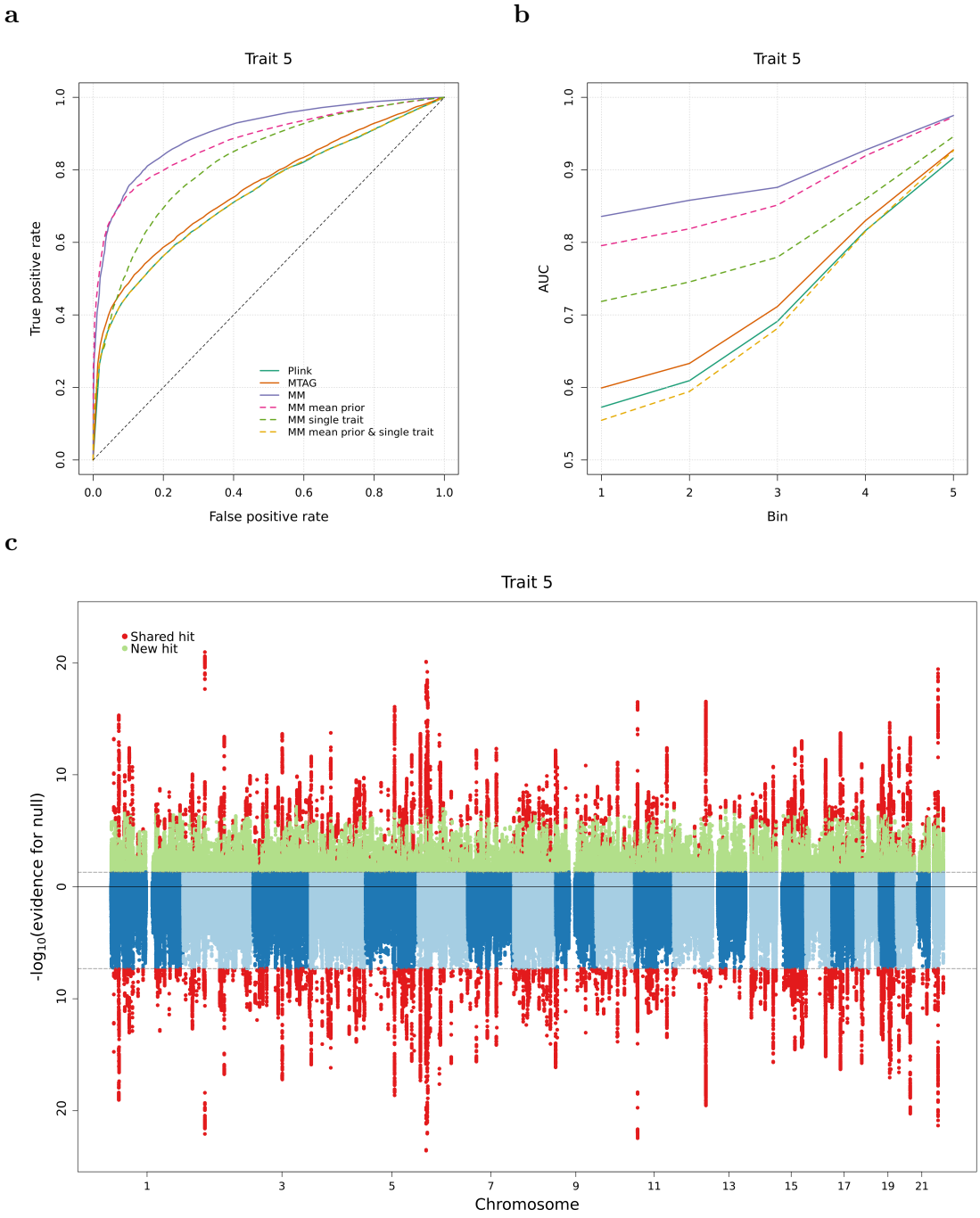


Figure 3.20: Bayesian GWAS for trait 5 from the selected dataset from scenario 2. **a** and **b** share the same legend. AUC is the area under the ROC curve. In **c**, the upper part is based on the mixture model, and the lower part based on Plink. Evidence for null is a p-value for Plink, and a posterior probability of non-association for the mixture model. Shared hits are SNPs that are significant with Plink and our mixture model, whereas new hits are SNPs that are only significant with the mixture model. The horizontal dashed lines are the thresholds for significance, which are $-\log_{10}(5 \times 10^{-8})$ for Plink and $-\log_{10}(0.05)$ for the mixture model).

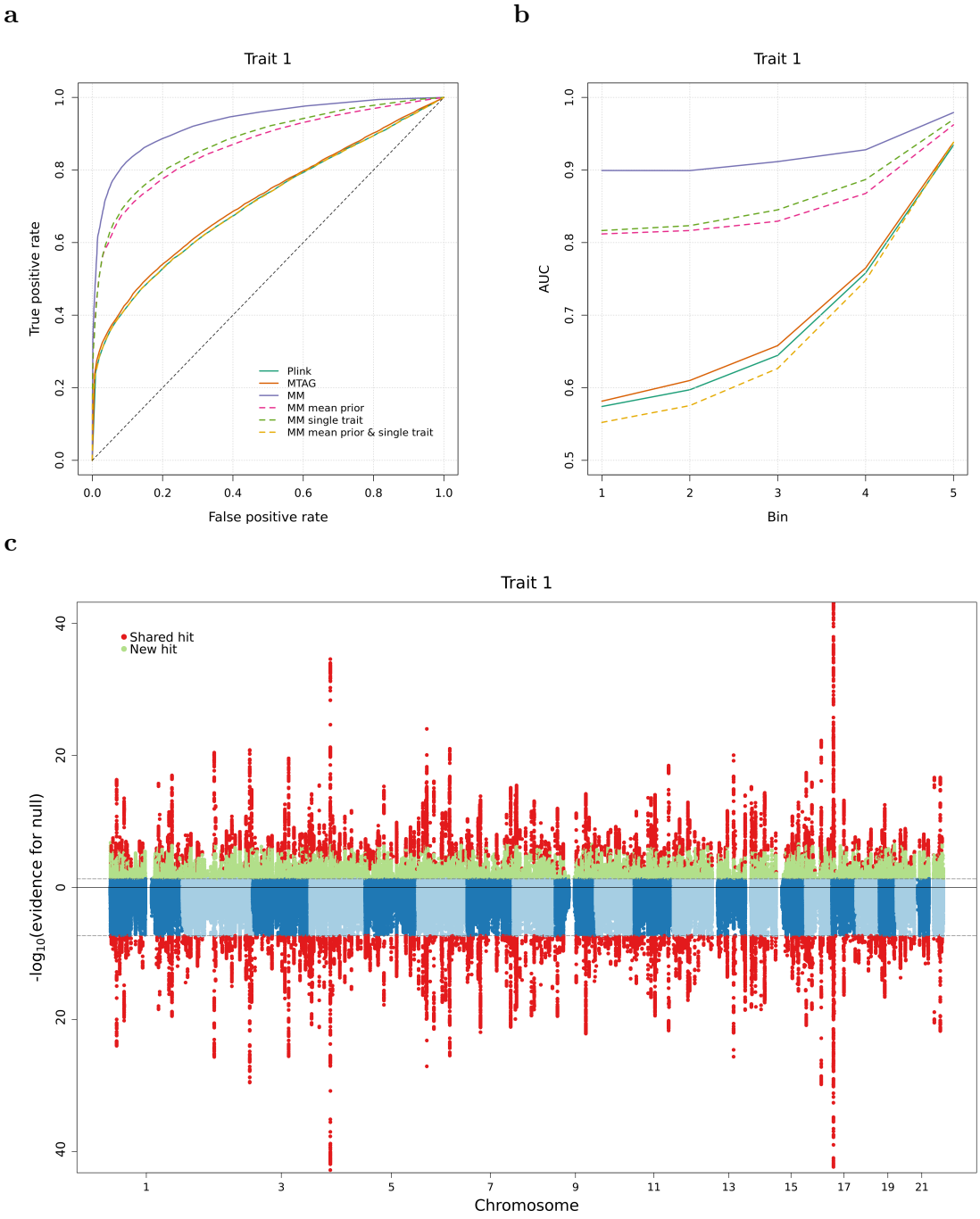


Figure 3.21: Bayesian GWAS for trait 1 from the selected dataset from scenario 1. **a** and **b** share the same legend. AUC is the area under the ROC curve. In **c**, the upper part is based on the mixture model, and the lower part based on Plink. Evidence for null is a p-value for Plink, and a posterior probability of non-association for the mixture model. Shared hits are SNPs that are significant with Plink and our mixture model, whereas new hits are SNPs that are only significant with the mixture model. The horizontal dashed lines are the thresholds for significance, which are $-\log_{10}(5 \times 10^{-8})$ for Plink and $-\log_{10}(0.05)$ for the mixture model).

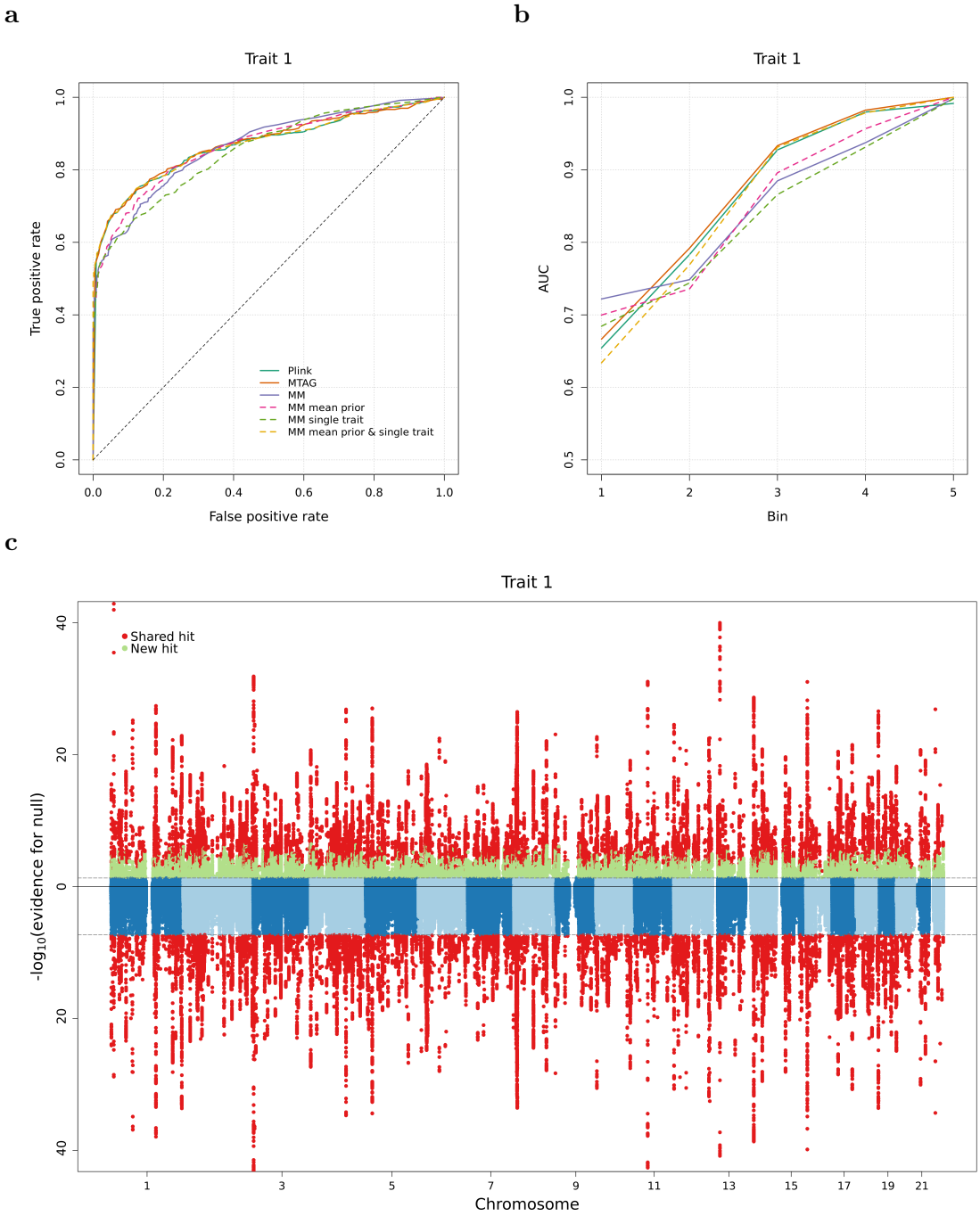


Figure 3.22: Bayesian GWAS for trait 1 from the selected dataset from scenario 2. **a** and **b** share the same legend. AUC is the area under the ROC curve. In **c**, the upper part is based on the mixture model, and the lower part based on Plink. Evidence for null is a p-value for Plink, and a posterior probability of non-association for the mixture model. Shared hits are SNPs that are significant with Plink and our mixture model, whereas new hits are SNPs that are only significant with the mixture model. The horizontal dashed lines are the thresholds for significance, which are $-\log_{10}(5 \times 10^{-8})$ for Plink and $-\log_{10}(0.05)$ for the mixture model).

Table 3.1 shows the true positive rate (TPR) and false discovery rate (FDR)[1] for a significance threshold of 0.05 for the Bayesian GWAS and of 5×10^{-8} for Plink and MTAG. For the MM, the false discovery rate is always below 0.05, as it should be, and the true positive rate is greater than for Plink and MTAG for comparable FDRs (with the exception of the not very polygenic trait 1 from scenario 2).

Table 3.1: TPR and FDR for the MM, Plink and MTAG, based on fixed significance thresholds (0.05 for the MM and 5×10^{-8} for Plink and MTAG.)

		MM		Plink		MTAG	
		TPR	FDR	TPR	FDR	TPR	FDR
Scenario 1	Trait 1	0.19	0.002	0.04	0.008	0.03	0.009
	Trait 3	0.24	0.003	0.03	0.005	0.03	0.002
Scenario 2	Trait 1	0.39	0.004	0.19	0	0.19	0
	Trait 5	0.28	0.009	0.003	0.01	0.03	0.003

To further illustrate the source of the increased power of our Bayesian GWAS, we investigate the hits that are no longer hits after turning off either the use of annotations or the use of non-focal traits (i.e. hits that depend on, or are driven by, either annotations or non-focal traits). Hits that are mainly driven by the z-scores of non-focal traits have on average smaller absolute z-scores for the focal trait and larger absolute z-scores for the non-focal traits than the rest of the hits (Figures 3.23 and 3.24, a-b and d-e). The reason is that when the z-score for the focal trait and the prior probabilities together are not enough to make a SNP pass the significance threshold, large z-scores for other traits can make the difference.

Similarly, hits that depend on the annotations have on average larger prior probabilities for non-null components than the rest of the hits (Figures 3.23 and 3.24, c and f). This is because prior probabilities can help to allocate SNPs to non-null components when absolute z-scores are small, helping SNPs to pass the significance threshold.

[1]Recall that the true positive rate is the proportion of true positives correctly classified as positive, whereas the false discovery rate is the proportion of positives that are false positives (i.e. type-I error).

