2018

Using functional annotation to characterize genome-wide association results

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

USING FUNCTIONAL ANNOTATION TO CHARACTERIZE GENOME-WIDE ASSOCIATION RESULTS

by

VIRGINIA APPLEGATE FISHER

A.B., Harvard University, 2008
M.A., Boston University, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2018

Approved by

First Reader

Ching-Ti Liu, Ph.D.
Associate Professor of Biostatistics

Second Reader

L. Adrienne Cupples, Ph.D.
Professor of Biostatistics

Third Reader

Paola Sebastiani, Ph.D.
Professor of Biostatistics

DEDICATION

To Sonia and Liam

USING FUNCTIONAL ANNOTATION TO CHARACTERIZE GENOME-WIDE ASSOCIATION RESULTS

VIRGINIA APPLEGATE FISHER

Boston University Graduate School of Arts and Sciences, 2018

Major Professor: Ching-Ti Liu, Associate Professor of Biostatistics

ABSTRACT

Genome-wide association studies (GWAS) have successfully identified thousands of variants robustly associated with hundreds of complex traits, but the biological mechanisms driving these results remain elusive. Functional annotation, describing the roles of known genes and regulatory elements, provides additional information about associated variants. This dissertation explores the potential of these annotations to explain the biology behind observed GWAS results.

The first project develops a random-effects approach to genetic fine mapping of trait-associated loci. Functional annotation and estimates of the enrichment of genetic effects in each annotation category are integrated with linkage disequilibrium (LD) within each locus and GWAS summary statistics to prioritize variants with plausible functionality. Applications of this method to simulated and real data show good performance in a wider range of scenarios relative to previous approaches. The second project focuses on the estimation of enrichment by annotation categories. I derive the distribution of GWAS summary statistics as a function of annotations and LD structure and perform maximum likelihood estimation of enrichment coefficients in two simulated scenarios. The resulting estimates are less variable than previous methods, but the asymptotic theory of standard errors is often not applicable due to non-convexity of the likelihood function. In the third project, I investigate the problem of selecting an optimal set of tissue-specific annotations with greatest relevance to a trait of interest. I consider three selection criteria defined in terms of the mutual information between functional annotations and GWAS summary statistics. These algorithms correctly identify enriched categories in simulated data, but in the application to a GWAS of BMI the penalty for redundant features outweighs the modest relationships with the outcome yielding null selected feature sets, due to the weaker overall association and high similarity between tissue-specific regulatory features.

All three projects require little in the way of prior hypotheses regarding the mechanism of genetic effects. These data-driven approaches have the potential to illuminate unanticipated biological relationships, but are also limited by the high dimensionality of the data relative to the moderate strength of the signals under investigation. These approaches advance the set of tools available to researchers to draw biological insights from GWAS results.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BMI	Body mass index
CMI	Conditional mutual information
CNS	Central nervous system
GWAS	Genome-wide association study
LD	Linkage disequilibrium
MI	Mutual information
SNP	Single nucleotide polymorphism

CHAPTER 1. INTRODUCTION

1.1. Background

As the volume and types of available genomic and biological data increase, there is a need to develop methods for integrative analysis across data types to extract clinically meaningful conclusions. Genome-wide association studies (GWAS) have been extremely successful in identifying genetic loci associated with numerous heritable diseases and complex traits. However, these studies identify trait-associated loci that are hundreds of kilobases in size, rather than pinpointing specific causal variants. Furthermore, variants with moderate effect size may only reach statistical significance in large sample sizes from meta-analysis in consortia, which have access to association summary statistics but not individual-level genotype data. Thus, to take full advantage of the potential for biological discoveries from GWAS results, novel methods are needed to further analyze marginal summary statistics of genetic associations in order to prioritize causal variants and investigate the mechanisms by which they affect phenotypic outcomes.

Both the strength and limitations of GWAS methodology are due in large part to the presence of linkage disequilibrium (LD), or correlation between single nucleotide polymorphisms (SNPs) located in nearby genomic regions due to joint inheritance of contiguous sections of chromosome. This phenomenon allows GWAS to identify loci associated with the trait of interest even when the true causal variant is not observed, but also complicates interpretation of results. Recently developed methods in post-GWAS analysis have proposed various approaches to exploit knowledge of this correlation structure in order to study the genetic architecture of traits. Several fine mapping methods use LD information to elucidate local patterns of genetic effects. These include GCTA, which calculates conditional association from summary statistics [57, 56], CAVIARBF which compares Bayes factors for all possible models with one causal SNP [13], PAINTOR which uses an E-M algorithm to estimate the prior probability of causal sets of variants conditional on functional annotation [34, 32], and bfGWAS which performs Bayesian variable selection by partitioning loci of interest into distinct LD blocks [58]. On the genome-wide scale, LD score regression combines GWAS summary statistics with a reference LD panel to estimate systematic properties of the distribution of marginal associations, due to population structure or enriched heritability among certain functional categories [10, 20]. This method has been applied to GWAS of complex traits including schizophrenia, Crohn's disease, and BMI to identify relevant regulatory mechanisms.

Parallel developments in genomic annotation have illuminated the functional elements of human genetic variation, at the nucleotide scale. These resources describe the locations of regulatory elements, including promoters, enhancers, transcription factor binding sites, and histone modification in intergenic loci where large proportions of GWAS results have been found [22, 31, 15]. These measures collectively provide extensive information at the single nucleotide scale, regarding the biological function of a given variant. Tissue-specific annotation facilitates the investigation of even more targeted biological mechanisms underlying the observed GWAS signals [41, 36].

In this thesis, I present three projects that use genomic functional annotation to gain greater understanding of the biology responsible for GWAS results.

1.2. Previous Literature: Fine Mapping

Genetic fine mapping has been an area of active research in this decade. I will first review existing methods based on GWAS summary statistics, then methods that integrate functional annotation. The models described below apply to the most general case of a continuous phenotype *Y* generated by an additive genetic model, at the population level:

$$y = \boldsymbol{x\beta} + \epsilon$$

Here, $\boldsymbol{\beta}$ is a (column) vector of length *M* equal to the number of SNPs included in the analysis, with non-zero entries corresponding to causal SNPs. This depends upon an assumption that all causal SNPs in a locus of interest are observed through direct genotyping or imputation, so that the genetic effect is modeled directly rather than through LD tagging, the genotypes and phenotypes have been standardized to mean zero and unit variance, and the phenotype has already been adjusted for all relevant non-genetic covariates. The genotypes for a specific individual form a (row) vector $\boldsymbol{x}$ of length *M*. In the most basic GWAS study design, the genotypes and phenotypes of *N* unrelated individuals are observed, and represented as a matrix $\boldsymbol{X}$ of dimension $N \times M$, with the genotype vectors $\boldsymbol{x_i}$ for individuals $i = 1, \dots, N$ and the phenotypes as a column vector $\boldsymbol{y}$ of length *N*. This model assumes that the true causal SNPs are observed, either directly genotyped or imputed in the analysis sample. For each individual, $\epsilon_i \sim N(0, \sigma_R^2)$ represents the phenotype variation due to non-genetic factors, assumed independent and identically distributed, with variance σ_R^2. The methods based only on summary association statistics are often directly applicable to cases beyond a continuous phenotype (e.g. logistic regression for binary outcomes), though in some cases a scale transformation is required.

Many fine mapping studies use a simple conditional regression analysis to identify the number of distinct genetic signals in a trait-associated locus. Suppose a GWAS study has been performed in a sample of size *N*, resulting in marginal tests of association between phenotype vector $\boldsymbol{y}$ and genotype vector X_j (i.e. a column of ***X)*** of length *N* containing all sampled genotypes at SNP *j*:

$$\boldsymbol{y} = X_j\beta_j + \boldsymbol{\epsilon}$$

This univariate test of association is performed at a large number of SNPs ($j = 1, \ldots, M$). If multiple SNPs in a given locus attain genome-wide significance, a conditional analysis tests the null hypothesis that all observed associations are due to LD with the SNP with lowest p-value in the region. If X_{top} is the vector of genotypes across individuals at the most significant SNP, then the conditional regression is

$$\boldsymbol{y} = \widetilde{\beta_j} X_j + \widetilde{\beta_{top}} X_{top} + \boldsymbol{\epsilon_{cond}}$$

If any coefficients $\widetilde{\beta_j}$ are significantly different from zero there is evidence of multiple association signals in the locus, and the conditioning is iteratively repeated. Genome-wide complex trait analysis (GCTA [56]) presents a method to estimate these conditional tests of association from marginal effect estimates and an LD matrix approximated from an external reference panel. In particular, the conditional effect estimate

$$\widetilde{\beta_2}|\widetilde{\beta_1} = (X_2^T X_2)^{-1} X_2^T \boldsymbol{y} - (X_2^T X_2)^{-1} X_2^T X_1 (X_1^T X_1)^{-1} X_1^T \boldsymbol{y}$$

is calculated by the approximations $X_j^T X_{j'} \approx NCov(X_j, X_{j'})$ and $X_j^T X_j \approx NVar\left(X_j\right) = N2p_j(1 - p_j)$ for the j^{th} and j'^{th} genotype vectors, both of which may be estimated from a population-matched reference panel.

The primary shortcoming of the conditional analysis approach is its reliance on the marginal or conditional p-values as statistics for the evidence of causality at a given SNP. In cases of measurement error due to imputation, sampling variation, or complex patterns of LD that lead to cancelling of effects with opposite direction, the variant with lowest p-value may be in LD with a causal SNP but not itself functional. For this reason, conditional analysis may be used to count the number of distinct genetic signals in the locus, but is not useful for prioritizing variants for follow-up studies of causality.