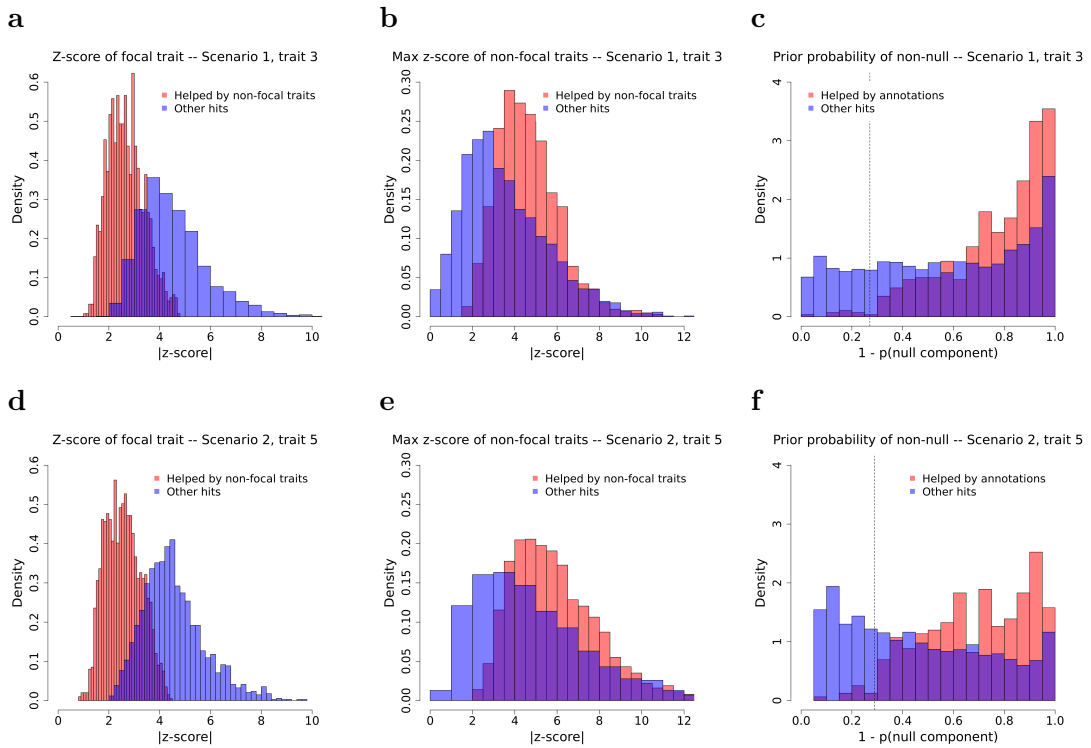


Figure 3.23: Properties of hits driven by either non-focal traits or annotations, for traits 3 and 5 from scenarios 1 and 2, respectively. The y-axes show probability densities, and the x-axes show either the absolute values of z-scores or the prior probabilities for belonging to any of the non-null components. The vertical dashed lines in **c,f** are the mean prior probabilities for belonging to the null component.

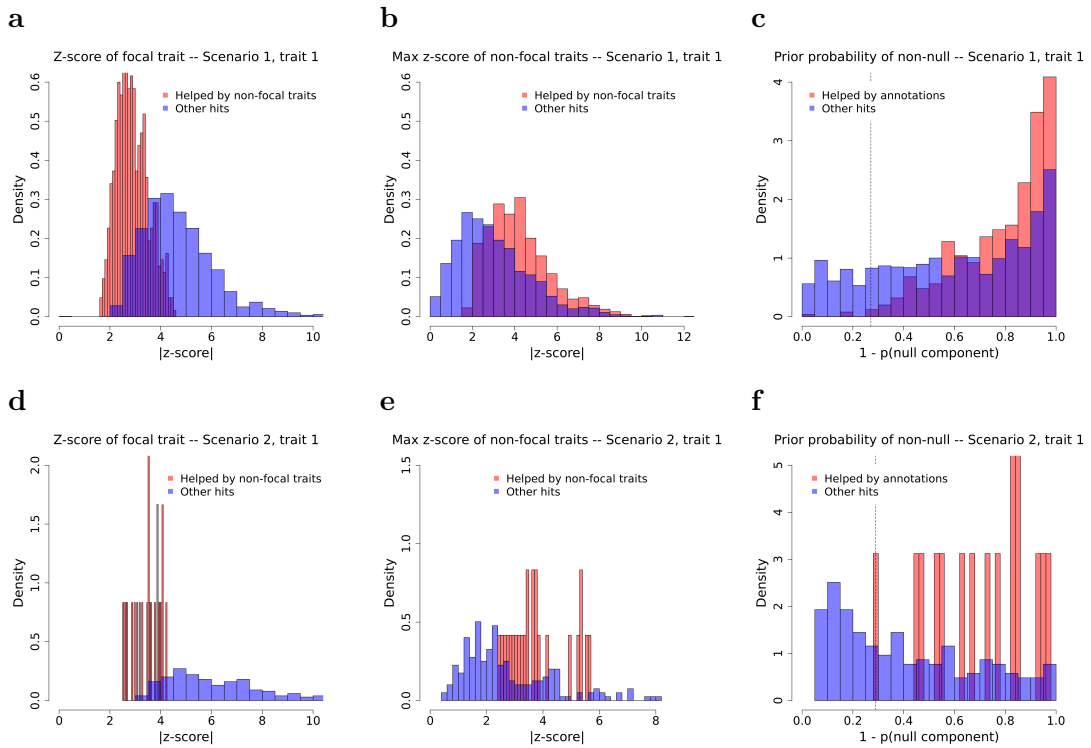


Figure 3.24: Properties of hits driven by either non-focal traits or annotations, for traits 1 and 1 from scenarios 1 and 2, respectively. The y-axes show probability densities, and the x-axes show either the absolute values of z-scores or the prior probabilities for belonging to any of the non-null components. The vertical dashed lines in **c,f** are the mean prior probabilities for belonging to the null component.

3.6 Assignment of hits to functional modules

After classifying SNPs as either null or non-null for all the studied traits with the Bayesian GWAS, we can allocate hits to the different components using their posterior probabilities over component memberships.

To assess our ability to correctly allocate hits to components, we used the intersection of hits with truly causal SNPs, so that we have a ground truth that is approximately well-defined (see below). Most hits are assigned to the right component in both scenarios (Figure 3.25) and, as expected, none of the hits is allocated to the null component (numbers are rounded for clarity, but rows do add up to one exactly). Note that ground truth here is given by the simulated component memberships and not by the observed z-scores and LD annotations, so the small fractions of misclassified hits are, at least in part, likely artefacts of this fact. That is to say, the model is probably correctly allocating those SNPs based purely on the observed data, with the true labels being impossible to recover.

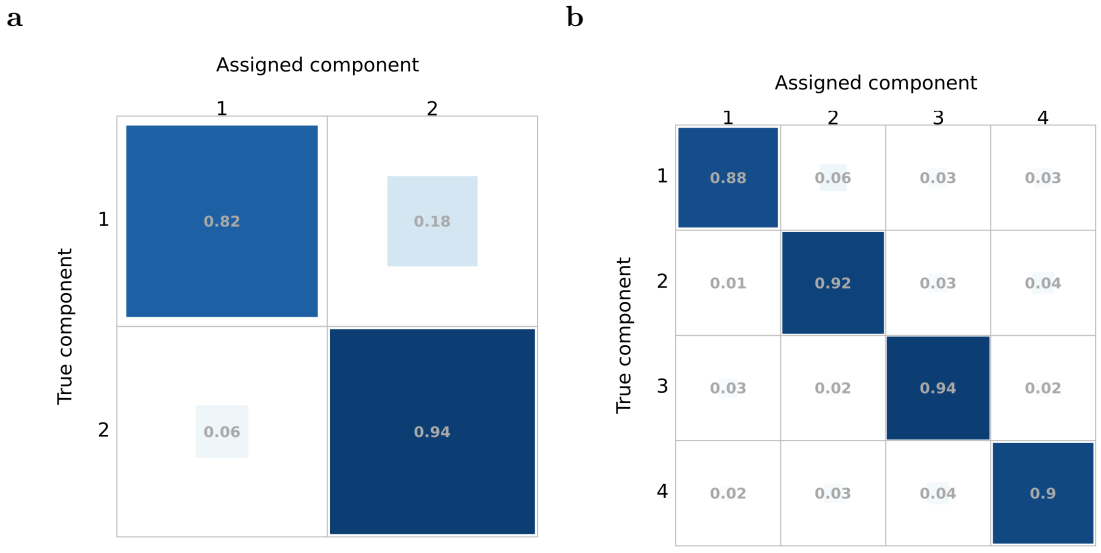


Figure 3.25: Confusion matrices for the allocation of hits to components. **a** Is for the selected dataset from scenario 1, and **b** is for the selected dataset from scenario 2. Numbers have been rounded (rows add up exactly to one) and represent proportions of hits. For example, in **a**, 82% of the hits that truly belong to component 1 are assigned to the inferred component 1, and the rest to the inferred component 2. The inferred components are matched to the true ones so that the numbers in the diagonals correspond to the proportions of correctly assigned hits.

4

Application to CAD and ASD

Contents

4.1 Overview

In this chapter, we use the inference and analysis pipeline validated in Chapter 3 to investigate the biological mechanisms driving Coronary Artery Disease (CAD) and Autism Spectrum Disorder (ASD). we will introduce these two traits and provide some background below. However, in brief, we chose CAD because it has several relatively well known risk-conferring mechanisms that we can seek to replicate, and we chose ASD as a contrasting example of a much more challenging and poorly understood trait.

4.2 Analysis of CAD

4.2.1 Background

Coronary Artery Disease (CAD) is the insufficient supply of blood to the heart due to the narrowing of the coronary arteries, typically due to the build-up of atherosclerotic plaque in their walls [21]. When the coronary arteries become completely blocked, CAD leads to a heart attack, which is globally the leading cause of mortality [72]

The heritability of CAD based on a linear model is estimated to be around 0.48 [73], and about 200 independent loci have been associated to CAD to date [12] as a result of increasingly large GWAS[1].

GWAS hits seem to be converging to pathways related to lipids metabolism (LDL and triglycerides), blood pressure, inflammation, and the development and proper functioning of blood vessels. Whether additional pathways are implicated in modulating CAD risk remains to be seen [21]. Current strategies to reduce CAD risk target the LDL and blood pressure pathways, in addition to lifestyle changes [75].

Here we sought to validate our method by recovering two well-known mechanisms driving CAD: increased serum levels of LDL and past smoking (coded as SMK, a binary trait that encodes whether individuals have ever smoked regularly or not). Both LDL and SMK have an estimated additive heritability of about 0.08[2] [13]. A comprehensive study of CAD will require a richer dataset and further work to be able to confidently detect subtler components, as discussed in chapter 5.

4.2.2 Description of the dataset

Table 4.1 lists the size and source of the GWAS from which we downloaded the z-scores for CAD, LDL and SMK. For inference, we kept the 1,085,237 SNPs with

[1] 60,801 cases and 130,681 controls in 2015 by the CARDIoGRAMplusC4D consortium [74], to which new 10,898 cases and 76,535 controls from the UK Biobank were added in 2017 by others [75], and making it to the 181,522 cases and 984,168 controls in the latest GWAS in 2022 by adding nine additional studies, all with European ancestry [12] (further reaching 210,842 cases and 1,167,328 controls in a cross-ancestry meta-analysis with the Japan Biobank).

[2] `https://nealelab.github.io/UKBB_ldsc/h2_browser.html`.

available z-scores for the three traits and aligned their effect alleles across traits[3]. The annotations are the same 515 introduced in Section 2.5 (396 Roadmap, 53 SEG based on GTEx, and 66 baseline).

Table 4.1: Summary of the GWAS summary statistics used for the CAD analysis. All the GWAS are based on individuals with European ancestry. LDL is a continuous trait and therefore cases are not defined (N/A: not applicable).

Trait	**N**	$\mathbf{N}_{cases}$	**Source**
CAD	184,305	60,801	Nat Genet. 2015 Oct; 47(10):1121-1130 [74]
LDL	440,546	N/A	PLoS Med. 2020 Mar 23; 17(3):e1003062 [77]
SMK	607,291	311,629	Nat Genet. 2019 Feb; 51(2):237-244 [13]

Figure 4.1 shows the z-scores used for inference for pairs of traits (around 1m HapMap3 SNPs), and Figure 4.2 the GWAS QQ-plots showing the expected enrichments in small p-values characteristic of well-powered GWAS.

[3]All the effect alleles had been previously aligned to the same strand of the hg19 reference genome by `https://gwas.mrcieu.ac.uk/` [76], from which we downloaded the summary statistics originally produced by the studies in Table 4.1.

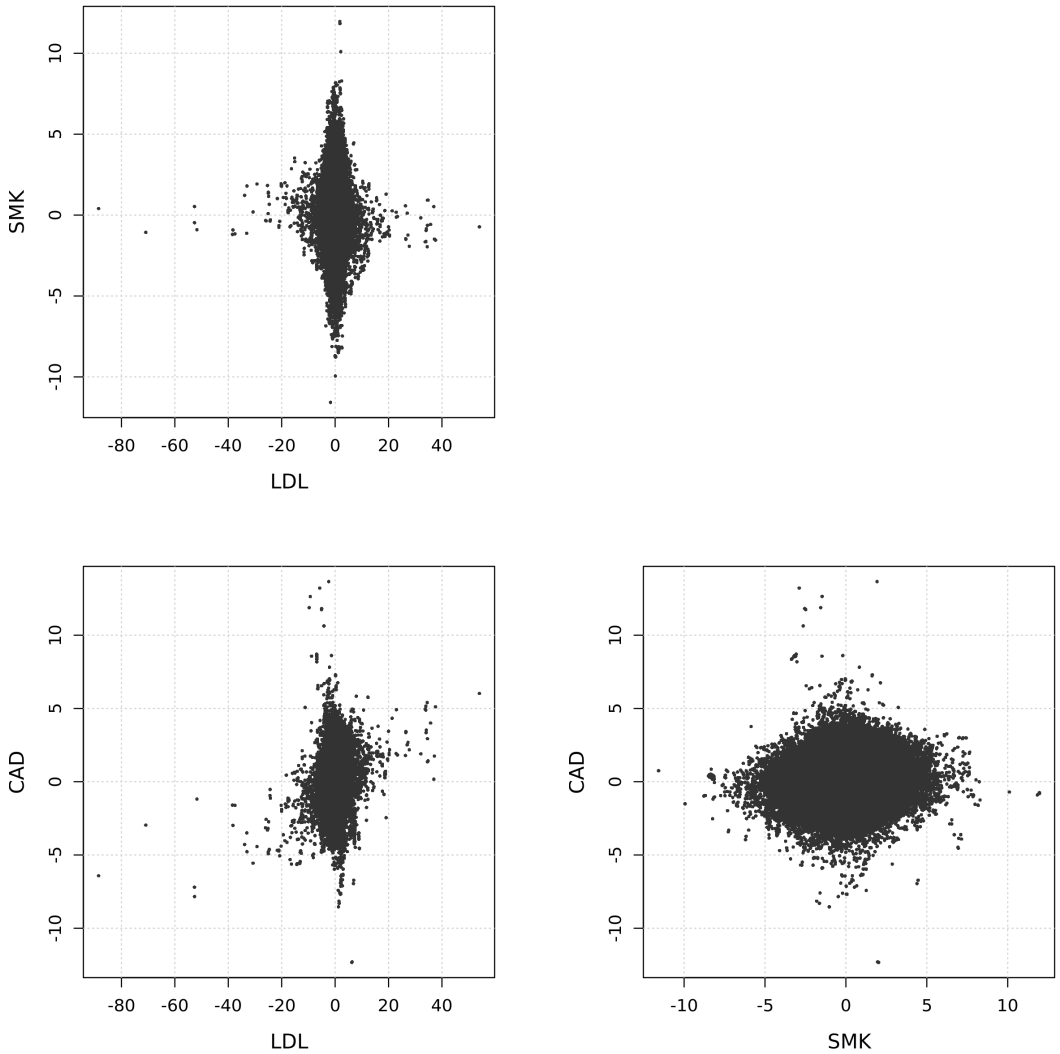


Figure 4.1: Z-scores used for inference from the CAD dataset, for pairs of traits.

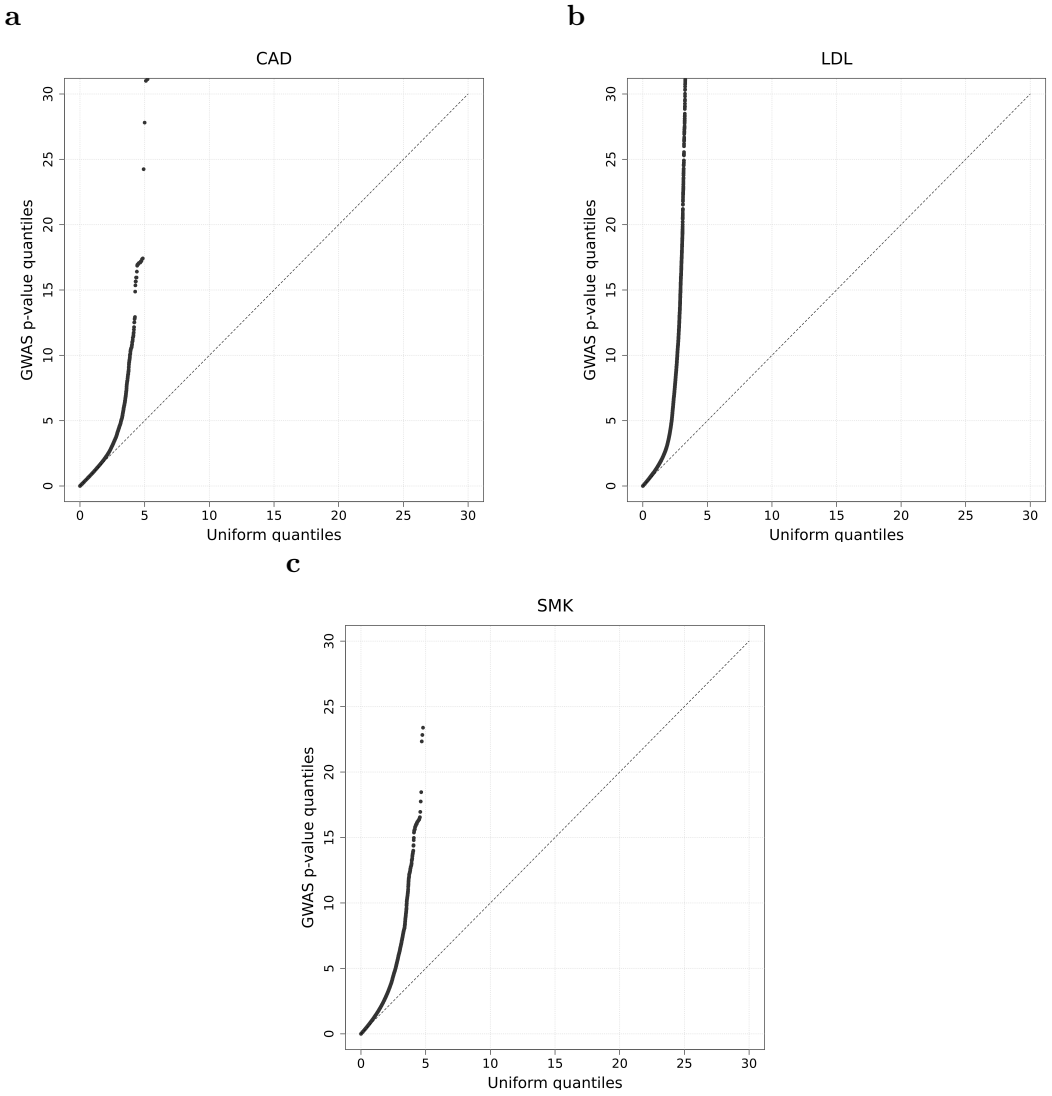


Figure 4.2: QQ-plots for the $-\log_{10}$ p-values of the CAD dataset. The x-axes are the quantiles of a uniform distribution ranging from 0 to 1 in $-\log_{10}$ space. The y-axes are the quantiles of the empirical distribution of p-values, also in $-\log_{10}$ space.

4.2.3 Results

Choice of the number of components

Figure 4.3 shows the 5-fold cross-validation likelihood for the CAD dataset. We chose three non-null components for the interpretation of functional modules (since the fourth one was a combination of the other three) and four for the Bayesian GWAS (to benefit from a slightly better modelling of the distribution of the z-scores).

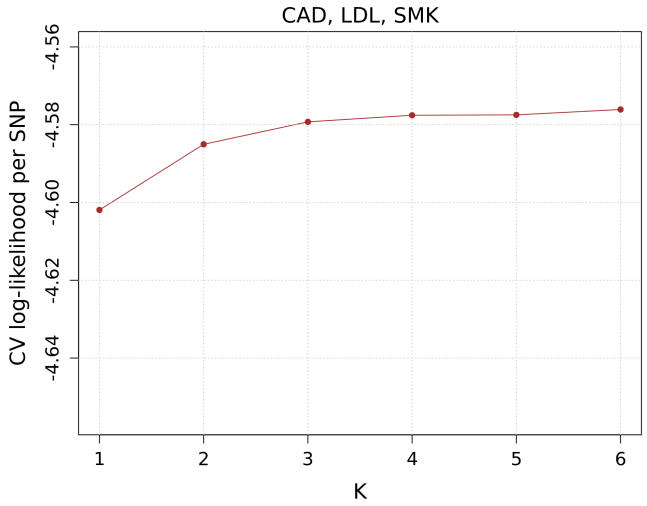


Figure 4.3: 5-fold cross-validation (CV) likelihood per SNP as a function of the number of non-null components for the CAD dataset.

Functional interpretation of the inferred components

Figure 4.4 shows the functional signatures of the three inferred non-null components. The first component captures the expected LDL pathway that confers risk for CAD. The relatively few SNPs in it have large z-scores for LDL and significantly correlated effects for LDL and CAD (consistent with the known causal link between LDL and CAD). This contrasts with the third component, for which weaker SNP effects on LDL are independent of SNP effects on CAD (due to additional factors controlling CAD risk; Figure 4.4 a). The top five annotation enrichments are related to the LDL metabolism, implicating the liver, the adrenal gland and coding regions (Figure 4.4 b). The rest of the enrichments are less interpretable (hippocampus, fibroblasts or spleen), and they may be explained by correlations of the LD annotations with other truly relevant tissues or cell types not included in our dataset. Finally, this

component does not account for almost any heritability of SMK, consistent with our expectation about the approximate independence of the LDL and SMK pathways.

The second component consists of the greatest expected number of SNPs, and accounts for most of the heritability of past smoking (Figure 4.4 a). SNP effects on SMK and CAD are significantly correlated, consistent with the known causal link between smoking and increased risk for CAD. The annotation enrichments point to the fetal brain and to coding and conserved regions (Figure 4.4 c). The enrichment in base LD scores may be due to the existence of relevant annotations that are not well approximated by a combination of the annotations in our dataset.

The third component accounts for the remaining heritability of the three traits. Notably, most of the heritability of CAD comes from this component, suggesting that the LDL and SMK pathways account for a relatively small fraction of CAD risk (Figure 4.4 a). This component probably captures the more peripheral parts of the LDL and SMK pathways (supported by SNPs with weaker effects), as well as one or several remaining mechanisms driving CAD. Mixing SNPs from multiple mechanisms attenuates the annotation enrichments, but we still found several enrichments: adipose tissue is probably due to LDL, whereas lung and placenta may be pointing to a mechanism that regulates blood pressure. Indeed, including blood pressure in the analysis (Figure B.1) showed that a similar third component explains most of the heritability of blood pressure, has very correlated SNP effects for blood pressure and CAD, and is also enriched in adipose and lung annotations (plus fetal muscle, also likely related to blood pressure). Overall, the third component from the main analysis is a representative example of how limited coverage of relevant traits and annotations can limit our ability to pull apart and interpret functional modules.

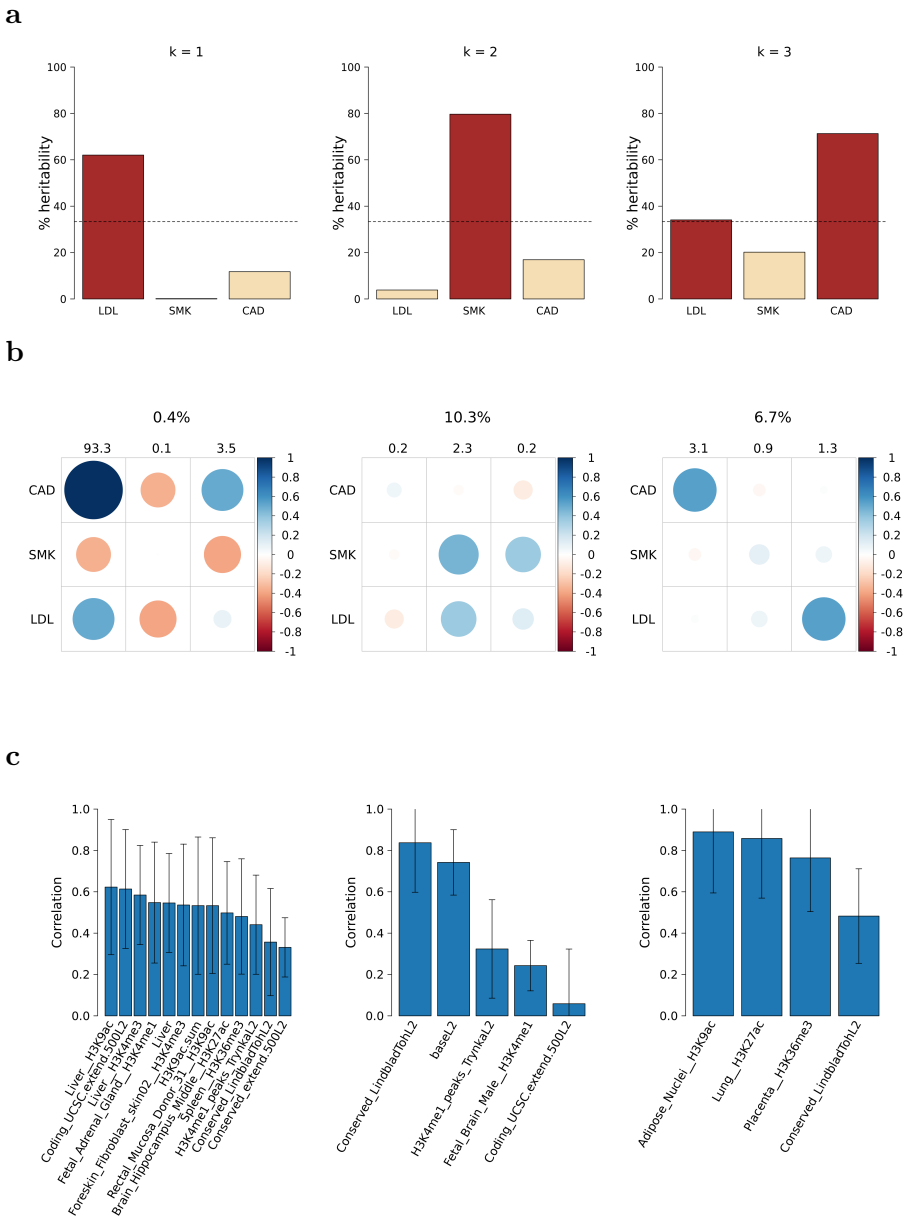


Figure 4.4: Interpretation of the three inferred non-null components for the CAD dataset. **a** Shows the percentage of heritability explained by each component, for each trait (so the bars add up to 100% for each trait). **b** Shows, for each component, genetic correlations in the off-diagonals, per-SNP heritabilities relative to the greatest one across components and traits in the diagonals, and z-score variances as column names (note that these depend on the GWAS sample sizes, whereas the per-SNP heritabilities do not). **c** Shows the Spearman correlations of the LD annotations that survived the last, component-specific filter with forward regression (see section 2.3.7), with the prior probabilities of belonging to the component. The error bars are standard deviations based on 100 bootstrap samples. See section 2.6.1 for further details.

Bayesian GWAS

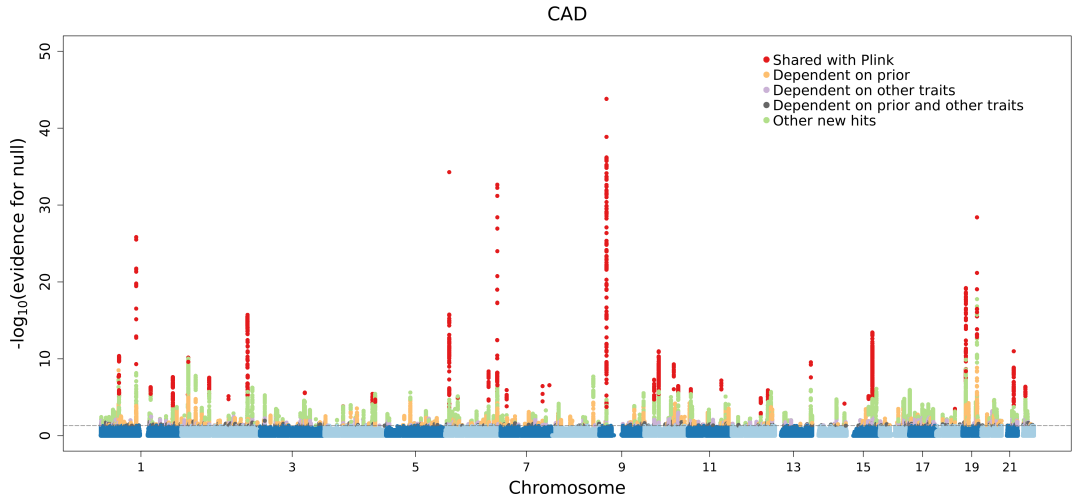
Our Bayesian GWAS replicated all the Plink hits and found 792 newly associated independent loci (as compared to Plink). Part a of Figure 4.5 shows the Manhattan plot for our Bayesian GWAS, colouring hits differently as a function of the source of evidence that drives them (see section 2.6.4). In this case, the strongest hits are also found by Plink (in red), implying that they have large absolute z-scores for CAD. Many of the newly associated SNPs are supported by both the annotations and the z-scores of other traits (hits in green). This is because they remain hits after turning off either the use of the priors or the use of other traits when calculating the probabilities of association (i.e. both sources suffice in isolation). Other hits are driven by strong priors (in orange) or by other traits (in purple). The weakest hits tend to need the simultaneous weak support from both sources of evidence (in grey), so that even if none of the sources suffices in isolation, their combined effect pushes the SNP above the significance threshold.

An in-depth investigation of the new hits constitutes an important area of future work. However, we will highlight that some of our new hits have also been recently associated to CAD for the first time in the newest GWAS for CAD [12] (Dec 2022). An example is rs6883598 (p-value = 0.0029 in 2015 [74] and 9.7×10^{-10} in 2022 [12], and PPA[4] = 0.96 in our GWAS), in chromosome 5, also associated to blood pressure, which we are able to assign with high confidence to our third inferred component. The hit has also been indirectly linked to connective tissue and vascular-related issues via a nearby gene, FBN2 [12]. In our GWAS, rs6883598 was driven by a combination of weak multi-trait z-scores and weak prior probabilities (the latter mainly due to high placental and conserved LD annotations). Another example is rs2207132 (p-value = 0.0001 in 2015 [74] and 6.7×10^{-10} in 2022 [12], and PPA = 0.99997 in our GWAS), in chromosome 20, which was also newly reported in the same paper and discovered by our Bayesian GWAS thanks to very large effects on LDL.