A series of methods approach the problem of quantifying the evidence for causality in favor of each SNP in a region by modeling the joint distribution of the vector of test statistics at all SNPs under various causal configurations. The CAusal Variants Identification in Associated Regions method (CAVIAR [30]) defines the non-centrality parameter $\lambda_c = \frac{\beta_c}{\sigma}\sqrt{N}$, where σ^2 is the residual variance from the GWAS regression, giving the asymptotic distribution of the Wald test statistic at a causal SNP *c* in the absence of LD:

$$Z_c = \frac{\widehat{\beta_c}}{SE(\widehat{\beta_c})} \sim N(\lambda_c, 1)$$

At a non-causal SNP *j* tagging a causal variant, the least squares effect estimate is weighted by the Pearson correlation between genotypes i and j:

$$\widehat{\beta_J} = \left(X_j^T X_j\right)^{-\mathbf{1}} X_j^T \boldsymbol{y} = \frac{1}{N} X_j^T (\boldsymbol{X\beta} + \boldsymbol{\epsilon}) = \sum_{\boldsymbol{j'=1}}^{\boldsymbol{M}} X_j^T X_{j'} \beta_{j'} + \boldsymbol{\epsilon}' = \sum_{\boldsymbol{j'=1}}^{\boldsymbol{M}} r_{jj'} \beta_{j'} + \boldsymbol{\epsilon}' \quad (\mathbf{1.1})$$

where $\widehat{r_{JJ'}} = \frac{1}{N} X_j^T X_{j'}$ and $\boldsymbol{\epsilon}' = \frac{1}{N} X_j^T \boldsymbol{\epsilon} \sim N(0, \frac{1}{N}\sigma_R^2)$. By the same argument, the Wald statistic distribution is

$$Z_j \sim N(\sum_{\boldsymbol{j'=1}}^{\boldsymbol{M}} \widehat{r_{jj'}}\lambda_{j'}, 1)$$

The authors define a causal indicator vector $\boldsymbol{c}$ of length *M* with value $c_j = 1$ for every causal SNP *j* in the locus and 0 otherwise. Then, the joint distribution of the vector of Wald test statistics is

$$\boldsymbol{Z} \sim MVN(\Sigma(\boldsymbol{\lambda} \circ \boldsymbol{c}), \Sigma) \qquad \qquad (\mathbf{1.2})$$

where $\Sigma = \frac{1}{N}\boldsymbol{X^T X}$ is the correlation matrix representing LD structure within the locus, and $\circ$ indicates element-wise multiplication. The LD correlation matrix Σ appears in both the mean and variance of the distribution of test statistics as a consequence of the so-called tagging of true genetic effects in the association tests of other SNPs in LD. The CAVIAR method uses maximum likelihood estimation based on the vector of test statistics $\boldsymbol{Z}$ observed from a GWAS, and Σ estimated from a population-matched reference panel to identify the set of causal SNPs $\boldsymbol{c}$ which yields greatest likelihood of the observed data.

In practice, enumeration over all 2^M causal sets is computationally prohibitive, so an upper limit *l* is placed on the number of causal SNPs considered simultaneously. The default value of *l*=6 is chosen as a compromise between model flexibility and expediency.

The CAVIARBF method adapts this model to estimate Bayes factors, quantifying the evidence of each causal set relative to the null model of no causal variants in the locus [13]. By placing independent normal priors on the distribution of each SNP effect, they derive a closed form estimator of the Bayes factor for a specific causal SNP model represented by the causal SNP vector $\boldsymbol{c}$ versus the null model with no causal SNPs, given a prior distribution of the SNP effects $\beta \sim N(0, \frac{\sigma_a^2}{\tau} I_M)$, where σ_a^2 depends upon the total trait heritability and I_M is the $M \times M$ identity matrix. These Bayes factors are used to calculate the posterior probability of each model $\boldsymbol{c}$ in the set C_l of models with at most l causal SNPs. From these posterior probabilities of causal SNP sets, the probability of including a given SNP as causal is then given by the posterior inclusion probability (PIP) which may be used to rank the evidence for each variant in the locus:

$$p\left(c_j = 1 \middle| Z, \Sigma, \sigma_a^2\right) = \sum_{Models\ \boldsymbol{c}: c_j = 1} p(\boldsymbol{c}|Z, \Sigma, \sigma_a^2)$$

The Bayes factor is approximately equivalent to the likelihood ratio statistic used in CAVIAR, where the former uses the complete data likelihood, and the latter evaluates the likelihood at its maximum value. Both of these methods are limited by the constraint on the number of causal variants considered in each model [30]. The authors present simulations showing that the results of estimation of SNP-specific causality over the constrained model space and full space of 2^M models are very similar when there are 1 or 2 true causal SNPs in the locus. However, neither paper investigates the validity of the assumption that trait-associated loci are expected to contain relatively few causal SNPs. In reality, the number of causal SNPs will vary by locus and by trait, but simulation studies could be designed to assess the sensitivity of model performance under a range of generative genetic models.

1.3.Functional Annotation

Additional, independent information regarding the biological function of SNPs within a locus may be used to aid fine mapping analysis and improve interpretability of results. Functional annotation provides the positions of known genomic elements, comprising both genes and regulatory regions such as histone modifications, transcription factor binding sites, DNAse hypersensitive sites and regions of enriched DNA methylation. Many of these regulatory elements show differential patterns of activation across tissue types, leading to the definition of tissue-specific functional annotation. Additionally, conservation of a given sequence across species gives evidence of its importance for survival, even if the functional mechanism is unknown. The ENCODE Project [22] and ROADMAP Epigenomics Consortium [15] have published extensive databases of functional annotation, and several methods have been developed to integrate these data into genetic fine mapping studies.

JAM (Joint Analysis of Marginal summary statistics, [42]) is based on a model similar to CAVIARBF, except that it models the residual variance σ_R^2 as an unknown parameter in the model, rather than conditioning on the median observed value, as CAVIARBF does. To facilitate computation, JAM uses a Cholesky decomposition of the LD matrix in terms of an upper triangular $M \times M$ matrix L such that

$$X^T X = L^T L$$

The existence of this decomposition depends upon $X^T X$ being positive definite, which cannot be the case when the locus under consideration contains more SNPs than the number of subjects in the reference panel. Thus, the LD matrix is rank-deficient. For this reason, JAM is only applicable to fine mapping studies after initial screening criteria reduce the number of SNPs under consideration. Functional annotation may be incorporated by means of SNP-specific parameters for the prior distribution of the proportion of causal SNPs.

The GenoWAP (Genome-Wide Association Prioritization) method calculates the posterior probability that each SNP is causal using Bayes' Theorem with prior probabilities of causality pre-computed as a function of annotation from NHGRI GWAS catalog training data [40]. This method does not model local LD structure, and assumes that the functional genetic architecture is consistent for all traits.

The Functional GWAS method (fGWAS [45]) is based on a hierarchical model where first the probability of a true association at each locus genome-wide is estimated, and then, the probability that each SNP is causal conditional on being located in an associated locus is estimated. The GWAS effect estimates and standard errors are used to calculate a Bayes factor comparing the model in which that SNP is causal to the null model in which the genetic effect is attributable to other SNPs. Prior probabilities of causality for each locus and SNP are defined in terms of locus-level and SNP-level functional annotation by means of penalized maximum likelihood estimation. This hierarchical model depends upon a partition of the genome into independent loci of equal size, and makes the assumption that each locus contains at most one causal SNP. These assumptions are generally not justifiable.

The Bayesian Functional GWAS method (bfGWAS [58]) method attempts to correct these shortcomings by including all SNPs within a locus in a multiple regression model for the phenotype, and modeling both the probabilities of causality $\boldsymbol{\pi}$, and the effect size distribution of causal SNPs within each annotation category.

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

The effect of a SNP j in functional annotation category k is modeled by a slab and spike prior distribution with annotation-level hyperparameters $\pi_j = \pi_k$ and $\sigma_j^2 = \sigma_k^2$, where δ_0 is a probability point mass at zero:

$$\beta_j \sim \pi_i N(0, \tau^{-1}\sigma_i^2) + (1 - \pi_i)\delta_0$$

A binary vector indicating causality at each SNP then has elements $c_i \sim Bernoulli(\pi_k)$ and the joint distribution of effect sizes at causal SNPs with $c_i = 1$ is given by the diagonal covariance matrix $\boldsymbol{V}_c = diag(\sigma_1^2, \ldots, \sigma_{|c|}^2)$ where $|\boldsymbol{c}|$ is the number of causal SNPs in model $\boldsymbol{c}$:

$$\boldsymbol{\beta}_c \sim MVN_{|c|}(0, \tau^{-1}\boldsymbol{V}_c)$$

Estimation of this model uses the EM-MCMC algorithm based on approximately independent genome-blocks containing 5,000-10,000 variants. These genome-blocks must be selected so that significantly trait-associated variants in LD with each other are in the same block. Genome-block selection, based on marginal association evidence, genomic distance and LD, is the first practical step in the implementation of this approach. The bfGWAS algorithm performs MCMC estimation within each block with fixed category-level parameter values $(\boldsymbol{\pi}, \boldsymbol{\sigma}^2)$ to obtain posterior estimates of $(\boldsymbol{\beta}, E[\boldsymbol{c}])$, then uses these posterior estimates from all loci genome-wide to maximize the posterior likelihoods of

$(\boldsymbol{\pi}, \boldsymbol{\sigma}^2)$ and update their estimates. The parameters $(\boldsymbol{\beta}, \boldsymbol{c})$ are the primary quantities of interest, representing the causal SNPs in each genome-block and their effect sizes.