Figure 4.6 shows similar Manhattan plots for the non-focal traits, LDL and SMK, also with many new hits that could be investigated in the future.

[4]PPA: posterior probability of association.

a



b

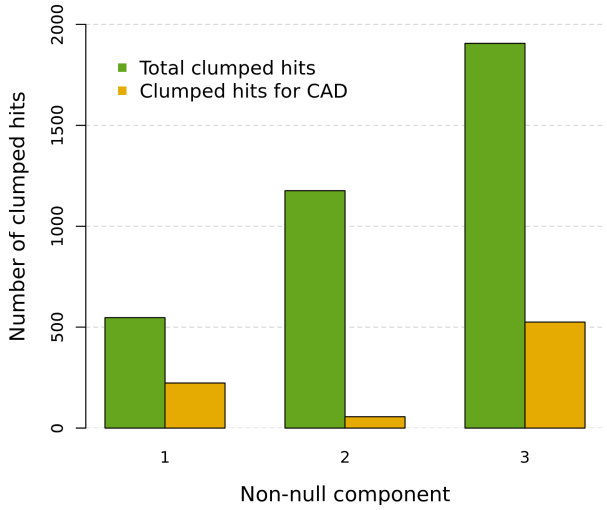


Figure 4.5: a. Manhattan plot for CAD based on four non-null components. **b.** The green bars show the distribution of confidently assigned independent hits for either CAD, LDL or SMK across non-null components. The orange bars show the same but focusing on independent hits for CAD.

Assignment of independent hits to components

After confidently associating SNPs to CAD, we can assign them to the different inferred components as a way to know the mechanisms through which they confer risk for CAD. Figure 4.5 b shows the number of independent loci associated to CAD that can be confidently assigned to each component (orange bars). It also shows how many of the independent loci associated to at least one trait (CAD, LDL or

SMK) can be confidently assigned to each component. The majority of independent hits for CAD belong to the third component, with the second component having the smallest number of independent hits for CAD. Without restricting the analysis to CAD hits, we can confidently populate the three inferred components with hundreds of hits for at least one of the studied traits (green bars).

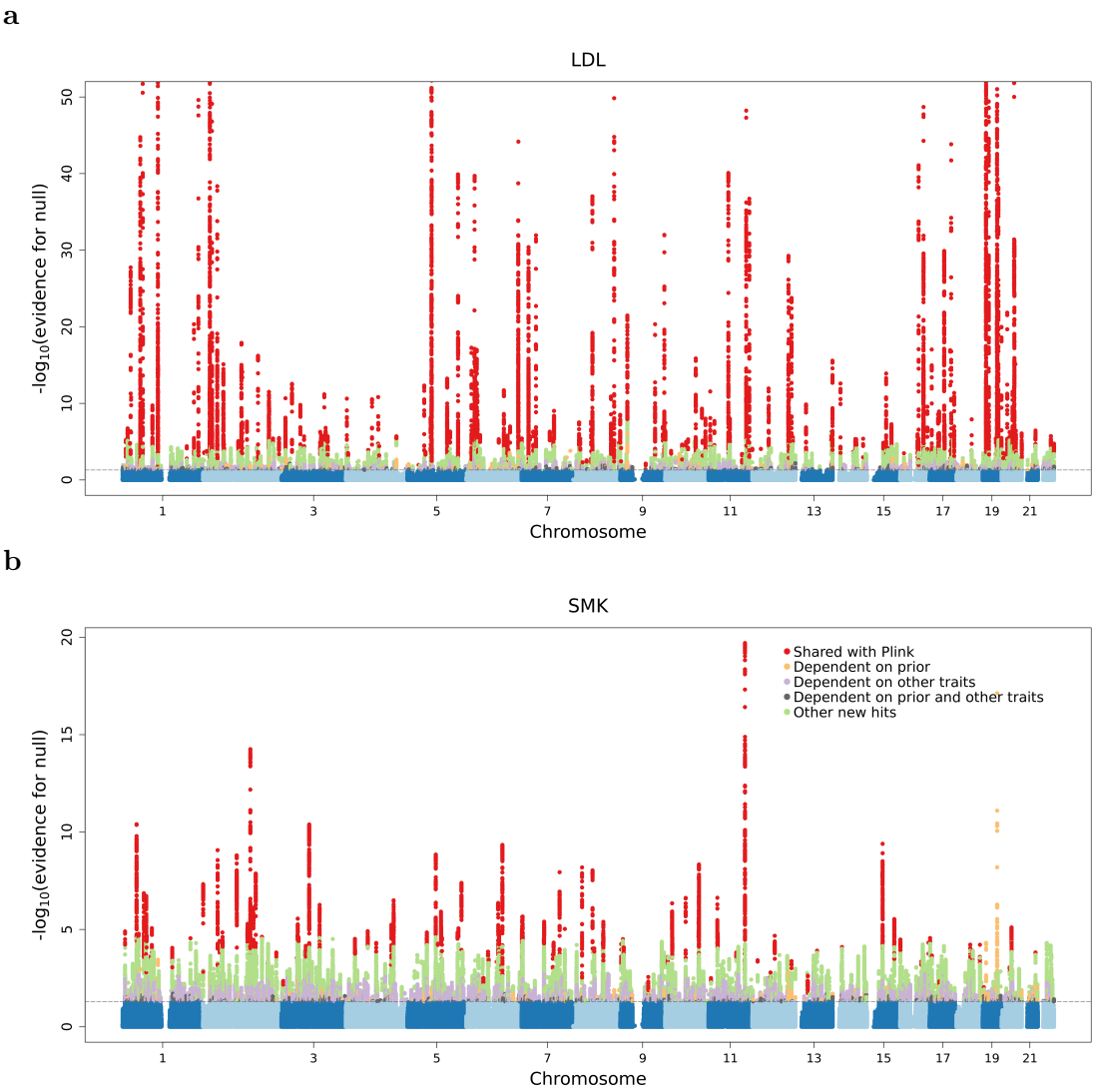


Figure 4.6: Manhattan plots for LDL and SMK based on four inferred non-null components.

4.3 Analysis of ASD

4.3.1 Background

Autism Spectrum Disorder (ASD) is a spectrum of phenotypes characterised by the simultaneous presence of persistent social communication deficits, restricted interests and repetitive behaviours. Such symptoms must appear at an early stage of development, cannot be better explained by intellectual disability or by global developmental delay, and must also 'cause clinically significant impairment in social, occupational, or other important areas of current functioning' [78]. Despite the significant impact of ASD on its carriers[5], there is currently no pharmacological treatment for its symptoms. This is partially explained by its psychiatric and neurodevelopmental nature, which complicate the mechanistic interpretation of the around 100 genetic associations known to date [25, 80]. Even if only 15 of the associated variants are common [81], most of the heritability of ASD is estimated to result from the additive effects of common variants, which is at least 0.1 [81–84].

ASD is genetically correlated with other psychiatric disorders and cognitive traits [54, 81, 83, 85], including the ones added to our analysis (educational attainment, major depressive disorder, attention deficit hyperactivity disorder and schizophrenia). Recent efforts have focused on explaining those genetic correlations as the result of different driving mechanisms. Some studies look at pairs of traits and attempt to find trait-specific and shared GWAS hits (e.g. ASD and ADHD [86] or ASD with neurodevelopmental disorders in general [80]). One study provided evidence in favour of at least two independent genetic components of ASD that approximate its two core symptoms [87]. Another study used a different collection of relatively low-level traits related to ASD (from the SPARK cohort) to support from two to three independent components using structural equation modelling (SEM) [85]. The authors of [59] also

[5]Autism Speaks, the largest non-profit organisation in the United States devoted to promoting solutions for ASD, has extensively documented the impact of ASD on cases and their families. Major problems are suffering of the ASD cases (as a result of professional limitations, bullying and co-morbid clinical conditions, among others), the significant attention demand from case relatives, and the overall economical cost associated to ASD (which is estimated to be about 60000 per year for children in U.S., triplicating for adults[6]). According to 2016 data from 11 U.S. states, ASD affects 1 in 54 (1.85%) 8-year old children [79], and there is currently no pharmacological treatment for the core symptoms of ASD.

use SEM to explain the genetic correlations of 11 psychiatric traits (including ASD), finding however four correlated components that are consequently hard to interpret.

The key ASD-related traits that have been used to show multiple independent genetic components of ASD do not have publicly available GWAS results (systemising [87] and traits measured in the SPARK cohort [85]). Here we analyse ASD together with some of the main genetically correlated psychiatric traits as a starting point toward more elaborate future designs. Like for CAD, a comprehensive study of ASD, in addition to a richer dataset, will require further work to be able to confidently detect subtler components (see chapter 5 for a discussion).

4.3.2 Description of the dataset

Table 4.2 lists the size and source of the GWAS from which we downloaded the z-scores for Autism Spectrum Disorder (ASD), Educational Attainment (EA), Attention Deficit Hyperactivity Disorder (ADHD), Major Depressive Disorder (MDD) and Schizophrenia (SCZ). we kept the 907,970 SNPs with available z-scores for every trait and without strand-ambiguous alleles (that is, without A/T or C/G reference-alternative allele pairs). We used the same 515 annotations as for the simulated and CAD datasets (396 Roadmap, 53 SEG based on GTEx, and 66 baseline).

Table 4.2: Summary of the GWAS summary statistics used for the ASD analysis. All the GWAS are based on individuals with European ancestry.

Trait	**N**	**N_{cases}**	**Source**
ASD	46,350	18,381	Nat Genet. 2019 Mar; 51(3):431-444 [81]
EA	765,283	X	Nat Genet. 2022 Apr; 54(4):437-449 [88]
ADHD	53,293	19,099	Nat Genet. 2019 Jan; 51(1):63-75 [89]
MDD	173,005	59,851	Nat Genet. 2018 May; 50(5):668-681 [90]
SCZ	127,906	52,017	Nature 2022 Apr; 604(7906):502-508 [91]

Figure 4.7 shows the z-scores used for inference (∼1m HapMap3 SNPs) for pairs of traits. Figure 4.8 shows the GWAS QQ-plots, which reveal that the GWAS for ASD, ADHD and MDD find weaker associations than those for traits in the CAD dataset.

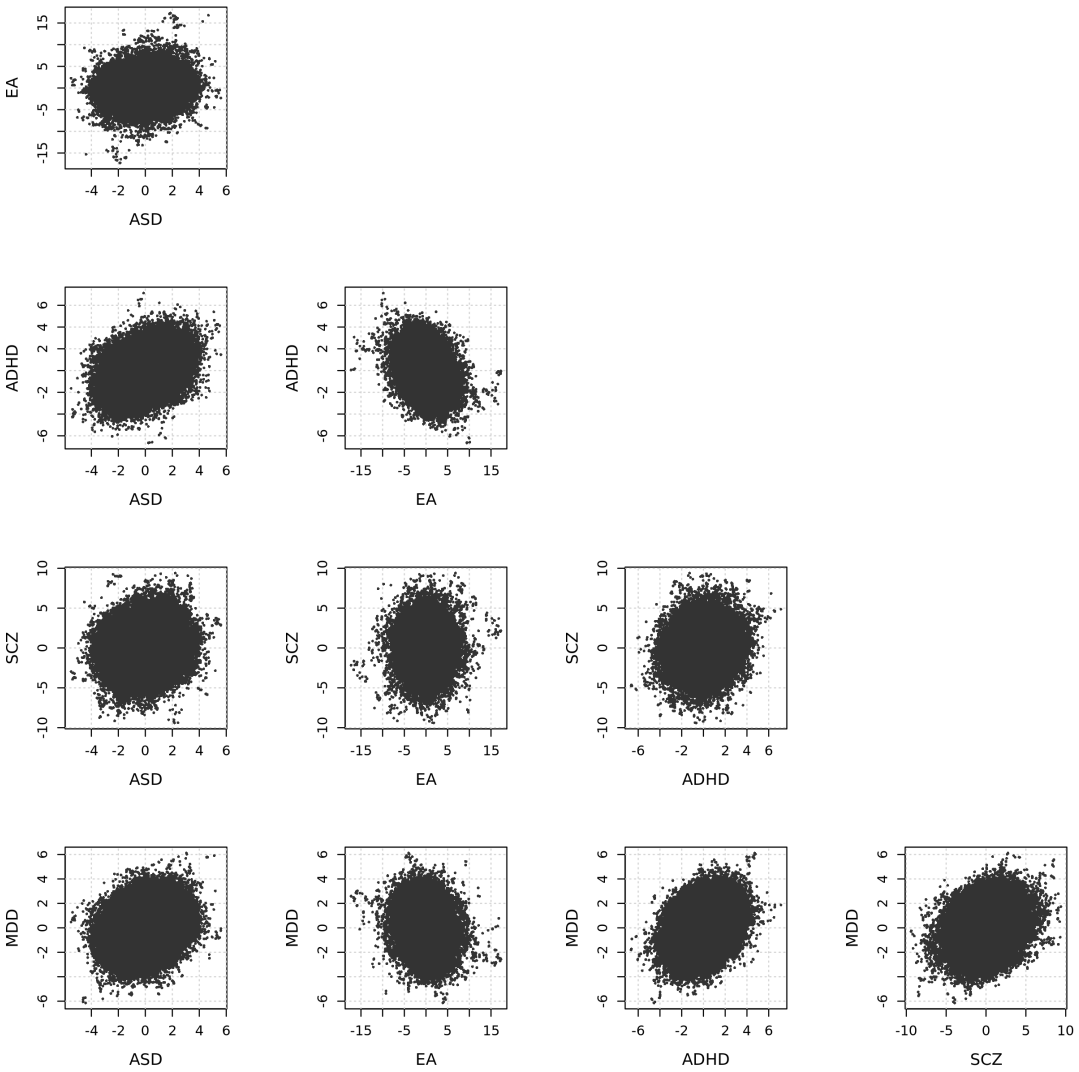


Figure 4.7: Z-scores used for inference from the ASD dataset, for pairs of traits.

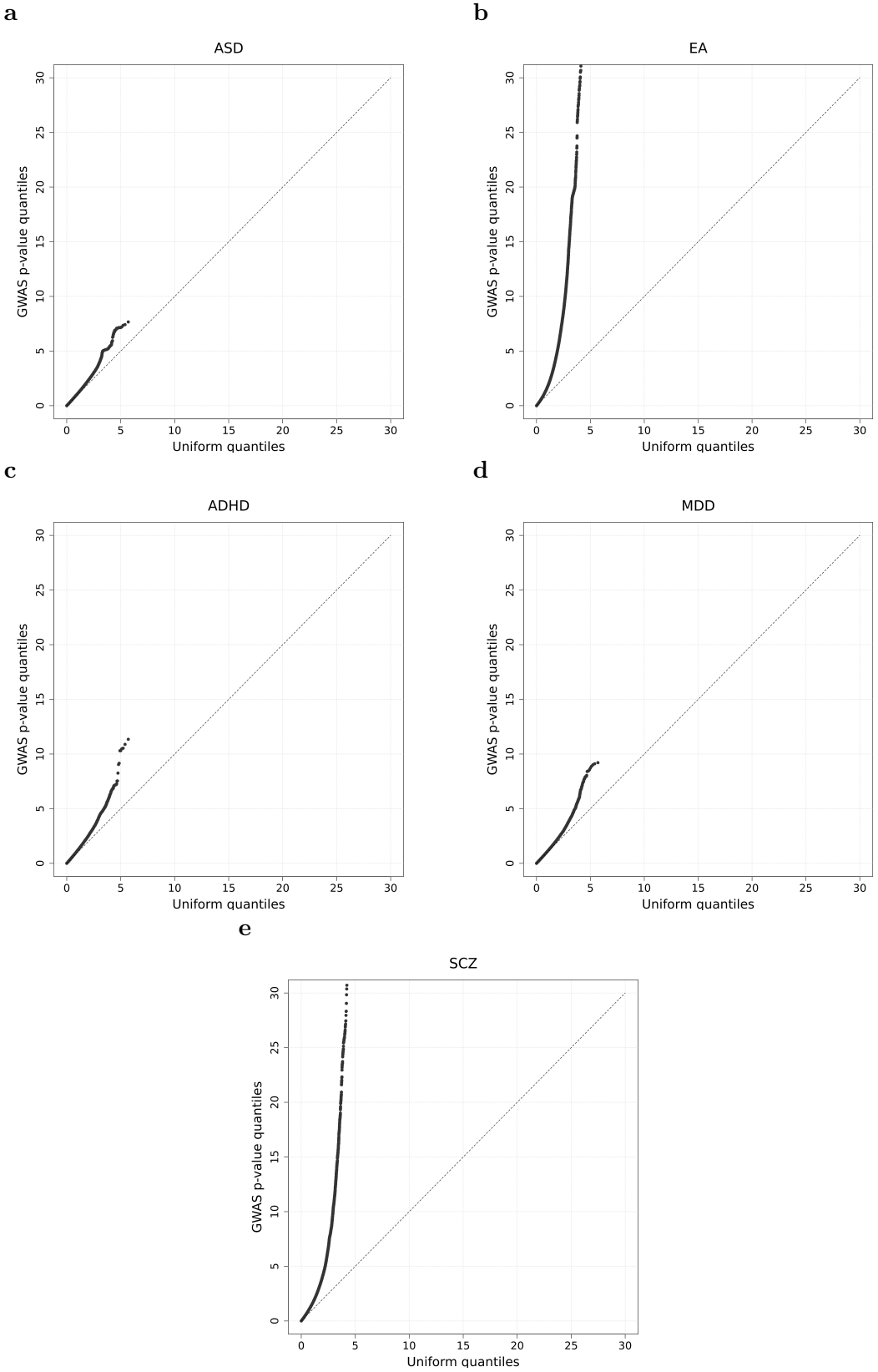


Figure 4.8: QQ-plots for the $-\log_{10}$ p-values of the ASD dataset. The x-axes are the quantiles of a uniform distribution ranging from 0 to 1 in $-\log_{10}$ space. The y-axes are the quantiles of the empirical distribution of p-values, also in $-\log_{10}$ space.

4.3.3 Results

Choice of the number of components

Figure 4.9 shows the 5-fold cross-validation likelihood for the ASD dataset. We chose only one non-null component for its interpretation, and three for the Bayesian GWAS (the additional components are combinations of the rest). We chose three for the GWAS to benefit from a slightly better modelling of the distribution of the z-scores. As we discuss in chapter 5, further work is needed to investigate the biological relevance of components that increase the CV likelihood only marginally, as these are likely a combination of biologically meaningful components and artefacts due to LD.

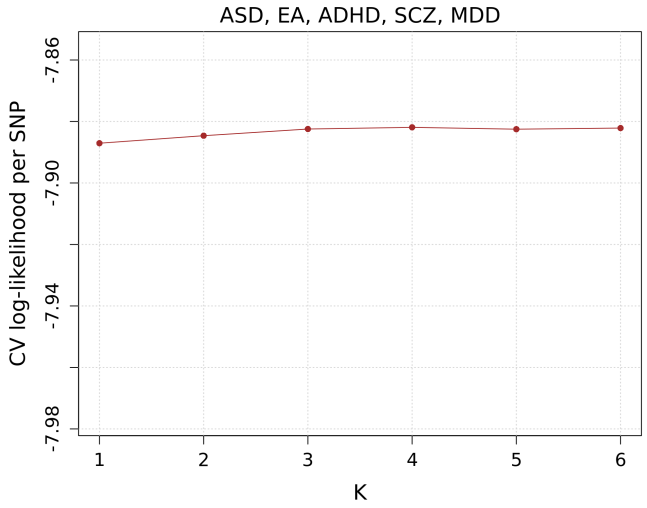


Figure 4.9: 5-fold cross-validation (CV) likelihood per SNP as a function of the number of non-null components for the ASD dataset.

Functional interpretation of the inferred components

Figure 4.10 shows the functional signature of the inferred non-null component. The inferred genetic correlations are consistent with previous estimates [54]: all the traits are genetically correlated with ASD, and educational attainment is negatively correlated with ADHD and MDD, and genetically independent of SCZ.

The top three enrichments are for broad annotations: conserved regions and their surroundings, as well as regions around transcribed regions. Implicated tissues include the fetal brain and adult brain (the frontal cortex, specifically), and the fetal thymus (which may be approximating a more interpretable annotation).

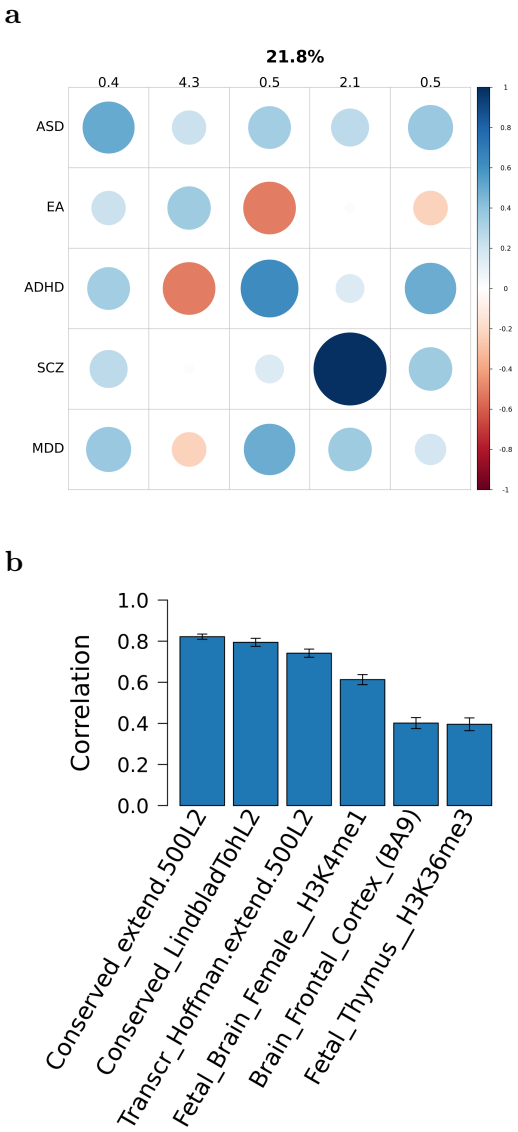


Figure 4.10: Interpretation of the inferred functional module for the ASD dataset. 21.8% of the training SNPs are expected to be non-null. In **a**, like before, column names are the variances of the z-scores for each trait, the diagonals are per-SNP heritabilities relative to their maximal value across traits and components, and the off-diagonals are genetic correlations. In **b**, the y-axis is the Spearman correlation of the LD annotations with the prior probabilities for belonging to the non-null component, and the error bars are standard deviations based on 100 bootstrap samples.

We found 415 new independent loci associated to ASD. These include the hits in chromosomes 8 and 20 found in the latest GWAS for ASD using Plink (not including their follow-up sample) [81], as well as hits found with MTAG by the same authors. An example of the latter is our strongest hit, rs2388334 (PPA=0.999994), in chromosome 6, which was identified by an MTAG meta-analysis of ASD and EA

(p-value $= 3.34 \times 10^{-12}$). In our Bayesian GWAS, rs2388334 is supported by both non-focal traits (including EA) and LD annotations. A second example is our second strongest hit, rs1452075 (PPA=0.99993), in chromosome 3, which was also associated to ASD in [81] by the MTAG meta-analysis with EA (p-value $= 3.17 \times 10^{-9}$). Again, in our GWAS, rs1452075 is supported by both non-focal traits and the prior probabilities of the SNP. Like for CAD, a proper investigation of the newly found hits is an important direction of future work.

Figure 4.12 shows the Manhattan plots for the remaining traits. 2010 new independent loci were associated to EA, 1301 to ADHD, 1790 to SCZ, and 719 to MDD, which will require extensive dedicated analysis in the future.

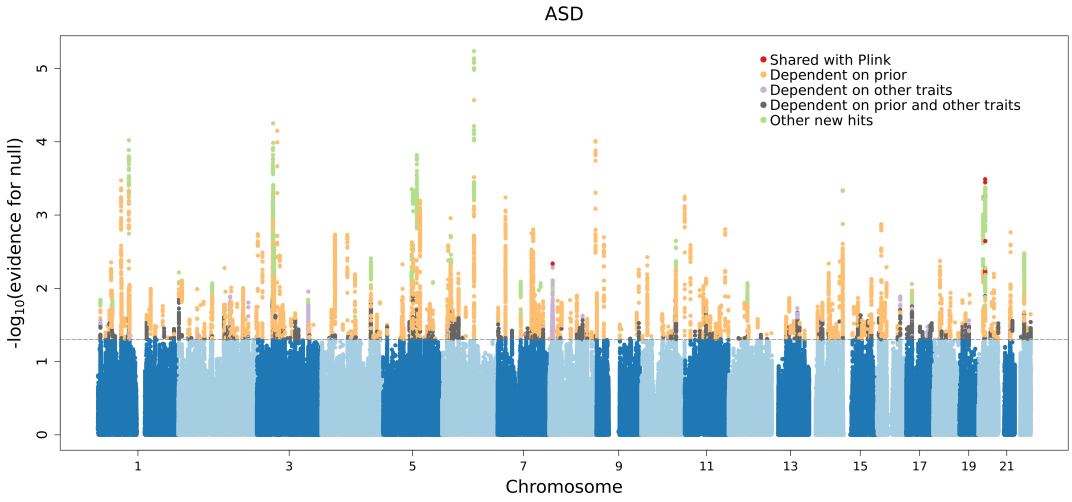


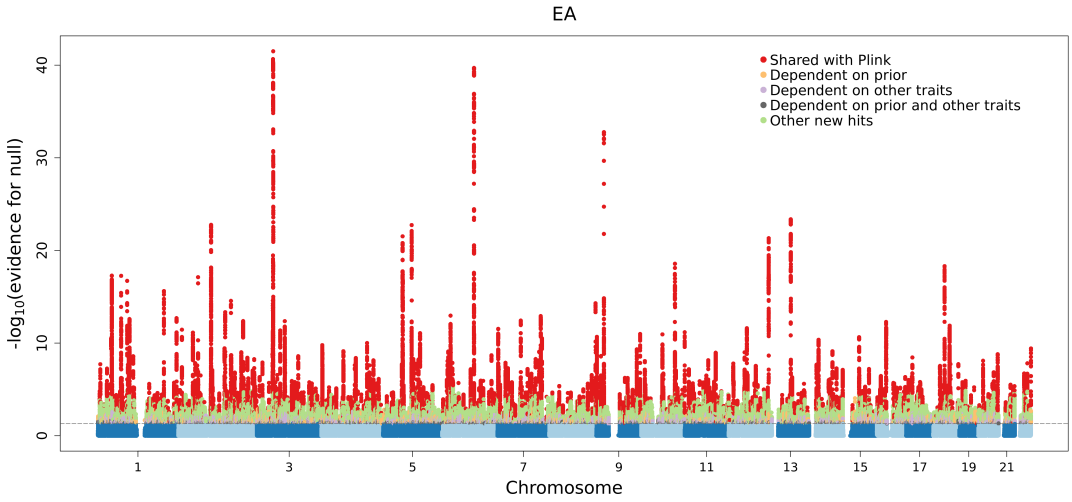
Figure 4.11: Manhattan plot for ASD based on three inferred non-null component.

Despite the promising results given by our Bayesian GWAS, it is important to note that our newly found hits have to be taken with caution until further guarantees of controlled type-I error are provided. Specifically, although we properly controlled type-I error in our simulated (Table 3.1), the ASD dataset diverges from our simulated datasets in that it includes binary traits with very low prevalence, like ASD (prevalence $\approx 1\%$[7]) or SCZ (prevalence $\approx 0.3\%$[8]). A very low prevalence may deviate the distribution of the z-scores from the assumed normality, which may in turn affect our ability to recover the true components or to output calibrated

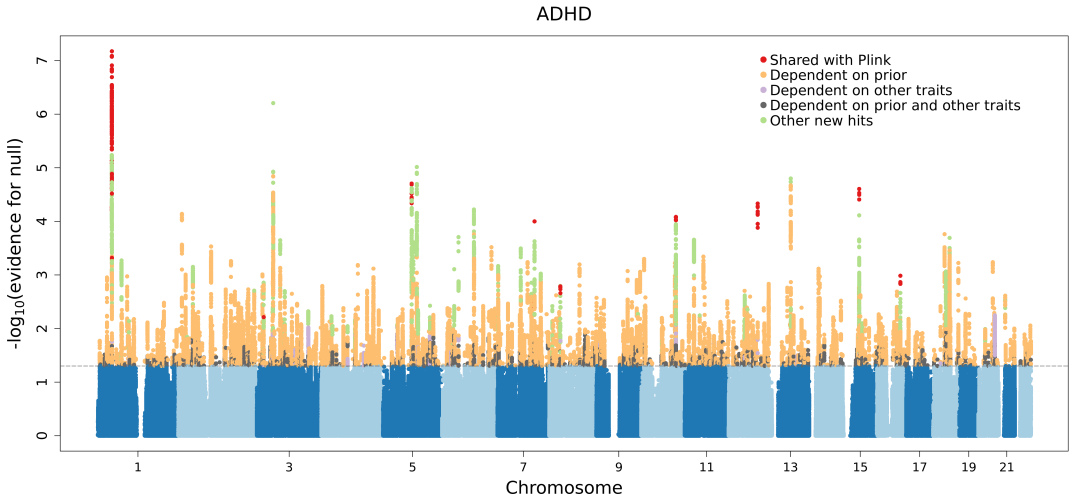
[7]https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders
[8]https://www.who.int/news-room/fact-sheets/detail/schizophrenia

GWAS results. In chapter 5, I further discuss the limitations of the simulations and suggest a more direct way of validating the new hits using polygenic risk scores.