This method addresses the primary weakness of fGWAS by removing the assumption on the number of causal SNPs per locus. However, the likelihood function used to summarize across loci in the M-step of the estimation algorithm requires the functional annotations to be defined so that each SNP is included in exactly one annotation category. It also uses individual-level genotype and phenotype data as input, though the authors write that they are developing an extension based on GWAS summary statistics.

Another method builds on the CAVIAR model described above. PAINTOR (Probabilistic Annotation INTegratOR, [34, 33]) defines prior probabilities for each causal SNP vector by independent logistic models for causality at each SNP. Given *K* binary annotation categories, the annotation at SNP j is represented as a vector A_j of length *K*+1 where $a_{kk} = 1$ if SNP j is in category k $(k = 1, \dots, K)$ and 0 otherwise, and $a_{j0} = 1$ is a baseline annotation representing the intercept of the logistic model

$$logit\left(P(c_j = 1)\right) = \gamma^T A_j$$

$$P(\boldsymbol{c}; \gamma, A) = \prod_{j=1}^{M} \left(\frac{1}{1 + \exp(\gamma^T A_j)}\right)^{c_j} \left(\frac{1}{1 + \exp(-\gamma^T A_j)}\right)^{1-c_j}$$

The coefficients γ represent the enrichment (or depletion) of causal SNPs within each annotation category. These are estimated with an E-M algorithm based on the complete data likelihood across all loci in the fine-mapping analysis:

$$L(Z; \gamma, \boldsymbol{\lambda}, A) = \prod_{loci} \sum_{c \in C} P(Z | \boldsymbol{c}, \boldsymbol{\lambda}) P(\boldsymbol{c}; \gamma, A)$$

Here, $P(Z|\boldsymbol{c},\boldsymbol{\lambda})$ is the probability density function of the observed test statistics, conditional on the causal SNP vector $\boldsymbol{c}$ and the vector of non-centrality parameters $\boldsymbol{\lambda}$ as defined in equation (1.2). To make the parameters γ identifiable, the non-centrality parameters $\boldsymbol{\lambda}$ are fixed at the observed values of the GWAS test statistics, with small magnitudes rounded up to 3.7 times the sign of the observed value, to ensure the inclusion of all SNPs in the locus for consideration.

Because PAINTOR iterates over causal sets in a similar manner to CAVIAR, the original method implemented shares the same limit on the permitted number of causal SNPs. In version 3 [33] this requirement is relaxed by the introduction of an Importance Sampling approach which concentrates computational resources on causal models with high probability, while drawing fewer samples from the likelihood of models with low probability, and therefore smaller contribution to the posterior.

The proposed E-M algorithm is only practicable when the dimension of γ is low, allowing the incorporation of no more than 5 or 10 annotation categories. The authors suggest an iterative procedure for selection of annotation categories, based on likelihood ratio tests of nested models adding each available annotation to those already selected, requiring repeated fitting of computationally demanding estimation algorithms. Version 3 provides an option for users to supply externally-derived estimates of γ.

1.4.Partitioning Heritability by Functional Annotation Enrichment

Incorporation of functional annotation data into genetic fine mapping studies requires estimation of enrichment coefficients which quantify the extent to which each annotation category contributes to the phenotype of interest. Two recent methods aim to estimate these coefficients from genome-wide summary statistics.

Finucane et al [20] build on the LD score regression framework, initially proposed to distinguish confounding due to population structure from polygenicity where large numbers of SNPs have weak effects on the phenotype. This approach treats the observed GWAS χ^2 statistics as the dependent variable in a regression model, and statistics calculated from LD as independent variables.

For the purpose of annotation enrichment estimation, the authors model the SNP effects $\boldsymbol{\beta}$ as a random vector with mean 0 and diagonal covariance matrix, so the random effects are pairwise independent but not necessarily identically distributed. The variance of each SNP effect depends on enrichment coefficients of the annotation categories to which that SNP belongs:

$$Var(\beta_j) = \sum_{k:a_{jk}=1} \tau_k$$

From this and equation (1.1), and using the fact that the phenotype vector has been standardize, and each individual SNP effect is expected to be much smaller than the total trait variance, so that $SE(\widehat{\beta_J}) \approx \frac{1}{\sqrt{N}}$ they derive the expected value of $\chi_j^2 \approx N\widehat{\beta_j^2}$:

$$E[\chi_j^2] = N\sum_{k=1}^{K} \tau_k \sum_{j':a_{j'k}=1} \widehat{r_{jj'}^2} + \sigma_R^2$$

$$= N\sum_{k=1}^{K} \tau_k \sum_{j':a_{j'k}=1} r_{jj'}^2 + \sum_{j'=1}^{M} \sum_{k:a_{j'k}=1} \tau_k + \sigma_R^2 = N\sum_{k=1}^{K} \tau_k \ell(j,k) + 1$$

The second equality uses the fact that $E\left[\widehat{r_{jj'}^2}\right] = r_{jj'}^2 + \frac{1}{N}$ and the third uses the fact that the phenotype is standardized so that $\sum_{j'=1}^{M} \sum_{k:a_{j'k}=1} \tau_k + \sigma_R^2 = \sum_{j'=1}^{M} Var(\beta_{j'}) + \sigma_R^2 = 1$. The category-specific stratified LD scores, defined as $\ell(j,k) = \sum_{j':a_{j'k}=1} r_{jj'}^2$, where $a_{j'k} = 1$ if SNP j' is in category k $(k = 1, \ldots, K)$ and 0 otherwise, quantify the extent to which SNP j tags variation in category k (stratified refers to the cumulative LD with the set of SNPs in a given annotation category). These scores may be calculated from a population-matched reference panel external to the GWAS analysis sample.

The LD score regression software estimates the coefficients τ_k by means of a weighted linear regression model. The weights are incorporated to adjust for the non-independence and heteroscedasticity of the χ_j^2 statistics. This linear regression may give negative estimates of $\widehat{\tau_k}$, despite their definition as variance components which must be between 0 and $Var(Y) = 1$. The authors recommend truncating these estimates to the permissible range when mean squared error is more important than unbiasedness (as when reporting the results from a single analysis), or using the original estimates when unbiasedness is a desired property (as in a simulation study).

MQS is an alternative Method of Moments approach for estimation of partitioning genetic variance components by annotation categories using GWAS summary statistics [60]. This method assumes the annotation categories are disjoint, so that each SNP j is included in at most one annotation. Then the genetic effects on the phenotype may be summarized by annotation category as $g_k = \sum_{j:a_{jk}=1} X_j \beta_j$

$$\boldsymbol{y} = \sum_{j=1}^{M} X_j \beta_j + \boldsymbol{\epsilon} = \sum_{k=1}^{K} g_k + \boldsymbol{\epsilon}$$

For each category k, g_k is a vector of length *n* modeled by the multivariate normal distribution $g_k \sim MVN(0, \sigma_k^2 R_k)$ where $R_k = X_k X_k^T / m_k$ is the genetic relationship matrix estimated from only the m_k SNPs in that category. This model parallels the standard formulation of variance components estimation in mixed effect models. Restricted maximum likelihood (REML) estimation, which is commonly used in this context, is statistically efficient in terms of the variance of the estimator, but requires individual level genotype and phenotype data, and is computationally demanding in large samples. The method of moments approach to variance component estimation depends upon a system of equations

$$E\left(y^T A_j y\right) = \sum_{i=1}^{K} tr\left(A_j K_i\right) \sigma_i^2 + tr\left(A_j\right) \sigma_E^2$$

Zhou shows that the matrices A_j may be derived from a weighting function on the observed SNPs, and that different choices of SNP weights give estimators that are equivalent to LD score regression and IBS Haseman-Elston regression as special cases.[12, 60]

Comparison between partitioned heritability results from MQS and LDSC requires a reparametrization of variance components. For a given category *k,* the parameter σ_k^2 in the MQS model represents the total trait variance due to all SNPs in that category, while the LDSC parameter τ_k represents the average per-SNP heritability. These parameters are related by the function $\sigma_k^2 = m_k \tau_k$.

Both the MQS and LDSC estimators are theoretically unbiased for the true variance components. However, both estimators can also yield negative estimates for the variance components, which are outside the parameter space. Additionally, the linear regression model with GWAS summary statistics as a function of stratified LD scores is misspecified, thereby reducing the efficiency of the estimator. The method of moments approach implemented in MQS makes an intentional trade-off of computational efficiency at the expense of statistical efficiency. Its inability to analyze overlapping annotation categories is a substantial impediment for applications to real-world data, where SNPs likely fall into more than one annotation category.