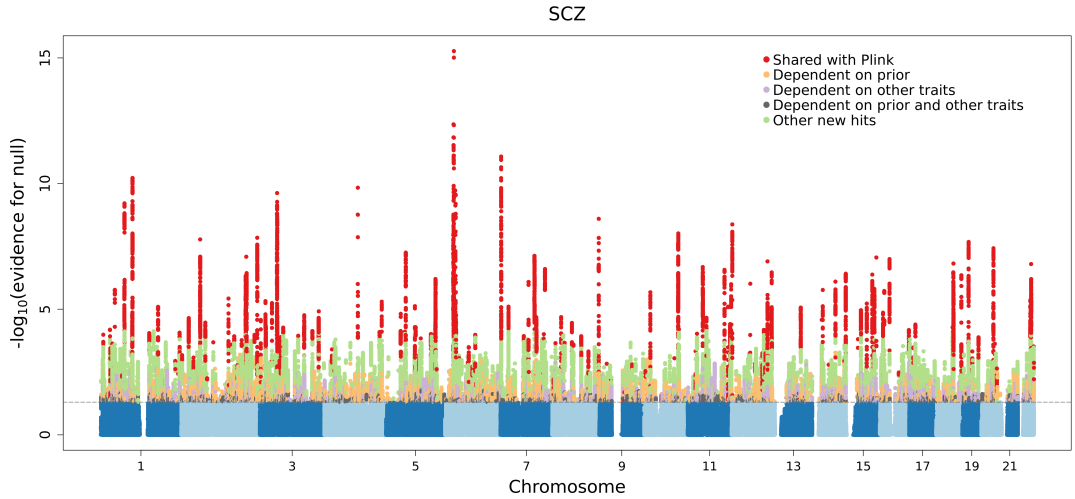
a



b



c



d

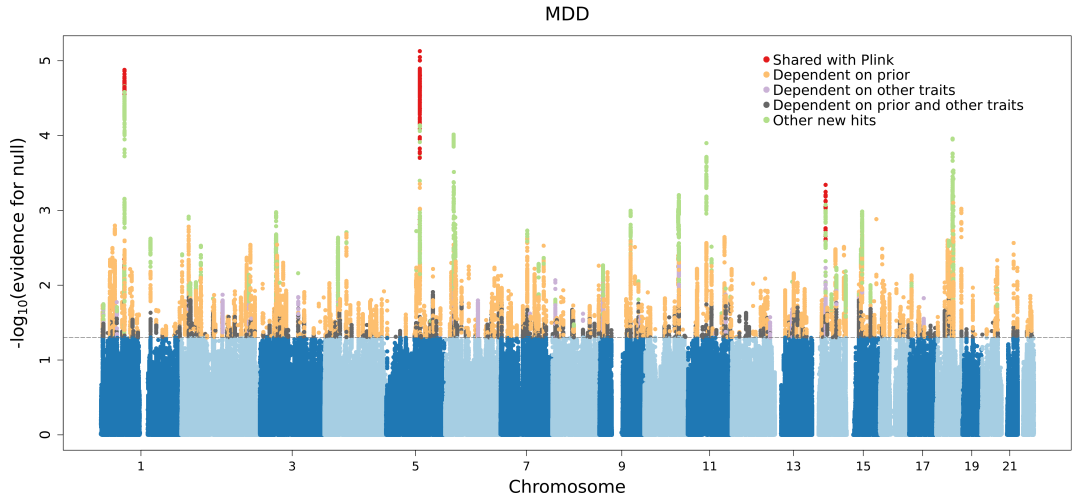


Figure 4.12: Manhattan plots for EA, ADHD, SCZ and MDD.

5 Discussion and future work

Contents

5.1 Summary of our contribution and results

In this thesis, we introduce a new mixture model that boosts GWAS power and clusters SNPs based their functional similarities using two sources of information: multi-trait GWAS z-scores and SNP functional annotations. This development responds to the need for methods that integrate the fast-growing catalogues of GWAS summary statistics and functional annotations of SNPs to uncover the mechanisms by which common genetic variation impacts complex traits. More generally, our method responds to the need for a better understanding of the

heritable component of complex traits, in order to catalyse the development of pharmacological interventions [23, 24].

In the methods chapter, we show how our mixture model specifically decomposes the additive heritability of complex traits by deriving the model from an assumed linear relationship between genotypes and complex traits. We validate our inference procedure with a range of simulated realistic datasets, and focus on two of them for validating our analysis pipeline.

The main goal of the simulations analysis was to interpret the inferred components functionally and to populate them with confidently assigned hits for the different traits. From the two selected datasets, we managed to recover and interpret correctly the simulated components. We also associated many more truly causal SNPs to each trait than linear regression (with Plink) or MTAG for the same false positive rate and similar false discovery rate, and assigned most of the true hits to the right component. We were also able to identify the source of evidence driving each hit (either the focal trait, the non-focal traits, the annotations, or some combination of these).

Our analysis concluded with two real datasets that focus on Coronary Artery Disease (CAD) and Autism Spectrum Disorder (ASD). We recovered three non-null components for the CAD dataset, which can be interpreted as conferring risk to CAD indirectly by regulating LDL, the probability of having smoked regularly, and systolic blood pressure, respectively. For ASD, we found only one non-null component implicating the fetal and adult brain. We found many independent hits for CAD and ASD, some of which replicate the new findings of a newer GWAS for CAD or of one of the latest meta-analysis of ASD with other traits [12, 81]. Some of these examples were supported by more than one source of evidence in our Bayesian GWAS.

Overall, despite the limitations that will be discussed below, our main contribution is a method with which to integrate the increasingly available GWAS and genomic functional data to generate insights about the biology of complex traits. We expect this method to be increasingly useful as more data becomes available,

providing a mechanism to produce novel insights into the biology of CAD and ASD, and any other complex trait of interest.

5.2 Future improvements of the analyses

The main future improvements of the analyses presented in this thesis consist of (i) the validation of the new hits found with the Bayesian GWAS, (ii) the re-analysis of CAD and ASD with more informative datasets, (iii) a more complete comparison to other methods, and (iv) the identification of subtler, biologically meaningful components.

5.2.1 Validation and investigation of the new GWAS hits

Based on our simulation results, the 0.05 significance threshold used for the Bayesian GWAS should keep the false discovery rate below 0.05. Prioritising hits that are supported by both z-scores and annotations may further reduce this number. To have increased confidence that the majority of new hits found for the analysed real traits are true, we could compare the predictive power of a polygenic score built with our hits to other existing polygenic scores. Additionally, showing that type-I error is properly controlled in a wider range of simulations, such as when traits are binary and have low prevalence or when genetic variants are rare, would provide further guarantees that the newly found hits are true positives. The simulations should also be extended to include more traits and components, a wider range of polygenicity and heritability, components with different sizes and strengths, as well as weaker and missing annotations.

Once the new hits have been validated, the individual investigation of prioritised hits (by linking them to likely affected functional elements in the genome) could add to the interpretation of the inferred components and ultimately provide further insights into the driving mechanisms of the studied complex traits.

5.2.2 Analysis of CAD and ASD with more informative datasets

Better data mainly means more relevant traits and annotations. Relevant traits are those that, alone or in combination, help to pull apart components with some heritability for the focal trait by having z-scores that are distributed differently across components. Obviously, the greater the power of their GWASes, the better as well. In practice, an heuristic way for selecting potentially relevant traits consists of choosing non-redundant (genetically uncorrelated) traits that are genetically correlated with the focal trait. For CAD, we could further include in the analysis traits related to the correct functioning of artery walls (e.g. vascular endothelial growth factor [92]) and to inflammation (e.g. IL-6 [93] or more broadly body mass index [94]), as these have been linked to CAD [21] and are generally genetically correlated with it (figure B.2). For ASD, we could use responses to diagnostic or mental health questionnaires (e.g. from iHART, MSSNG, SFARI SPARK and UK Biobank) and other low-level cognitive and personality traits like intelligence [95], systemising [87], neuroticism [76], chronotype [76], and others (figure B.3). In general, the lower level traits are, the more likely they will be non-redundant: consider for example the differences between responses to diagnostic questionnaires and the psychiatric traits used in this thesis, which combine overlapping lower-level symptoms. This was the approach followed by Lucía de Hoyos *et al.* in [85], who found two to three components for ASD but using a range of relatively low-level traits related to development (e.g. age of crawling or of self-feeding), language (e.g. language level), repetitive behaviour (e.g. self-injurious behaviour), and others.

Relevant annotations are those that mark as accurately as possible, alone or in combination, the functional elements where causal SNPs are located. Future work will involve compiling more relevant annotations for CAD and ASD, at the single-cell resolution and across developmental stages for ASD if possible. In our study, most of the brain-related annotations were for adult brains and broad brain regions, with the only fetal annotations mostly aggregating the entire brain and only distinguishing between male and female fetal brains (Table A.2).

5.2.3 Comprehensive comparison to other methods

In chapter 1, we compared the functionality of several methods that build on GWAS results to better understand the biology of complex traits. we claimed that our mixture model is more complete than the rest, and in chapter 3 we showed how its features generally translate into greater GWAS power than linear regression and MTAG. In the context of boosting GWAS power, however, there are more powerful and popular association methods that can be use in practice, like Bolt-LMM [64]. Using Bolt-LMM as baseline instead of Plink, is a straightforward substitution and would lead to a more realistic assessment of the power gained by our mixture model.

Given the relatively few resources needed by our mixture model, other comparisons should probably only be prioritised if they competed in convenience or popularity. We highlight g-SEM for its recent gain in popularity [59, 85]. Comparing its inferred functional modules to the ones found by our mixture model in cases where g-SEM is well-defined, and assessing the extent to which the identifiability constraints of g-SEM are limiting in practice, would be useful to guide practice and to better understand both frameworks. PDR [60] could also be an insightful choice, as it directly competes with g-SEM and resembles our MM in that it also uses a mixture of multivariate Gaussians. While PDR does not use functional annotations, it allows SNPs to belong to multiple components, which is biologically plausible but not supported by our MM (although it is accounted to some extent by our inferred posterior probabilities).

On a final note and significantly deviating from our focus on relatively simple methods that rely on GWAS summary statistics and a linear association model, it would be perhaps useful to explore the prospects of other emerging frameworks based on non-linear models and individual-level data, like for example Deep Structured Phenotype Network (DSPN) [49]. DSPN integrates multi-omics and multi-trait data in an interpretable deep learning framework, and finds mechanisms in the form of implicated combinations of pre-defined SNP-gene-enhancer regulatory networks, changes in cell-type proportions, and gene expression modules.

5.2.4 Identification of subtler, meaningful components

So far, we have estimated the number of non-null components using the cross-validation (CV) likelihood: we chose the smallest number of components that precedes the saturation of the likelihood and that does not include an artefact due to LD (see the next section). However, the saturation of the CV likelihood does not imply the absence of additional biologically meaningful components, as these could be either weak or strong but made of small proportions of SNPs, either way increasing the overall likelihood only marginally. An important next step would be to run CV with a greater range of components and to devise an automated way of deciding which components are not artefacts due to LD, as it may not be obvious by visual inspection. So far, we have suspected that components are artefacts when their covariance matrices could be explained by a combination of the covariance matrices of the other components. However, as we increase the number of components, a component may for example split into two components with similar genetic correlations but scaled per-SNP heritabilities to capture two different mechanisms. Taking into account the annotations and the size of the components should help to identify the likely biologically meaningful components among artefacts. Finally, further simulations with subtler components would be needed to investigate this so far unexplored type of realistic scenario.

5.3 Limitations of our method

The main limitations of our method relate to its assumptions and to its ability to screen collections of annotations for relevant and non-redundant ones.

5.3.1 Model assumptions and scalability

The main assumptions that we are making are that:

1. The SNPs used for inference are independent (only for the likelihood function).
2. The variability of z-score variance within a component can be explained by the variability of the LD annotations of SNPs (through the prior probabilities).

3. All the GWAS use individuals with the same ancestry.

4. The z-scores are normally distributed.

5. The variance of the true SNP effects is proportional to $[p(1-p)]^{\alpha}$, with $\alpha = -1$ and where p is the SNP frequency.

6. A SNP can only belong to one component.

We reduce the impact of the first assumption by weighting the contribution of SNPs to the overall likelihood as a function of their LD scores (so that the greater the LD score, the smaller the weights). This is the same approach used by LD score regression [67], seems to work well with our simulations, and keeps inference fast.

The second assumption leads in practice to extra components to explain the z-scores of the affected SNPs (what we refer to as 'artefacts due to LD'). This can happen when multiple causal SNPs cluster in the same locus or when SNPs are in weak LD with strong hits. Based on our simulations with realistic numbers of causal SNPs and heritabilities, the assumption seems to be well tolerated and the strongest artefacts are easily identified (as they have all the z-score variances scaled up and are combinations of other components). The latter may however no longer be true as we consider larger numbers of components that include both artefacts and weak, biologically meaningful components. As noted in the previous section, further simulations with weaker true components may help to explore the impact of LD artefacts on our ability to infer them correctly.

Violating the third assumption would result in noisier LD annotations, reducing the ability of the prior to assign SNPs to components. Dealing with multiple ancestries may be possible by treating z-scores and LD annotations based on different ancestries as different traits and annotations and analysing all of them jointly. This is another area of future research.

The fourth assumption may be significantly violated when considering rare genetic variants (MAF $< 0.1\%$) and binary traits with low prevalence ($< 10\%$). An important next step consists of assessing and addressing (if needed) the impact of

low MAF and low prevalence in the ability of the MM to infer the right components and to output calibrated posterior probabilities of association.

The fifth assumption was implicitly made by constraining the true effects of standardised SNPs in each component to have constant variance. Although we assumed that $\alpha = -1$, α is probably closer to -0.35 for many complex traits, as it has been previously estimated [27, 96]. In brief, this means that our model derivation relies on an overestimation of the contribution of rare variants to the heritability of complex traits. Although we do not expect this approximation to affect the results significantly, this would be easy to test by re-doing the simulated GWAS after standardising the SNPs using different powers of their standard deviation (i.e. with different values of α).

Regarding the sixth assumption, it is biologically plausible that two different mechanisms are supported by two overlapping groups of SNPs. In such cases, we expect our model to explain the intersection of SNPs by either allocating an extra component to it or by inferring posterior probabilities that reflect the ambiguity about group memberships. The robustness of our MM to the existence of overlapping groups of SNPs could be easily tested by simulating correlated prior probabilities and sampling group memberships with replacement.

Finally, note that we are locally optimising the (approximate) likelihood of the model, relying on an heuristic initialisation of the parameters. This has worked well with simulations so far, but it may need some revision as we scale up the number of traits, components or annotations.

5.3.2 Screening of large catalogues of SNP annotations

Even if we had access to idealised catalogues of GWAS summary statistics and SNP annotations, we would still have to commit to a reduced relevant dataset to make inference manageable. Our current approach is to screen traits for genetic correlations with the focal trait of interest, and SNP annotations for some evidence of heritability enrichment for any of the analysed traits. Redundant traits are those that are very genetically correlated with other traits and should be discarded;

redundant LD annotations are also very correlated ones and are discarded via our 'stratified forward LD score regression', which starts with only one annotation and sequentially adds more annotations until these stop explaining much more heritability. The proposed screening for traits is probably enough for now, but our annotation screening strategy is likely too strict, discarding annotations of weak and component-specific relevance. One reason for this is that stratified LD score regression assumes a single overall component, so weak and component-specific annotation enrichments may be diluted and not detectable when doing so. Other reasons are that we currently require annotations to have p-values for heritability enrichment smaller than 0.05, we initially use a sub-sample of 300,000 SNPs to speed up the otherwise too slow screening procedure, and we require annotations to improve predictions by 1% in each step of the forward regression to be included in the model. Further testing our annotation screening strategy in more challenging scenarios (e.g. where many weak annotations matter) and improving it if needed is an important area of future work.

Appendices

A
Annotation tables

Table A.1: Summary of baseline annotations. There is an additional annotation for each of these for 500kb windows around them.

Base ('1' for all SNPs)	H3K4me3 (Trynka)
Coding (UCSC)	H3K9ac peaks (Trynka)
Conserved (LindbladToh)	H3K9ac (Trynka)
CTCF (Hoffman)	Intron (UCSC)
DGF (ENCODE)	Promoter-Flanking (Hoffman)
DHS peaks (Trynka)	Promoter (UCSC)
DHS (Trynka)	Repressed (Hoffman)
Enhancer (Andersson)	Super Enhancer (Hnisz)
Enhancer (Hoffman)	TFBS (ENCODE)
FetalDHS (Trynka)	Transcr (Hoffman)
H3K27ac (Hnisz)	TSS (Hoffman)
H3K27ac (PGC2)	UTR-3 (UCSC)
H3K4me1 peaks (Trynka)	UTR-5 (UCSC)
H3K4me1 (Trynka)	Weak Enhancer (Hoffman)
H3K4me3 peaks (Trynka)	

Table A.2: Summary of tissue-specific annotations sorted alphabetically by tissue. Each of these annotations in practice expands into up to five annotations with different epigenetic marks (DHS, H3K4me1, H3K4me3, H3K36me3, H3K9ac, H3K27ac), depending on availability. PB: peripheral blood; PC: primary cultured; SC: stem cell; CB:cord blood; CV: cardiovascular; GI: gastrointestinal; CNS: central nervous system.

Annotation name	Tissue group	Source
Adipose Subcutaneous	adipose	GTEx
Adipose Visceral Omentum	adipose	GTEx
Adipose Nuclei	adipose	Roadmap
Cells EBV transformed lymphocytes	blood/immune	GTEx
Spleen	blood/immune	GTEx
Whole Blood	blood/immune	GTEx
Primary monocytes PB	blood/immune	Roadmap
Primary B cells PB	blood/immune	Roadmap
Primary T cells CB	blood/immune	Roadmap
Primary T cells PB	blood/immune	Roadmap
Primary NK cells PB	blood/immune	Roadmap
Primary hematopoietic SC Female	blood/immune	Roadmap
Primary hematopoietic SC Male	blood/immune	Roadmap
Primary T helper memory cells PB 2	blood/immune	Roadmap
Primary T helper naive cells PB	blood/immune	Roadmap
Primary T helper memory cells PB 1	blood/immune	Roadmap
Primary T helper cells PMA.I_stimulated	blood/immune	Roadmap
Primary T helper 17 cells PMA.I stimulated	blood/immune	Roadmap
Primary T helper cells PB	blood/immune	Roadmap
Primary T regulatory cells PB	blood/immune	Roadmap
Primary T cells effector.memory enriched PB	blood/immune	Roadmap
Primary T killer naive cells PB	blood/immune	Roadmap
Primary T killer memory cells PB	blood/immune	Roadmap
Primary mononuclear cells PB	blood/immune	Roadmap

Continued on next page

Table A.2 – continued from previous page

Annotation name	Second column	Third column
Thymus	blood/immune	Roadmap
Spleen	blood/immune	Roadmap
Primary neutrophils PB	blood/immune	Roadmap
Primary B cells CB	blood/immune	Roadmap
Primary hematopoietic SC	blood/immune	Roadmap
Primary hematopoietic SC short-term culture	blood/immune	Roadmap
Transformed fibroblasts	bone/connective	GTEx
Foreskin Fibroblast Primary Cells skin 01	bone/connective	Roadmap
Foreskin Fibroblast Primary Cells skin 02	bone/connective	Roadmap
AdulT Dermal FibroblasT Primary Cells	bone/connective	Roadmap
Lung FibroblasT Primary Cells	bone/connective	Roadmap
Osteoblast Primary Cells	bone/connective	Roadmap
Brain Amygdala	CNS	GTEx
Brain Anterior cingulate cortex (BA24)	CNS	GTEx
Brain Caudate basal ganglia	CNS	GTEx
Brain Cerebellar Hemisphere	CNS	GTEx
Brain Cerebellum	CNS	GTEx
Brain Cortex	CNS	GTEx
Brain Frontal Cortex_.BA9.	CNS	GTEx
Brain Hippocampus	CNS	GTEx
Brain Hypothalamus	CNS	GTEx
Brain Nucleus accumbens basal ganglia.	CNS	GTEx
Brain Putamen basal ganglia	CNS	GTEx
Brain Spinal cord cervical (c.1)	CNS	GTEx
Brain Substantia nigra	CNS	GTEx
Nerve Tibial	CNS	GTEx

Continued on next page

Table A.2 – continued from previous page

Annotation name	Second column	Third column
Brain Angular Gyrus	CNS	Roadmap
Brain Anterior Caudate	CNS	Roadmap
Brain Cingulate Gyrus	CNS	Roadmap
Brain Hippocampus Middle	CNS	Roadmap
Brain Inferior Temporal Lobe	CNS	Roadmap
Brain Dorsolateral Prefrontal Cortex	CNS	Roadmap
Brain Substantia Nigra	CNS	Roadmap
Cortex derived PC neurospheres	CNS	Roadmap
Ganglion Eminence derived PC neurospheres	CNS	Roadmap
Brain Germinal Matrix	CNS	Roadmap
Artery Aorta	CV	GTEx
Artery Coronary	CV	GTEx
Artery Tibial	CV	GTEx
Heart Atrial Appendage	CV	GTEx
Heart LefT Ventricle	CV	GTEx
Aorta	CV	Roadmap
LefT Ventricle	CV	Roadmap
RighT Atrium	CV	Roadmap
RighT Ventricle	CV	Roadmap
Adrenal Gland	endocrine	GTEx
Pancreas	endocrine	GTEx
Pituitary	endocrine	GTEx
Thyroid	endocrine	GTEx
Pancreas	endocrine	Roadmap
Pancreatic Islets	endocrine	Roadmap
Skin Not Sun Exposed Suprapubic	epithelial	GTEx

Continued on next page

Table A.2 – continued from previous page

Annotation name	Second column	Third column
Skin Sun Exposed Lower leg	epithelial	GTEx
Breast Mammary Epithelial Cells	epithelial	Roadmap
Foreskin Keratinocyte Primary Cells skin 02	epithelial	Roadmap
Foreskin Melanocyte Primary Cells skin 01	epithelial	Roadmap
Mammary Epithelial Primary Cells	epithelial	Roadmap
Epidermal Keratinocyte Primary Cells	epithelial	Roadmap
Foreskin Keratinocyte Primary Cells skin 03	epithelial	Roadmap
Foreskin Melanocyte Primary Cells skin 03	epithelial	Roadmap
BreasT Myoepithelial Primary Cells	epithelial	Roadmap
Cervix Ectocervix	female	GTEx
Cervix Endocervix	female	GTEx
Fallopian Tube	female	GTEx
Ovary	female	GTEx
Uterus	female	GTEx
Vagina	female	GTEx
Ovary	female	Roadmap
Fetal Adrenal Gland	fetal	Roadmap
Fetal Brain Male	fetal	Roadmap
Fetal Brain Female	fetal	Roadmap
Fetal Heart	fetal	Roadmap
Fetal Intestine Large	fetal	Roadmap
Fetal Intestine Small	fetal	Roadmap
Fetal Kidney	fetal	Roadmap
Fetal Lung	fetal	Roadmap
Fetal Muscle Trunk	fetal	Roadmap
Fetal Muscle Leg	fetal	Roadmap

Continued on next page

Table A.2 – continued from previous page

Annotation name	Second column	Third column
Placenta	fetal	Roadmap
Fetal Stomach	fetal	Roadmap
Fetal Thymus	fetal	Roadmap
Placenta Amnion	fetal	Roadmap
Colon Sigmoid	GI	GTEx
Colon Transverse	GI	GTEx
Esophagus Gastroesophageal Junction	GI	GTEx
Esophagus Mucosa	GI	GTEx
Esophagus Muscularis	GI	GTEx
Small Intestine Terminal Ileum	GI	GTEx
Stomach	GI	GTEx
Gastric	GI	Roadmap
Small Intestine	GI	Roadmap
Colonic Mucosa	GI	Roadmap
Colon Smooth Muscle	GI	Roadmap
Duodenum Smooth Muscle	GI	Roadmap
Esophagus	GI	Roadmap
Rectal Mucosa Donor 29	GI	Roadmap
Rectal Mucosa Donor 31	GI	Roadmap
Rectal Smooth_Muscle	GI	Roadmap
Sigmoid Colon	GI	Roadmap
Stomach Smooth Muscle	GI	Roadmap
Duodenum Mucosa	GI	Roadmap
Stomach Mucosa	GI	Roadmap
Kidney Cortex	kidney	GTEx
Liver	liver	GTEx

Continued on next page

Table A.2 – continued from previous page

Annotation name	Second column	Third column
Liver	liver	Roadmap
Lung	lung	GTEx
Lung	lung	Roadmap
Prostate	male	GTEx
Testis	male	GTEx
BreasT Mammary Tissue	other	GTEx
Minor Salivary Gland	other	GTEx
Muscle Skeletal	skeletal muscle	GTEx
Psoas Muscle	skeletal muscle	Roadmap
Skeletal Muscle Female	skeletal muscle	Roadmap
Skeletal Muscle Male	skeletal muscle	Roadmap
Bladder	urinary	GTEx

B

Supplementary figures

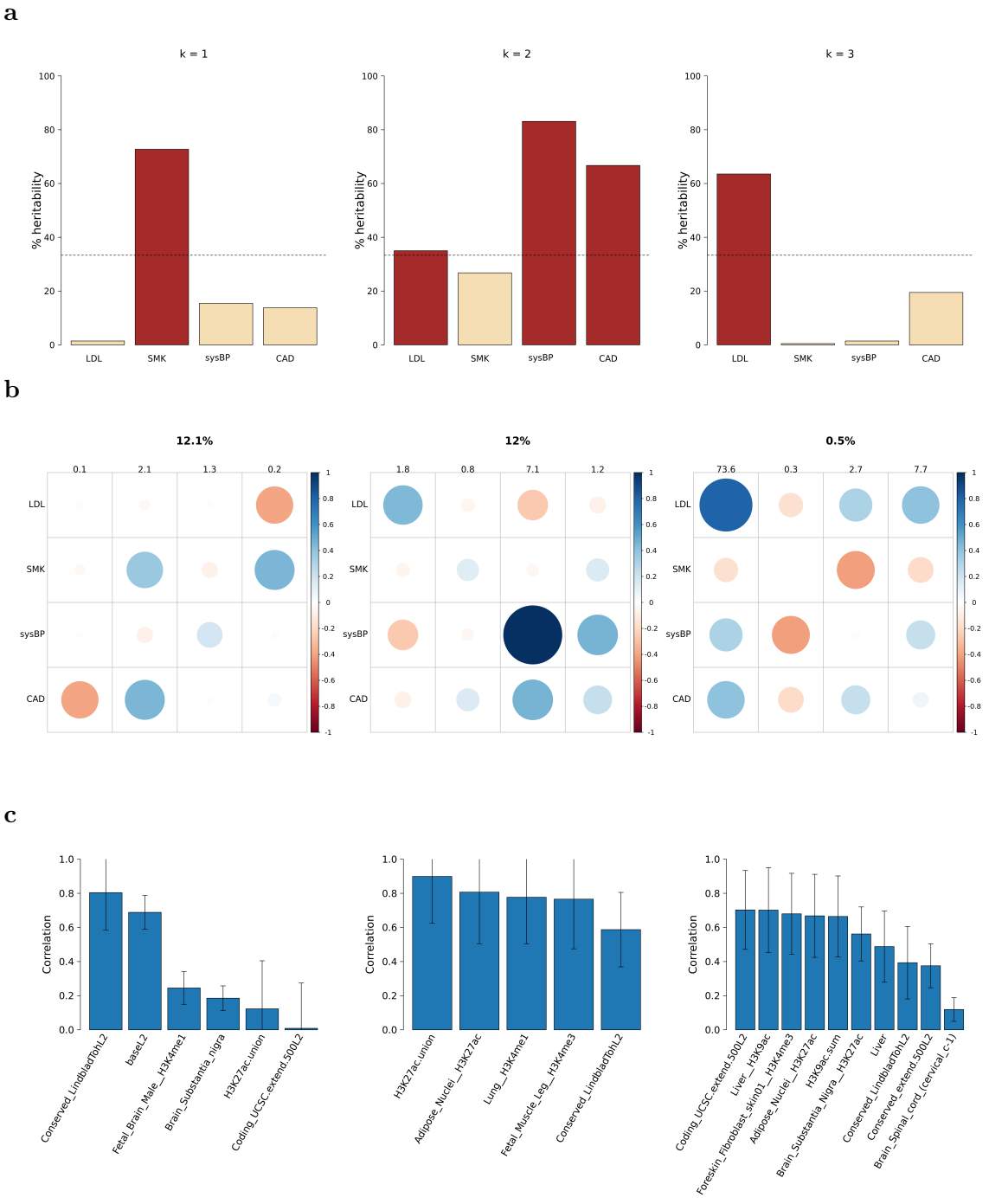


Figure B.1: Interpretation of functional modules for an extended CAD dataset with systolic blood pressure (from a 2018 GWAS based on 757,601 European individuals [97]) and a larger GWAS for CAD from 2017 (122,733 cases) [98].

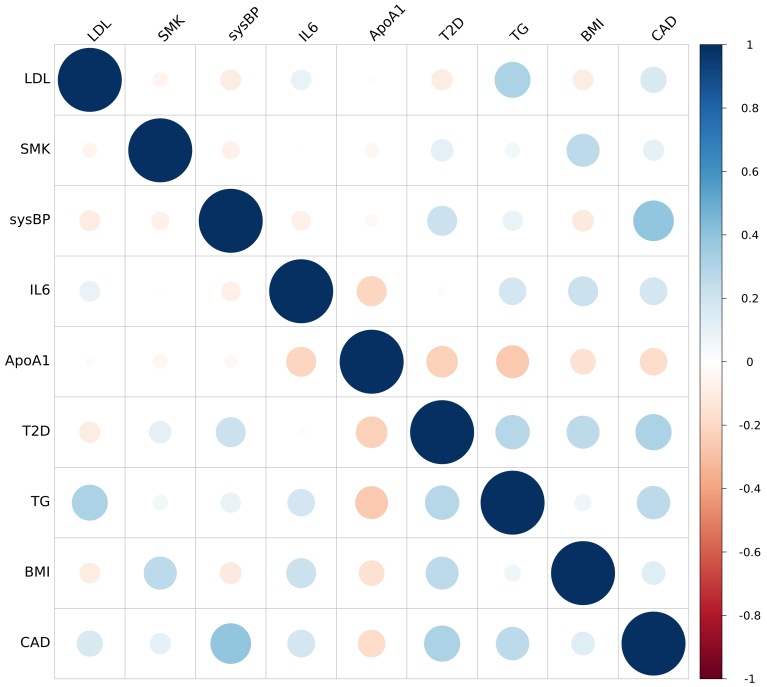


Figure B.2: Genetic correlations of multiple CAD-related traits, estimated with our mixture model using one single non-null component ($K = 1$). LDL: low-density lipoprotein cholesterol [77]; SMK: past smoking [13]; sysBP: systolic blood pressure [97]; IL6: interleukin 6 [93]; ApoA1: apolipoprotein A-1 [77]; T2D: type-2 diabetes [99]; TG: triglycerides [77]; BMI: body-mass index [94]; CAD: coronary artery disease [98].

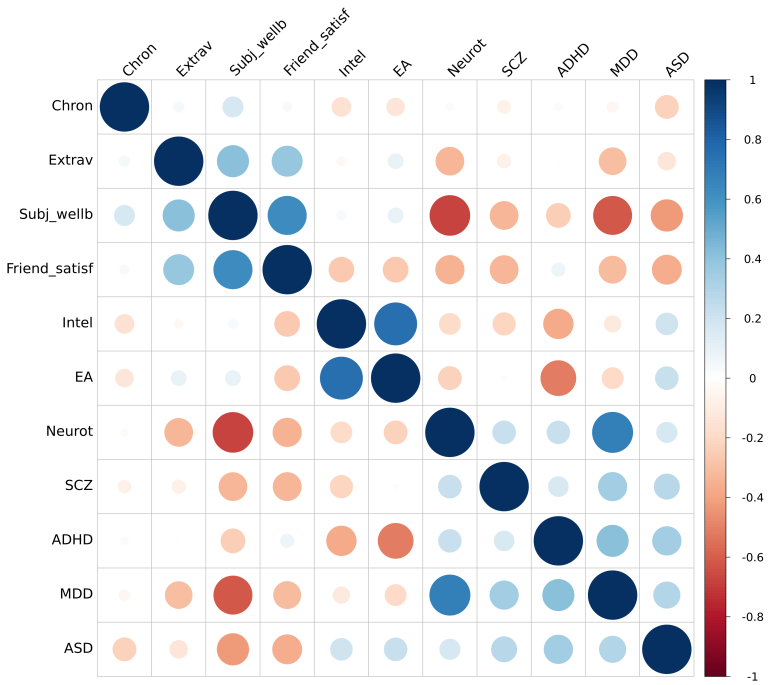


Figure B.3: Genetic correlations of multiple ASD-related traits, estimated with our mixture model using one single non-null component ($K = 1$). Chron: chronotype (ukb-b-4956 dataset in the the ieu-GWAS database [76]); Extrav: extroversion [100]; Subj_wellb: subjetive well-being [101]; Friend_satisf: friendships satisfaction (ukb-a-371 dataset in ieu-GWAS [76]); Intel: intelligence [95]; EA: educational attainment [88]; Neurot: neuroticism (ukb-b-4630 in ieu-GWAS [76]); SCZ: schizophrenia [91]; ADHD: attention deficit hyperactivity disorder [89]; MDD; major depressive disorder [90]; ASD: autism spectrum disorder [81].