1.5.Outline of this Dissertation

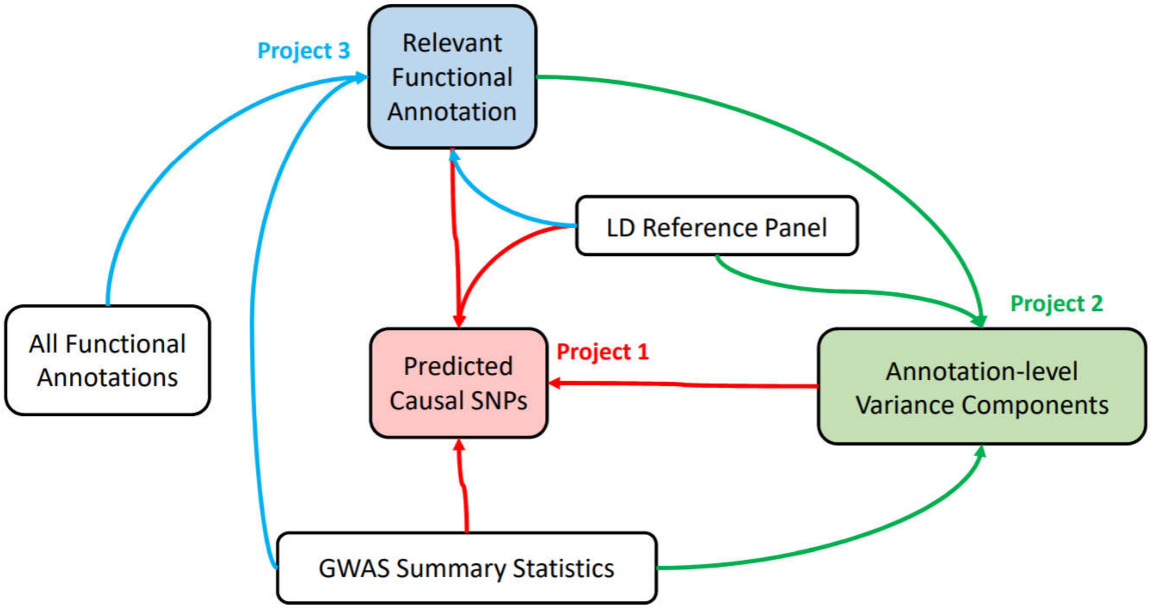
Project 3
Relevant Functional Annotation
LD Reference Panel
Project 2
All Functional Annotations
Predicted Causal SNPs
Project 1
Annotation-level Variance Components
GWAS Summary Statistics

Figure 1.1. Diagram representing inputs and outputs of the three projects in this dissertation.

In this thesis, I present three projects that integrate GWAS results with genomic functional annotation to improve upon existing methods for fine mapping, partitioning trait heritability, and selection of tissue-specific annotations relevant to a given trait. The common goal for these projects is to provide tools for investigators seeking to glean clinically relevant biological insights from genome-wide association results.

Chapter 2 develops a fine mapping method to prioritize potentially causal variants in a locus of interest based on GWAS association statistics, patterns of linkage disequilibrium (LD) among all variants in the region, and functional annotation. Unlike the CAVIAR and PAINTOR approaches, this method does not place an upper bound on the number of causal SNPs per locus. In contrast to the previous approaches described above, I separate the estimation of annotation effects from that of SNP effects in the loci of interest. The annotation effects are estimated from genome-wide summary statistics, taking advantage of low-level systematic enrichment of association signals throughout the genome including those outside of regions reaching genome-wide significance. In the implementation I present in Chapter 2, these estimates are obtained by stratified LD score regression [20], though the fine mapping model can accept alternative estimates of these quantities. The annotation enrichment coefficients are then used to define variant-specific distributions for the effect size in a multiple regression model accounting for all variants in the locus. This approach is analogous to a penalized regression model in which larger penalties are applied to variants in annotation categories that are not enriched for trait heritability. I apply the proposed method to both simulated data and a published GWAS of body mass index (BMI) [37].

In Chapter 3, I focus on the estimation of functional enrichment coefficients from genome-wide summary data. Previous approaches to this problem make simplifying assumptions that result in estimators with high variance, limiting their ability to identify weak levels of enriched association. By modeling the noncentrality parameter of the GWAS χ^2 test statistics as a function of annotation-specific LD scores and enrichment coefficients, I am able to estimate the proportion of heritability attributable to each annotation category with maximum likelihood estimation. I explore reparametrization of the likelihood in order to ensure non-negative estimates of the variance components. However, this induces symmetry between positive and negative values in the new parameter space, so that the transformed likelihood function is non-convex and difficult to optimize

Chapter 4 explores the application of mutual information based feature selection algorithms to the problem of identifying functional annotation categories that are relevant to a given trait of interest. Feature selection refers to methods for the identification of optimal sets of predictors for a given outcome, when the available predictors (also called features) are highly correlated with each other, or simply too numerous to model simultaneously. In Chapter 3 I demonstrated that estimation of enriched heritability by functional annotation is improved by including the correct set of annotations in the model. Previous approaches have dealt with this problem by selecting annotation categories that show significant enrichment when modeled singly, as in PAINTOR, or by considering all available annotations simultaneously, as in LDSC. To my knowledge, there have not been any previous applications of statistical feature selection criteria to functional annotation for

GWAS statistics. The approach presented in Chapter 4 is based upon mutual information, a non-parametric measure of dependence between random variables which does not require distributional assumptions, in contrast to the highly model-specific approach of Chapter 3. I apply three feature selection criteria to a data set consisting of tissue-specific histone marks, in hopes of identifying both relevant tissues and specific regulatory mechanisms involved in the genetics of BMI.

These three projects demonstrate several themes that appear in various ways throughout the field of statistical genetics. They each work with GWAS summary statistics rather than requiring individual-level genotype and phenotype data, to facilitate implementation in consortium-based studies and maximize the available sample sizes. Generally, I take an exploratory perspective, assuming that the underlying trait biology is largely unknown, rather than incorporating targeted biological hypotheses, as for example in a candidate gene study. Finally, due to the scale of these data sets, I recognize computational efficiency as critical to the practical applicability of any novel statistical method. Chapter 5 contains discussion of these themes and directions for future research.

The methods I develop in this thesis address the challenges that investigators face in using findings from GWAS to understand the genetic etiology of complex traits, and translate this understanding into advancements in the prevention and treatment of disease, and ultimately improve public health.

CHAPTER 2. GENETIC FINE MAPPING WITH FUNCTIONAL ANNOTATION: A RANDOM EFFECTS APPROACH

2.1. Introduction

Large-scale genotyping studies have great potential to enhance our understanding of the genetic etiology of human complex traits. Genome-wide association studies (GWAS), in which a large number of single nucleotide polymorphisms (SNPs) are individually tested for association with an outcome of interest, have been the primary study design for such investigations [53]. However, the findings from such analyses typically suggest genomic loci with hundreds of kilobases in size, often containing hundreds of SNPs that exceed the genome-wide significance threshold.

Patterns of linkage disequilibrium (LD) within such loci present a challenge for researchers seeking to identify the true causal variants. Functional validation in experimental organisms is necessary to confirm findings from epidemiological studies, but these experiments are costly and time-consuming. Statistical fine mapping methods can be useful to prioritize variants for follow-up. In order to take advantage of large sample sizes available in meta-analysis of GWAS within consortia of studies, there is demand for methods that need only summary association statistics rather than individual-level data [43].

Functional annotation, which describes both protein-coding genes and epigenetic regulatory elements such as promoters, enhancers, and transcription factor binding sites, provides valuable information about the potential biological relevance of SNPs within a trait-associated locus.

Several methods to integrate functional annotation into fine mapping of loci identified by GWAS were described in Chapter 1. These methods attempt to jointly model the influence of functional annotation categories and the individual-level SNP effects, requiring computationally demanding iterative algorithms or resampling procedures, and limiting the number of functional categories that can be modeled simultaneously. Of these previously developed methods, PAINTOR is the only one that uses summary statistics as input and incorporates both functional annotation and LD structure into the fine mapping analysis. Earlier versions of PAINTOR placed an upper bound on the number of causal SNPs per locus, while version 3 uses Importance Sampling to explore the space of potential causal sets [33].

In this chapter, I propose AnnoRE, a random effects model for genetic fine mapping that integrates the genome-wide heritability attributable to each functional annotation category to prioritize SNPs in each locus identified by GWAS. Specifically, the effect distribution of each SNP is defined by annotation-specific functional variance components. For the implementation presented here, I estimated annotation-specific functional variance components by the LD score regression method of partitioning heritability [20], but other estimators may be substituted. This method uses GWAS summary statistics to estimate enrichment of association signals across a large number of functional categories. The sum of these variance components yields high-resolution SNP effect distributions, and by conditioning on the variance components we obtain an efficient closed form solution within the fine mapping locus. In the Methods section, we present the details of the proposed model. Simulation studies compare the performance of AnnoRE to previously published fine-mapping methods. Finally, we apply AnnoRE to a large-scale GWAS meta-analysis of body mass index (BMI) [37]. And we conclude with a discussion of the strengths and potential limitations of this approach.

2.2. Methods

The parameters of the proposed approach, the AnnoRE model, may be estimated either directly from subject-level genotype and phenotype data, or from marginal genetic association statistics in combination with allele frequency and LD information from an ancestry-matched reference panel. For computational efficiency we restrict fine mapping analysis to variants exceeding a nominal significance threshold of p<0.05 in the preliminary GWAS analysis.

2.2.1. Random Effects Model

In each fine mapping locus, with *M* SNPs genotyped or imputed in *N* subjects, we assume an additive genetic model in a multiple regression equation with all candidate SNPs included as predictors

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

Given an unrelated sample, the subject-level residuals $\boldsymbol{\epsilon}$ follow a multivariate normal distribution, $\boldsymbol{\epsilon} \sim MVN(0, \sigma_R^2 I_N)$, where σ_R^2 is the residual variation in the phenotype not attributable to additive genetic effects.

We assume that the $N \times M$ genotype matrix $\boldsymbol{X}$ is standardized so that each column has mean zero and unit variance. Annotation describing C functional categories is summarized in a $(M \times C)$ matrix $\boldsymbol{A}$ of binary indicators $a_{jk} = 1$ if SNP j is included in functional category k and zero otherwise.

We define a random effect for each standardized SNP by a Gaussian distribution with mean zero and variance $Var(\beta_j) = k\ \sum_k a_{jk}\tau_d$ where the sum is taken over all annotation categories containing SNP *j.* The annotation-specific variance component τ_k represents the coefficient of expected per-SNP heritability in category *k.* These variance components may be estimated by the LD score regression method for partitioning heritability [20], with negative estimates truncated at zero. The factor h_{loc} is defined as the expected heritability per SNP within each fine mapping locus, relative to the genome-wide heritability per SNP. Inclusion of this factor adjusts the random effect variances for the strength of the observed genetic association within the locus.

$$h_{loc} = \frac{(heritability\ in\ locus)/(\#\ of\ SNPs\ in\ locus)}{(total\ trait\ heritability)/(\#\ of\ SNPs\ in\ GWAS)}$$

The vector of SNP effects $\boldsymbol{\beta} \sim N(0, \boldsymbol{H})$ is modeled as random effects with independent Gaussian distributions, with $\boldsymbol{H} = diag(h_{loc}\boldsymbol{A\tau})$. Then, the best linear unbiased predictor (BLUP), obtained by maximizing the joint distribution of *Y* and β conditional on *H* and σ_R^2 is given by

$$\widetilde{\boldsymbol{\beta}} = \left[\frac{\boldsymbol{X^T X}}{\sigma_R^2} + \boldsymbol{H^{-1}}\right]^{-1} \frac{\boldsymbol{X^T y}}{\sigma_R^2}$$

and its variance-covariance matrix, conditionally on the parameters $\boldsymbol{\tau}$

$$Cov(\widetilde{\boldsymbol{\beta}}|\boldsymbol{H}) = \left[\frac{\boldsymbol{X^T X}}{\sigma_R^2} + \boldsymbol{H^{-1}}\right]^{-1}$$

This is a well-known result in the theory of random effects models and coincides with the estimator implemented in SAS PROC MIXED [28].

Note that, under this model, genotype vector standardization encodes the assumption that less frequent variants will have larger effect sizes, as we would expect due to negative selection [54].

From the diagonal elements of this covariance matrix we may construct a Wald test statistic, conditional on $\boldsymbol{\tau}$, for the effect of each SNP in the locus: $\tilde{\beta}_j / SE(\tilde{\beta}_j)$.

2.2.2. Estimation from Summary Statistics

The AnnoRE model may also be estimated from GWAS summary statistics and local LD information from a population-matched reference panel as an approximation of the genetic correlation structure. Suppose that a given fine mapping locus contains *M* SNPs, with allelic effect estimates $\widehat{\beta_j}$, standard errors $SE(\widehat{\beta_j})$ and allele frequencies $\widehat{p_j}$ available from a study with N_j unrelated subjects contributing to analysis at SNP *j*. Additionally, suppose that the LD matrix $\widehat{\boldsymbol{\Sigma}}$ of pairwise Pearson correlations between all *M* SNPs is available from a reference panel of the same ancestral population as the GWAS sample.

These summary statistics are calculated in terms of the SNP genotypes X_j, the length-*N* column vectors of the matrix $\boldsymbol{X}$, with a simple linear regression model at each SNP j:

$$\boldsymbol{y} = X_j \beta_j + \boldsymbol{\epsilon_j}$$

By assumption, the subjects are unrelated, so $\boldsymbol{\epsilon_j} \sim N(0, \sigma_j^2 I_{N_j})$.

If all genotypes were independent, i.e. if there was no tagging due to LD and the GWAS effect estimates at a given SNP represented only the causal effect of that SNP and independent residual error, the asymptotic distributions of the least squares estimators $\widehat{\beta_J} = \left(X_j^T X_j\right)^{-1} X_j \boldsymbol{y}$ would give rise to the GWAS test statistics:

$$Z_j = \frac{\widehat{\beta_J}}{SE(\widehat{\beta_J})} \sim N\left(\beta_j \sqrt{\frac{N_j}{\sigma_j^2}}, 1\right)$$

Following the notation in Hormozidari *et al* [30], define the non-centrality parameter $\lambda_j = \beta_j \sqrt{\frac{N_j}{\sigma_j^2}}$, which is related to the statistical power to detect a significant association between genotype *j* and the trait of interest.

Allowing for genotype correlation within the locus gives the LD-induced non-centrality parameter $\Lambda_j = \sum_{j'} r_{jj'} \lambda_{j'}$ where $r_{jk} = \frac{1}{\sqrt{N_j N_{j'}}} X_j X_{j'}^T$ is the LD (Pearson correlation) between SNPs *j* and *k*.

Then the multivariate distribution of the vector of Wald statistics across the locus is

$$\boldsymbol{Z}|\boldsymbol{\beta} \sim MVN(\boldsymbol{\Lambda}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is the LD matrix, which may be approximated from a reference panel. Define $\boldsymbol{S} = diag\left(\sqrt{\frac{N_j}{\sigma_j^2}}\right)\boldsymbol{\Sigma}$, so that $\boldsymbol{\Lambda} = \boldsymbol{S\beta}$, yielding a multiple regression model

$$\boldsymbol{Z} = \boldsymbol{S\beta} + \boldsymbol{\eta}$$

where $\boldsymbol{\eta} \sim N(0, \boldsymbol{\Sigma})$. Using the same random effect variance matrix $\boldsymbol{H}$ defined in Section 2.2.1, the BLUP for this regression is

$$\widetilde{\boldsymbol{\beta}} = \ [\boldsymbol{S^T}\boldsymbol{\Sigma^{-1}}\boldsymbol{S} + \boldsymbol{H^{-1}}]^{-1}\boldsymbol{S^T}\boldsymbol{\Sigma^{-1}}\boldsymbol{Z}$$

Variances of the individual SNP effect estimates are given by the diagonal elements of the matrix $[\boldsymbol{S^T}\boldsymbol{\Sigma^{-1}}\boldsymbol{S} + \boldsymbol{H^{-1}}]^{-1}$, which may be used to define Wald statistics for the estimators. These test statistics $W_j = \widetilde{\beta_j}/SE(\widetilde{\beta_j})$ follow an asymptotic standard normal distribution under the null hypothesis $\beta_j = 0$, conditional on the random effect variances $\hat{\boldsymbol{\tau}}$ [47, 28]. The details of this derivation are given in Appendix A.

2.3. Simulation Study

2.3.1. Design of Simulations

I simulated GWAS test results in a 1MB region from 28,000,000 to 29,000,000 base pairs on chromosome 21. Using HAPGEN2 [50], I constructed synthetic samples of size *N*=10,000 from haplotypes of the 379 individuals of European ancestry included in the Phase 1 of the 1000 Genomes Project [14]. This reference panel contained *M*=2,159 SNPs in the locus of interest. Annotation-specific variance components $\hat{\boldsymbol{\tau}}$ were estimated from the GIANT Consortium GWAS of BMI [37] for 52 regulatory annotation categories from the ENCODE project [22], to define the random effect distributions for all SNPs in the locus. Negative estimates $\widehat{\tau_d}$ were truncated at zero, effectively removing 28 annotation categories from the model, representing no evidence of enriched heritability in these categories.

Each simulation scenario had one causal SNP, and these 20 causal SNPs were used to define a range of simulation scenarios. For each simulation scenario, or model with one of the 20 SNPs as causal, I created 1000 phenotype replicates under a genetic model, fixing

SNP-specific heritability at $h^2 = 0.001$ and calculating the genetic effect size to account for different allele frequencies between the scenarios.

Causal SNPs were selected in scenarios of both high and low total LD with other variants in the region. High and low LD SNPs were defined as those with $l_j = \sum_k r_{jk}^2$ in the top or bottom quartile for the locus. To assess the contribution of annotation across methods, I considered scenarios with one causal SNP near the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th and 100th percentiles of the distribution of random effects variances $\sum_d a_{jd}\tau_d$, where the sum is taken over all annotation categories in the model. Characteristics of the selected causal variants are reported in **Table 1**.

In these simulated samples, I compared the ranking by AnnoRE with (1) a naïve GWAS approach based solely on marginal association p-values, (2) LASSO regression using subject level genotype and phenotype data, (3) the PAINTOR method [34] with the top 5 most highly enriched annotation categories (conserved regions, extended H4K5me1 peaks, extended H3K9ac peaks, H3K9ac peaks, and extended Super Enhancers [20]) and (4) the GenoWAP method, using prior probabilities of causality pre-computed from NHGRI GWAS catalog training data [39, 40].

2.3.2. Results of Simulation Study

AnnoRE gave a higher ranking on average across 1000 replicates to the true causal SNP than comparison methods for all selected causal SNPs, except in three scenarios where the causal SNP was in at least the 90th percentile of functional annotation and PAINTOR gave higher average ranking (**Table 2.1 and Figure 2.1**). The causal SNPs that PAINTOR ranked higher than AnnoRE were located in 3 or 4 of the annotation categories considered in the PAINTOR estimation, suggesting that this method is powerful when the selected annotations contain the true causal SNPs, but is sensitive to the choice of annotation categories included in the model. In fact, PAINTOR often failed to converge entirely for SNPs not included in any of the annotation categories it considered. AnnoRE is more robust to scenarios of moderate evidence from annotation because it summarizes the effects of a large number of annotation categories which are estimated from genome-wide summary statistics.