References

[1] Tinca JC Polderman et al. “Meta-analysis of the heritability of human traits based on fifty years of twin studies”. In: *Nature genetics* 47.7 (2015), pp. 702–709.

[2] Tian Ge et al. “Phenome-wide heritability analysis of the UK Biobank”. In: *PLoS genetics* 13.4 (2017), e1006711.

[3] Kyoko Watanabe et al. “A global overview of pleiotropy and genetic architecture in complex traits”. In: *Nature genetics* 51.9 (2019), pp. 1339–1348.

[4] 1000 Genomes Project Consortium et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), p. 68.

[5] Peter M Visscher et al. “10 years of GWAS discovery: biology, function, and translation”. In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22.

[6] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. “An expanded view of complex traits: from polygenic to omnigenic”. In: *Cell* 169.7 (2017), pp. 1177–1186.

[7] Emil Uffelmann et al. “Genome-wide association studies”. In: *Nature Reviews Methods Primers* 1.1 (2021), pp. 1–21.

[8] Iftikhar J Kullo et al. “Polygenic scores in biomedical research”. In: *Nature Reviews Genetics* (2022), pp. 1–9.

[9] Annalisa Buniello et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019”. In: *Nucleic acids research* 47.D1 (2019), pp. D1005–D1012.

[10] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562.7726 (2018), pp. 203–209.

[11] Loıc Yengo et al. “A saturated map of common genetic variants associated with human height”. In: *Nature* 610.7933 (2022), pp. 704–712.

[12] Krishna G Aragam et al. “Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants”. In: *Nature Genetics* (2022), pp. 1–13.

[13] Mengzhen Liu et al. “Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use”. In: *Nature genetics* 51.2 (2019), pp. 237–244.

[14] James J Lee et al. “Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals”. In: *Nature genetics* 50.8 (2018), pp. 1112–1121.

[15] Michael D Gallagher and Alice S Chen-Plotkin. "The post-GWAS era: from association to function". In: *The American Journal of Human Genetics* 102.5 (2018), pp. 717–730.

[16] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. "From genome-wide associations to candidate causal variants by statistical fine-mapping". In: *Nature Reviews Genetics* 19.8 (2018), pp. 491–504.

[17] Christiaan A De Leeuw et al. "The statistical properties of gene-set analysis". In: *Nature Reviews Genetics* 17.6 (2016), pp. 353–364.

[18] Christiaan A de Leeuw et al. "MAGMA: generalized gene-set analysis of GWAS data". In: *PLoS computational biology* 11.4 (2015), e1004219.

[19] Tune H Pers et al. "Biological interpretation of genome-wide association studies using predicted gene functions". In: *Nature communications* 6.1 (2015), pp. 1–9.

[20] Nasa Sinnott-Armstrong et al. "GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background". In: *Elife* 10 (2021).

[21] Amit V Khera and Sekar Kathiresan. "Genetics of coronary artery disease: discovery, biology and clinical translation". In: *Nature Reviews Genetics* 18.6 (2017), pp. 331–344.

[22] Allison B Goldfine. "Statins: is it really time to reassess benefits and risks?" In: *New England Journal of Medicine* 366.19 (2012), pp. 1752–1755.

[23] Matthew R Nelson et al. "The support of human genetic evidence for approved drug indications". In: *Nature genetics* 47.8 (2015), pp. 856–860.

[24] Emily A King, J Wade Davis, and Jacob F Degner. "Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval". In: *PLoS genetics* 15.12 (2019), e1008489.

[25] Lilia M Iakoucheva, Alysson R Muotri, and Jonathan Sebat. "Getting to the Cores of Autism". In: *Cell* 178.6 (2019), pp. 1287–1298.

[26] Elizabeth K Ruzzo et al. "Inherited and de novo genetic risk for autism impacts shared networks". In: *Cell* 178.4 (2019), pp. 850–866.

[27] Jian Zeng et al. "Signatures of negative selection in the genetic architecture of human complex traits". In: *Nature genetics* 50.5 (2018), pp. 746–753.

[28] Dominic Holland et al. "Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model". In: *PLoS Genetics* 16.5 (2020), e1008612.

[29] Matthew T Maurano et al. "Systematic localization of common disease-associated variation in regulatory DNA". In: *Science* 337.6099 (2012), pp. 1190–1195.

[30] Aviv Regev et al. "Science forum: the human cell atlas". In: *elife* 6 (2017), e27041.

[31] Peter M Visscher, William G Hill, and Naomi R Wray. "Heritability in the genomics era—concepts and misconceptions". In: *Nature reviews genetics* 9.4 (2008), pp. 255–266.

[32] Molly Gasperini, Jacob M Tome, and Jay Shendure. "Towards a comprehensive catalogue of validated and target-linked human enhancers". In: *Nature Reviews Genetics* 21.5 (2020), pp. 292–310.

[33] Hilary K Finucane et al. “Partitioning heritability by functional annotation using genome-wide association summary statistics”. In: *Nature genetics* 47.11 (2015), pp. 1228–1235.

[34] Hilary K Finucane et al. “Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types”. In: *Nature genetics* 50.4 (2018), pp. 621–629.

[35] Xiang Zhu, Zhana Duren, and Wing Hung Wong. “Modeling regulatory network topology improves genome-wide analyses of complex human traits”. In: *Nature communications* 12.1 (2021), pp. 1–15.

[36] Joseph K Pickrell. “Joint analysis of functional genomic data and genome-wide association studies of 18 human traits”. In: *The American Journal of Human Genetics* 94.4 (2014), pp. 559–573.

[37] Bradley E Bernstein et al. “The NIH roadmap epigenomics mapping consortium”. In: *Nature biotechnology* 28.10 (2010), pp. 1045–1048.

[38] Anshul Kundaje et al. “Integrative analysis of 111 reference human epigenomes”. In: *Nature* 518.7539 (2015), pp. 317–330.

[39] Jian Zhou et al. “Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk”. In: *Nature genetics* 50.8 (2018), pp. 1171–1179.

[40] Carles A Boix et al. “Regulatory genomic circuitry of human disease loci by integrative epigenomics”. In: *Nature* 590.7845 (2021), pp. 300–307.

[41] Jason Ernst and Manolis Kellis. “Chromatin-state discovery and genome annotation with ChromHMM”. In: *Nature protocols* 12.12 (2017), pp. 2478–2492.

[42] Jeff Vierstra et al. “Global reference mapping of human transcription factor footprints”. In: *Nature* 583.7818 (2020), pp. 729–736.

[43] GTEx Consortium. “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. In: *Science* 369.6509 (2020), pp. 1318–1330.

[44] ENCODE Project Consortium et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), p. 57.

[45] Jennifer E Rood et al. “Impact of the Human Cell Atlas on medicine”. In: *Nature Medicine* (2022), pp. 1–11.

[46] Emelie Braun et al. “Comprehensive cell atlas of the first-trimester developing human brain”. In: *bioRxiv* (2022).

[47] Kimberly Siletti et al. “Transcriptomic diversity of cell types across the adult human brain”. In: *bioRxiv* (2022).

[48] Schahram Akbarian et al. “The psychencode project”. In: *Nature neuroscience* 18.12 (2015), pp. 1707–1712.

[49] Daifeng Wang et al. “Comprehensive functional genomic resource and integrative model for the human brain”. In: *Science* 362.6420 (2018), eaat8464.

[50] Wouter Van Rheenen et al. “Genetic correlations of polygenic disease traits: from theory to practice”. In: *Nature Reviews Genetics* 20.10 (2019), pp. 567–581.

[51] Adrian Cortes et al. “Identifying cross-disease components of genetic risk across hospital data in the UK Biobank”. In: *Nature genetics* 52.1 (2020), pp. 126–134.

[52] Brendan Bulik-Sullivan et al. “An atlas of genetic correlations across human diseases and traits”. In: *Nature genetics* 47.11 (2015), pp. 1236–1241.

[53] Joseph K Pickrell et al. “Detection and interpretation of shared genetic influences on 42 human traits”. In: *Nature genetics* 48.7 (2016), pp. 709–717.

[54] Brainstorm Consortium et al. “Analysis of shared heritability in common disorders of the brain”. In: *Science* 360.6395 (2018), eaap8757.

[55] Patrick Turley et al. “Multi-trait analysis of genome-wide association summary statistics using MTAG”. In: *Nature genetics* 50.2 (2018), pp. 229–237.

[56] Leland H Hartwell et al. “From molecular to modular cell biology”. In: *Nature* 402.6761 (1999), pp. C47–C52.

[57] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. “Network medicine: a network-based approach to human disease”. In: *Nature reviews genetics* 12.1 (2011), pp. 56–68.

[58] Karthik A Jagadeesh et al. “Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics”. In: *Nature Genetics* 54.10 (2022), pp. 1479–1492.

[59] Andrew D Grotzinger et al. “Genetic architecture of 11 major psychiatric disorders at biobehavioral, functional genomic and molecular genetic levels of analysis”. In: *Nature genetics* 54.5 (2022), pp. 548–559.

[60] Jenna Lee Ballard and Luke Jen O’Connor. “Shared components of heritability across genetically correlated traits”. In: *The American Journal of Human Genetics* 109.6 (2022), pp. 989–1006.

[61] Andrew D Grotzinger et al. “Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits”. In: *Nature human behaviour* 3.5 (2019), pp. 513–525.

[62] Travis T Mallard et al. “Multivariate GWAS of psychiatric disorders and their cardinal symptoms reveal two dimensions of cross-cutting genetic liabilities”. In: *Cell Genomics* 2.6 (2022), p. 100140.

[63] Andrew D Grotzinger et al. “Pervasive Downward Bias in Estimates of Liability-Scale Heritability in GWAS Meta-Analysis: A Simple Solution”. In: *Biological Psychiatry* (2022).

[64] Po-Ru Loh et al. “Mixed-model association for biobank-scale datasets”. In: *Nature genetics* 50.7 (2018), pp. 906–908.

[65] Wei Zhou et al. “Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies”. In: *Nature genetics* 50.9 (2018), pp. 1335–1341.

[66] Longda Jiang et al. “A resource-efficient tool for mixed model association analysis of large-scale data”. In: *Nature genetics* 51.12 (2019), pp. 1749–1755.

[67] Brendan K Bulik-Sullivan et al. “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies”. In: *Nature genetics* 47.3 (2015), pp. 291–295.

[68] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[69] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[70] Matthew Stephens and David J Balding. "Bayesian statistical methods for genetic association studies". In: *Nature Reviews Genetics* 10.10 (2009), pp. 681–690.

[71] Man Lu et al. "Targeting growth hormone function: strategies and therapeutic applications". In: *Signal transduction and targeted therapy* 4.1 (2019), pp. 1–11.

[72] Jamal S Rana et al. "Changes in mortality in top 10 causes of death from 2011 to 2018". In: *Journal of general internal medicine* 36.8 (2021), pp. 2517–2518.

[73] Eli A Stahl et al. "Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis". In: *Nature genetics* 44.5 (2012), pp. 483–489.

[74] "A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease". In: *Nature genetics* 47.10 (2015), pp. 1121–1130.

[75] Niek Verweij et al. "Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure". In: *Scientific reports* 7.1 (2017), pp. 1–9.

[76] Ben Elsworth et al. "The MRC IEU OpenGWAS data infrastructure". In: *BioRxiv* (2020).

[77] Tom G Richardson et al. "Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis". In: *PLoS medicine* 17.3 (2020), e1003062.

[78] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[79] Matthew J Maenner, Kelly A Shaw, Jon Baio, et al. "Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2016". In: *MMWR Surveillance Summaries* 69.4 (2020), p. 1.

[80] Michael E Talkowski and Stephan Sanders. *Diverse mutations in autism-related genes and their expression in the developing brain*. 2022.

[81] Jakob Grove et al. "Identification of common genetic risk variants for autism spectrum disorder". In: *Nature genetics* 51.3 (2019), pp. 431–444.

[82] Trent Gaugler et al. "Most genetic risk for autism resides with common variation". In: *Nature genetics* 46.8 (2014), pp. 881–885.

[83] Cato Romero et al. "Exploring the genetic overlap between twelve psychiatric disorders". In: *Nature Genetics* (2022), pp. 1–8.

[84] Elise B Robinson et al. "Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population". In: *Nature genetics* 48.5 (2016), pp. 552–555.

[85] Lucıa De Hoyos et al. "Structural models of genome-wide covariance identify multiple common dimensions in autism". In: *medRxiv* (2022).

[86] Manuel Mattheisen et al. “Identification of shared and differentiating genetic architecture for autism spectrum disorder, attention-deficit hyperactivity disorder and case subgroups”. In: *Nature Genetics* 54.10 (2022), pp. 1470–1478.

[87] Varun Warrier et al. “Social and non-social autism symptoms and trait domains are genetically dissociable”. In: *Communications biology* 2.1 (2019), pp. 1–13.

[88] Aysu Okbay et al. “Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals”. In: *Nature genetics* 54.4 (2022), pp. 437–449.

[89] Ditte Demontis et al. “Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder”. In: *Nature genetics* 51.1 (2019), pp. 63–75.

[90] Naomi R Wray et al. “Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression”. In: *Nature genetics* 50.5 (2018), pp. 668–681.

[91] Vassily Trubetskoy et al. “Mapping genomic loci implicates genes and synaptic biology in schizophrenia”. In: *Nature* 604.7906 (2022), pp. 502–508.

[92] Karsten Suhre et al. “Connecting genetic risk to disease end points through the human blood plasma proteome”. In: *Nature communications* 8.1 (2017), pp. 1–14.

[93] Lasse Folkersen et al. “Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease”. In: *PLoS genetics* 13.4 (2017), e1006706.

[94] Loic Yengo et al. “Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry”. In: *Human molecular genetics* 27.20 (2018), pp. 3641–3649.

[95] Jeanne E Savage et al. “Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence”. In: *Nature genetics* 50.7 (2018), pp. 912–919.

[96] Armin P Schoech et al. “Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection”. In: *Nature communications* 10.1 (2019), p. 790.

[97] Evangelos Evangelou et al. “Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits”. In: *Nature genetics* 50.10 (2018), pp. 1412–1425.

[98] Pim Van Der Harst and Niek Verweij. “Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease”. In: *Circulation research* 122.3 (2018), pp. 433–443.

[99] Angli Xue et al. “Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes”. In: *Nature communications* 9.1 (2018), p. 2941.

[100] Marleen HM De Moor et al. “Meta-analysis of genome-wide association studies for personality”. In: *Molecular psychiatry* 17.3 (2012), pp. 337–349.

[101] Aysu Okbay et al. “Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses”. In: *Nature genetics* 48.6 (2016), pp. 624–633.