		overall	**Anno.**	**Average rank at causal SNP**				
Causal SNP	**MAF**	**LD**	**Pctile**	**GWAS**	**LASSO**	**GenoWAP**	**PAINTOR**	**Anno.RE**
rs9974258	0.413	low	10	69.2	55.3	1191.3	NA	23.0
rs77960433	0.037	low	20	1602.2	48.1	2076.9	2129.2	14.4
rs433893	0.381	low	30	42.1	52.5	1466.8	NA	13.5
rs2830795	0.278	low	40	1931.4	49.8	946.1	2142.6	13.9
rs2830794	0.493	low	50	1414.6	52.1	1003.7	2098.7	19.3
rs189506146	0.305	low	60	52.1	50.6	1333.0	28.3	6.8
rs235952	0.193	low	70	59.0	45.2	1201.8	13.8	6.6
rs235938	0.467	low	80	59.9	50.8	1212.5	18.6	6.8
rs235936	0.495	low	90	54.4	50.9	1040.7	4.1	7.0
rs2830854	0.393	low	100	53.7	49.7	1704.9	1.8	6.5
rs381814	0.286	high	10	70.5	67.7	1090.4	2044.7	29.9
rs229093	0.299	high	20	75.0	22.6	981.3	NA	52.1
rs7281968	0.33	high	30	77.4	NA	32.7	NA	49.5
rs229087	0.303	high	40	72.0	NA	27.2	1938.1	53.8
rs229061	0.305	high	50	88.4	NA	505.6	NA	58.3
rs229060	0.305	high	60	72.6	NA	99.4	146.6	34.1
rs162497	0.331	high	70	76.1	58.4	86.3	154.2	28.3
rs229063	0.305	high	80	81.5	NA	473.9	165.8	30.7
rs371445	0.292	high	90	82.5	72.9	327.0	78.7	15.9
rs162508	0.282	high	100	72.3	NA	73.2	5.2	25.4

Table 2.1. Selected causal SNPs characteristics: minor allele frequency (MAF), overall LD $l_j = \sum_{k\in locus} r_{jk}$ in the bottom quartile (low) or top quartile (high) within the locus, and percentile of annotation score $\sum_d a_{jd}\tau_d$. Methods used to evaluate simulated data are compared in terms of average rank of the true causal variant across 1,000 replicates. LASSO was unable to estimate several SNPs due to high collinearity with other variants in the locus. Additionally, the PAINTOR

algorithm did not converge for 4 scenarios where the causal SNP was not located in any of the most relevant annotations included in the PAINTOR model.

All methods performed better in scenarios where the simulated causal variant was in low LD with all other SNPs in the locus (top row of **Figure 2.1**), except for GenoWAP, which does not model LD. However, the improved performance of GenoWAP at high LD variants is not consistent across scenarios. The prior probabilities of causality used by GenoWAP are different from the annotations used by AnnoRE and PAINTOR in these analyses, possibly accounting for the discrepancy. Additionally, the EM algorithm used by PAINTOR to estimate simultaneously the annotation-level heritability enrichment and the SNP-level effects was unable to converge in simulation scenarios where the true causal SNP was not located in any of the top 5 most highly enriched annotation categories, which were the only annotation data provided to PAINTOR. The developers of PAINTOR, in showing its application to practical data, used a stepwise selection procedure to choose a set of annotations for inclusion in the final model based on goodness-of-fit statistics from K fine mapping models, each with only one annotation category included. However, the annotation feature selection step is outside the scope of this project, so PAINTOR results are presented based only on the scenarios that reached convergence.

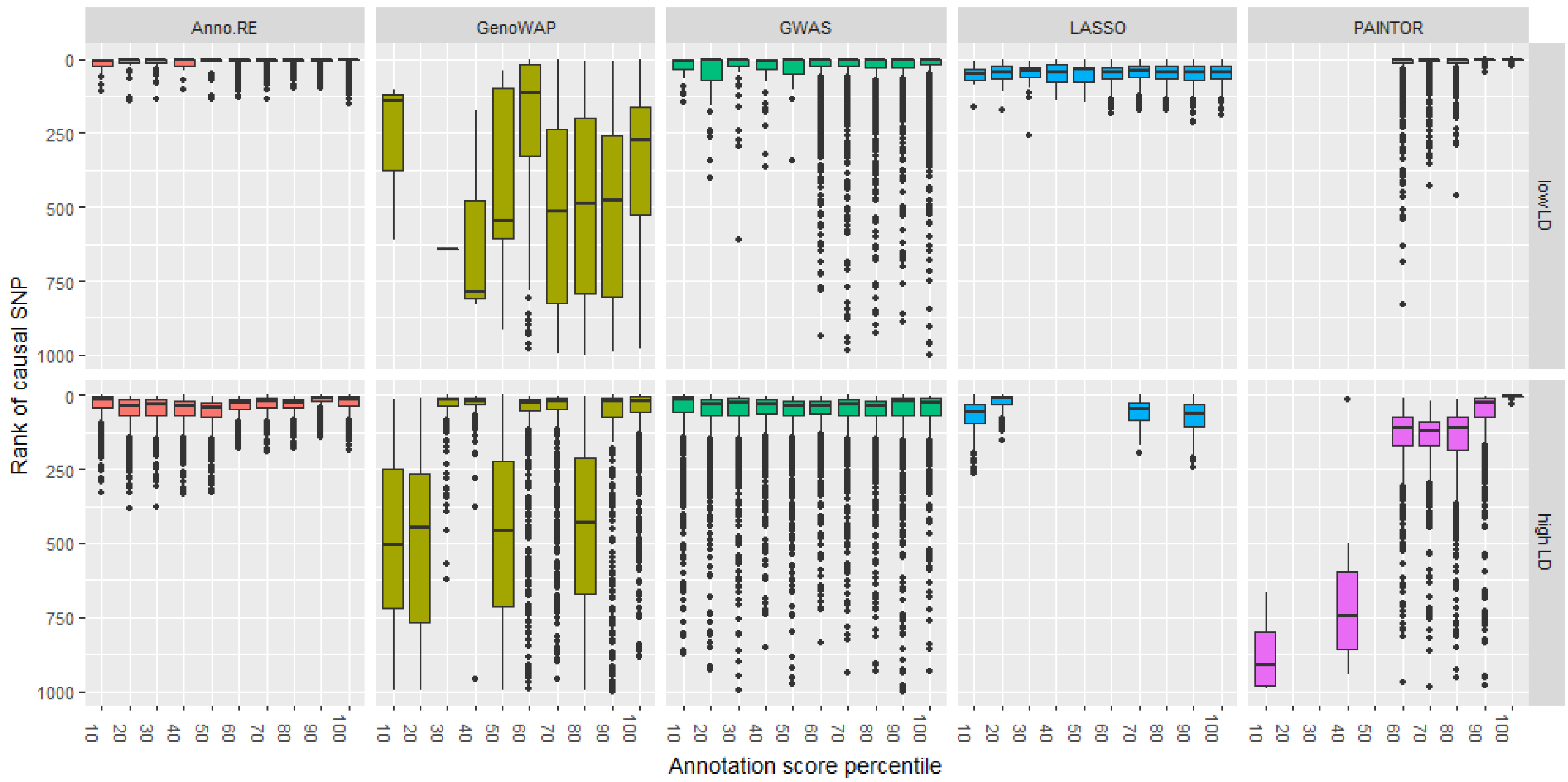

Figure 2.1. Box plots of ranking of simulated causal SNP within fine mapping locus by all methods under comparison, truncated to only show rankings in the top 500 SNPs. Note that LASSO was unable to obtain estimates in 6 high LD scenarios due to extreme collinearity, and PAINTOR estimation did not converge when the causal SNP was not included in any of the five annotation categories considered by that analysis.

2.4. Real Data Analysis

Body mass index (BMI) is an important risk factor for numerous diseases, such as Type 2 diabetes, hypertension, and heart disease. It is highly heritable, with twin study estimates of the genetic contributions accounting for 49-90% of trait variance [19]. The Genetic Investigation of Anthropometric Traits (GIANT) consortium is an international collaboration studying the genetic basis of anthropometric phenotypes including BMI. I performed AnnoRE analysis fine mapping with summary statistics from the GIANT GWAS meta-analysis of BMI in 322,154 subjects of European ancestry [37]. This study reported 77 loci with the strongest signals separated by at least 500kb, and reaching genome-wide significance ($p < 5 \times 10^{-8}$) in the European ancestry sample. Summary statistics for all reported SNPs within a 500kb radius of the most significantly associated markers were extracted for fine mapping, with LD information from the 1000 Genomes Project Phase 1 European ancestry reference panel.

Within each locus, I defined a locus-specific significance threshold by Bonferroni correction based on the number of SNPs included in the fine mapping analysis, to adjust for multiple testing of the BLUP Wald statistics. In the analysis of summary statistics from 77 loci with genome-wide significant associations with BMI in the GIANT consortium meta-analysis of European ancestry studies, 6 loci contained results exceeding the locus-specific significance threshold. In each of these cases, the top SNP selected by fine mapping was different from the SNP with lowest GWAS p-value, though all of the top fine mapping SNPs did exceed genome-wide significance in the GWAS results (**Table 2.2**).

SNP	Chr: position	Nearest gene	LD D' and r^2	MAF	GWAS $\hat{\beta}$	GWA.pval	AnnoRE $\tilde{\beta}$	Wald scores	AnnoRE pval	GWA rank	AnnoRE rank
rs1477196	16:53808258	FTO	1.0	0.28	-0.058	3.22E-72	-0.147	348.049	1.13E-77	50	1
rs1558902	16:53803574		0.41	0.45	0.082	7.51E-153	0.014	2.930	0.087	1	31
rs34358	5:74965122	LOC441087	0.96	0.35	0.023	2.31E-12	0.037	37.200	1.07E-09	11	1
rs2112347	5:75015242		0.87	0.38	-0.026	6.19E-17	-0.001	0.368	0.544	1	56
rs988748	11:27724745	BDNF	0.99	0.22	-0.041	5.90E-23	-0.027	33.935	5.70E-09	16	1
rs11030104	11:27684517		0.89	0.20	0.041	5.56E-28	0.003	0.729	0.393	1	43
rs11676272	2:25141538	ADCY3	1.0	0.48	0.032	1.12E-21	0.031	28.952	7.42E-08	3	1
rs10182181	2:25150296		0.99	0.50	-0.031	8.78E-24	-0.009	5.028	0.025	1	4
rs1048932	11:115044850	CADM1	0.90	0.50	-0.019	9.43E-10	-0.019	17.367	3.08E-05	2	1
rs12286929	11:115022404		0.57	0.43	0.022	1.31E-12	0.002	1.884	0.170	1	7
rs1800437	19:46181392	QPCTL	0.99	0.18	0.035	1.73E-17	0.022	12.912	3.27E-04	3	1
rs2287019	19:46202172		0.89	0.15	0.036	4.59E-18	0.002	1.159	0.282	1	10

Table 2.2. Top ranked SNPs from random effects fine mapping (AnnoRE) and GWAS analysis in each of the loci with AnnoRE results attaining locus-wide significance.

The strongest fine mapping signal, rs1477196, identified by the AnnoRE is in the FTO locus on chromosome 16 (**Figure 2.2**) which has been studied extensively in the genetics of obesity [59]. Identified in association with obesity in previous publications [46, 25], this SNP is located in a haplotype block with two other highly associated SNPs, rs17817449 and rs9939609. However, the random effect variance for rs1477196 is higher than that of the other SNPs on the haplotype due to its inclusion in the super enhancer and H3K27ac histone mark annotation categories, both of which were estimated to contain enriched heritability signal for BMI by LD score regression.

In the *ANKDD1B* locus on chromosome 5, the top GWAS signal is located in an intron of the gene, while the top SNP selected by AnnoRE, rs34358, is a stop gain mutation in an exon of the gene, located in a highly conserved region across vertebrate species, a ChIP-seq peak and DNaseI hypersensitive site.

The *BDNF* (brain-derived neurotrophic factor) gene on chromosome 11 has been implicated in numerous psychiatric and neurological diseases [11, 2]. Both the top GWAS hit in this locus, rs11030104, and the top SNP selected by AnnoRE fine mapping, rs988748, are located in conserved regions in introns of this gene. The SNP selected by AnnoRE fine mapping is included in the super enhancer, fetal DNaseI hypersensitive site, and H3K4me3 peak annotation categories, while the SNP with lowest GWAS p-value is not in these important categories.

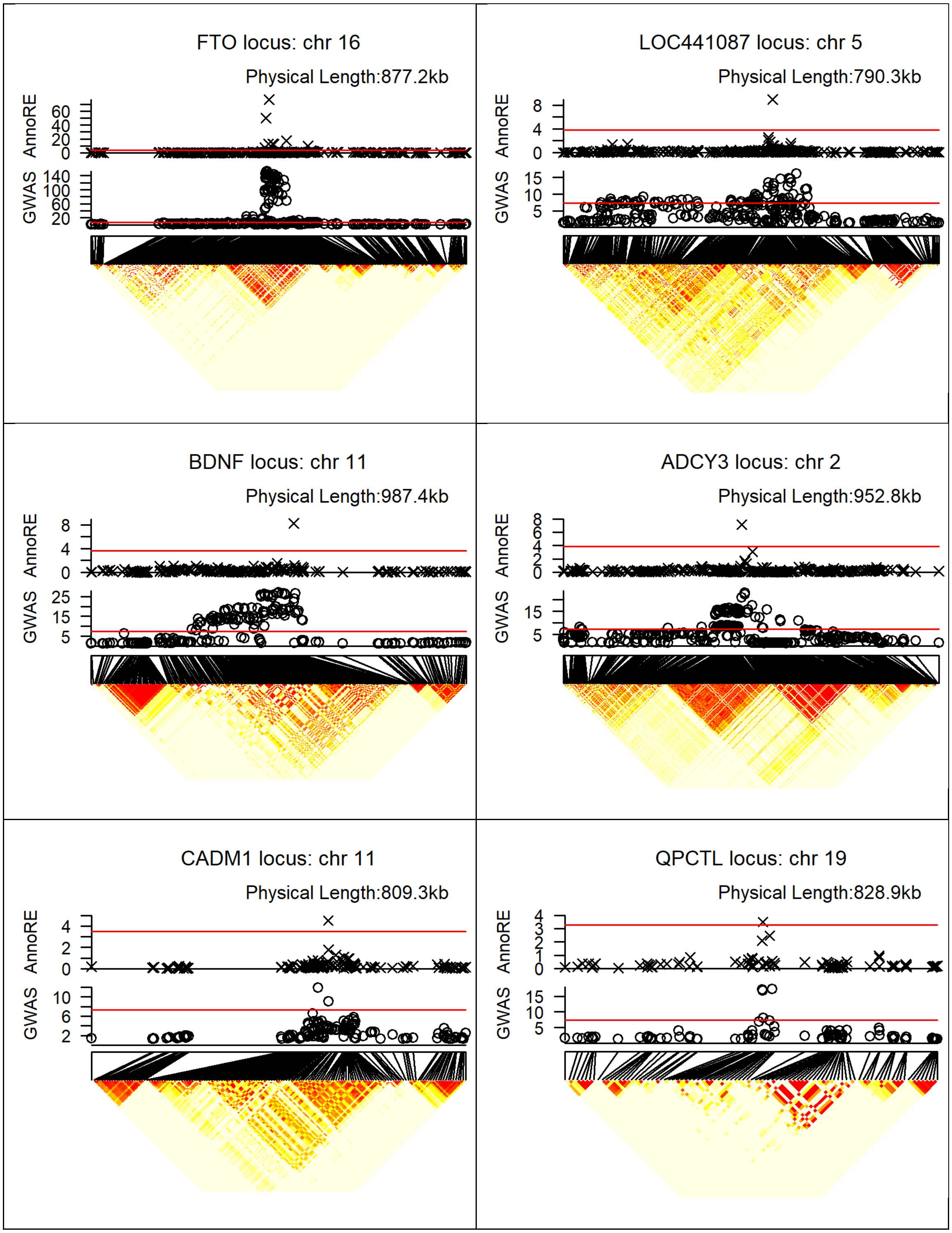

Figure 2.2. Plots of local LD structure, -log$_{10}$ GWAS p-values, and –log$_{10}$ AnnoRE p-values in six loci with significant fine mapping results. Horizontal red lines show genome-wide (GWAS) and locus-wide (AnnoRE) significance thresholds, respectively. Physical length in kilobase (kb) refers to the number of base pairs (in 1000s) from beginning to end of locus, defined by nominally associated SNPs less than 500kb from the strongest GWAS association.

AnnoRE identified the SNP with lowest GWAS p-value as the most likely causal variant in 13 (16.8%) of the 77 loci analyzed. Among the remaining loci, the median GWAS ranking of the top fine mapping SNP was 14, and the median AnnoRE fine mapping ranking of the top GWAS SNP was 9. Remarkably, only 39 (51%) of the top SNPs identified by AnnoRE attained genome-wide significance ($p < 5 \times 10^{-8}$) in univariate GWAS analysis.

To assess the ability of our fine mapping method to discriminate between SNPs in the locus, I computed the ratio between the random effect Wald statistics of the top ranked SNP, and the second and third ranked SNPs. In 17 (22%) of the loci, the signal at the top ranked SNP was more than twice as strong as the second ranked, and in 30 (39%) of loci, the top ranked signal was more than twice the third ranked signal.

2.5. Discussion and Conclusions

I present AnnoRE, a method for genetic fine mapping incorporating functional annotation. This model uses random effects to perform multiple regression with smoothing of the SNP effect estimates dependent on their functional annotation, so that SNPs in categories with enriched heritability receive less shrinkage relative to those without evidence of biological function.

AnnoRE makes no assumptions regarding the number of true causal signals, because all SNPs in a given locus are included as predictors in the random effects model. This approach is compatible with the omnigenic or infinitesimal models of inheritance, which hypothesize that complex traits are influenced by large numbers of genetic variants with small effect magnitudes [4, 7]. Under this model, I assume that a locus containing functional elements that affect trait outcomes may contain several causal SNPs modifying the protein product of genes or regulatory elements that control gene expression. However, existing fine mapping methods generally assume a model of inheritance in which only a few SNPs are truly causal, and all other association signals are artifacts of confounding due to LD. In regions of high LD, the tagging effect of numerous weak signals may substantially influence test statistics, in ways that cannot be captured by a model that assumes zero effect sizes at all but a few SNPs.

Using a two-stage procedure for the estimation of the annotation-level effects and the SNP-level effects allows for improvements in computational efficiency while exploiting information about the annotation effects from all genome-wide test statistics. Thus, AnnoRE is able to account for systematic enrichments of association signal below the genome-wide significance level to identify functional categories that are more likely to contain causal SNPs. On the other hand, the AnnoRE approach assumes that the distribution of trait heritability among functional categories is consistent genome-wide. If a given annotation category is important only in specific genomic locations, AnnoRE may fail to identify causal SNPs in that category.

The AnnoRE method shows superior performance to naïve GWAS ranking, LASSO penalized regression, and GenoWAP [40] in simulation studies across a range of LD structures and annotation scenarios. PAINTOR [34] showed superior performance when the five annotation categories provided for its model included at least three containing the causal variant. Because PAINTOR is only capable of considering a few annotation categories simultaneously, it is at a disadvantage for causal variants outside of those annotations and must be conduct several times to build a small set of strongly enriched annotations, requiring substantial computation time. AnnoRE ranked the true causal variant in the top 10 percent by ordering of the fine mapping statistics, on average across the 1000 simulation replicates, even in simulation scenarios where LASSO and PAINTOR were unable to obtain estimates at all.

In our analysis of loci identified by a large GWAS meta-analysis of BMI, I found six loci where the top variant identified by random effects fine mapping exceeded a locus-wide significance threshold. In all of these cases, the top SNP selected during fine mapping was in very high LD with the most significant GWAS signal, with D' statistic greater than 0.9 in all six loci. The D' statistic accounts for differences in allele frequency, whereas the standard correlation statistic r^2 cannot attain its maximum value of one between SNP genotypes with different allele frequencies. In these six loci, the SNPs selected as most significant by AnnoRE fine mapping are located in annotation categories with greater plausible functional relevance that the top SNPs selected by GWAS. In 25 additional loci, the AnnoRE test statistic at the top SNP is more than twice the magnitude at the third ranked SNP. These loci are promising candidates for further exploration, as the fine mapping analysis distinguishes one or two SNPs with stronger evidence of causality relative to others in the locus.

Our proposed method addresses the problem of collinearity due to high LD among sets of SNPs within a locus by defining random effect variances that depend on SNP annotation. The resultant estimator is similar to a penalized method such as ridge regression [29], with the smoothing penalty differing by SNP annotation. Thus, even if two SNPs are in perfect

LD, the one with stronger annotation evidence will receive a larger effect estimate. For this reason, it is desirable to include many annotation categories in the estimation of variance components to ensure that differences in annotation allow the estimator to prioritize the genetic variants within LD blocks. However, the variance component estimates may be unstable when there is high correlation among the annotation marks themselves, as may be the case for tissue-specific annotation of the same signal across similar tissues. Further work is required to identify optimal sets of annotation categories for use in fine mapping studies.

One limitation of AnnoRE is the fact that the distribution of the test statistics is conditional upon the variance components estimated by LD score regression and does not account for the uncertainty in these estimates. Large GWAS sample sizes reduce the variability of these estimates, making this simplifying assumption more acceptable. This method is designed to identify causal SNPs in common allele frequency ranges and may be less powerful in the case of rare variants, as is also the case for other existing fine mapping methods.

In summary, I have proposed a framework to prioritize variants with known biological relevance that are associated with the phenotype independently of other variants in the locus. Integrating local LD structure and functional annotation, this proposed approach can either be applied to individual level data or utilize GWAS summary statistics data. The resulting SNPs are promising candidates for the functional follow-up studies that are necessary to translate findings from genetic epidemiology towards increased understanding of human biology and clinical applications.

CHAPTER 3. PARTITIONING HERITABILITY BY GENOMIC ANNOTATION

3.1. Introduction

Heritability is a fundamental concept in statistical genetics. Informally, it represents the proportion of phenotype variation attributable to genetic effects. To understand the genetic architecture of a given trait, it is useful to partition the heritability by categories of genetic variants, to investigate which categories have greater effect on the trait than others. Such an analysis contributes to the interpretation of genome-wide association studies (GWAS) by translating results from the level of individual variants to that of genome-wide functional features. These category-level enrichment estimates may be used to estimate variant-level effect distributions, such as those presented in Chapter 2. In Chapter 1, I described two previously published methods for estimating partitioned heritability--stratified LD score regression (LDSC) [20], and MinQue for Summary statistics (MQS) [60]. Both of these methods can result in estimates of heritability outside of the permissible range of 0-1. In addition, LDSC estimates partitioned heritability from GWAS summary statistics, but its linear regression model is only a first-order approximation of the relationship between SNP categories and heritability, yielding highly variable estimates from GWAS of moderate sample size. The MQS method also does not require individual-level data, but does require summary statistics beyond those calculated for a standard GWAS, and is only capable of analyzing non-overlapping categories. For this project, I present an alternative approach based on maximum-likelihood estimation (MLE) from GWAS test statistics, and perform simulation studies to compare performance of these methods across different numbers of truly enriched categories.

3.2. Methods

Suppose the true genetic model for the continuous trait of interest is an additive linear combination of *M* genetic variants:

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

In this model for *N* individuals, *y* is a $N \times 1$ vector and $\boldsymbol{X}$ is a matrix of genotypes. For each individual, $\boldsymbol{x_i}$ represents a genotype row vector of length *M* for the i^{th} subject, β is a $M \times 1$ vector of random regression coefficients and ϵ is random residual error. For this model, genotypes are standardized. For a given SNP j, define G_{ij} to be the genotype of subject i at SNP j, coded as (0,1,2). In a sample of N unlrelated individuals, $\widehat{p_j}$ is the (sample) minor allele frequency, the standardized genotype vector X_j with entries $x_{ij} = \frac{1}{\sqrt{2\widehat{p_j}(1-\widehat{p_j})}}(G_{ij} - 2\widehat{p_j})$ has mean zero and $X_j^T X_j = N$. Standardizing the genotypes induces a scaling of the genetic effects. If γ_j is the per-allele effect of SNP j coded as G_j without standardization, then $\beta_j = \sqrt{2\widehat{p_j}(1-\widehat{p_j})}\gamma_j$ is the effect size of SNP j coded as the standardized genotype X_j. Without loss of generality, the vector *y* of phenotype values is also assumed to be standardized to have mean of zero and unit variance.

Under this model, where all causal SNPs are observed, and all genetic effects are linear (in particular, precluding gene-gene or gene-environment interactions), we may define the trait heritability attributable to a set of SNPs $G = \{1, \dots, M\}$ as $h_G^2 = \sum_{j \in G} \beta_j^2$, and the partitioned heritability attributed to annotation category C_k as $h_k^2 = \sum_{j \in C_k} \beta_j^2$. When the categories $(C_1, \dots, C_K)$ overlap, it may be the case that $h_G^2 \leq \sum_{k=1}^{K} h_k^2$. When the categories are disjoint (non-overlapping), it coincides with the definition of partitioned heritability used by Gusev *et al.,* who "define the h_g^2 for each functional category as the squared correlation r^2 between the true phenotype and the prediction only from SNPs in that functional category when all functional categories are jointly analyzed for a best linear prediction." [26]

3.2.1. Model and Likelihood

To aggregate association signals by functional category, the random effect of each (standardized) genotype X_j is modeled, independently of all other genetic effects, with the distribution

$$\beta_j \sim N\left(0, \sum_{k=0}^{K} a_{jk}\tau_k\right)$$

Here $a_{j1}, \dots, a_{jK}$ are binary annotation indicators for *K* categories, and $a_{j0} = 1$ represents background heritability not attributable to these categories. The category-specific variance components τ_k are the estimand of interest.

Linkage disequilibrium (LD) leads to tagging of causal signals at nearby markers which are associated with the phenotype solely due to correlation with the true causal SNPs. This is the same phenomenon as confounding or omitted-variable bias. In an unrelated sample with no population structure and $\boldsymbol{\epsilon} \sim MVN(0, \sigma_R^2 I_N)$ where σ_R^2 is the residual variance of the trait not attributable to additive genetic effects, GWAS tests of univariate association use the least squares estimator:

$$\widehat{\beta_J} = (X_j^T X_j)^{-1} X_j^T \boldsymbol{y} = \frac{1}{N} X_j^T (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

$$= \frac{1}{N}\sum_{j'=1}^{M} X_j^T X_{j'}\beta_{j'} + \frac{1}{N}X_j^T\boldsymbol{\epsilon} = \sum_{j'=1}^{M} \hat{r}_{jj'}\,\beta_{j'} + \boldsymbol{\epsilon}'$$

In this notation, $\hat{r}_{jj'} = \frac{1}{N}X_j^T X_{j'}$ is the maximum likelihood estimator of the correlation between SNPs j and j', and $\epsilon' = \frac{1}{N}X_j^T\boldsymbol{\epsilon} \sim N(0, \frac{1}{N}\sigma_R^2)$.

The GWAS Wald test statistics are defined in terms of the estimated residual variance, from the equality of $X_j^T X_j = N$, so that $\sqrt{\left(X_j^T X_j\right)^{-1}} = \frac{1}{\sqrt{N}}$ as a result of genotype standardization, and the uniform minimum variance unbiased estimator of the standard error of $\widehat{\beta}_J$ given as $\widehat{SE}(\widehat{\beta}_J)^2 = \frac{1}{N-2}\hat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\hat{\boldsymbol{\epsilon}}$ where $\hat{\epsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - X_J\widehat{\beta}_J = X_j\left(X_j^T X_j\right)^{-1}X_j^T\boldsymbol{y}$ and because this estimator is unbiased, it converges in expectation to $Var\left(\widehat{\beta}_J\right) = \frac{1}{N}\sigma_R^2$

$\hat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\hat{\boldsymbol{\epsilon}}$ converges in expectation to $N * Var(\hat{\boldsymbol{\epsilon}}_{\iota}) = \sigma_R^2\left(1 - \left[X_j\left(X_j^T X_j\right)^{-1}X_j^T\right]_{ii}\right)$

$$Z_j = \frac{\widehat{\beta}_J}{\widehat{SE}(\widehat{\beta_J})} = \frac{\left(X_j^T X_j\right)^{-1}X_j^T\boldsymbol{y}}{\sqrt{\left(X_j^T X_j\right)^{-1}\frac{1}{N-2}\hat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\hat{\boldsymbol{\epsilon}}}} = \frac{\frac{1}{\sqrt{N}}X_j^T\boldsymbol{y}}{\sqrt{\frac{1}{N-2}(\boldsymbol{y}-\hat{\boldsymbol{y}})^T(\boldsymbol{y}-\hat{\boldsymbol{y}})}}$$

$$\xrightarrow{in\ expectation} \sqrt{N}\frac{Cov(X_j, \boldsymbol{y})}{Var(\boldsymbol{\epsilon})}$$

When the trait is highly polygenic, each univariate GWAS model will only explain a small proportion of phenotypic variance, so that $Var(\boldsymbol{\epsilon}) \approx Var(y) = 1$ and $W_j \approx N\widehat{\beta}_J^{\,2}$. This relationship allows us to approximate the expectation of GWAS Z scores as a weighted sum of true effect sizes at all causal SNPs in LD with SNP j, with weighting according to the strength of LD correlation: $E\left[\widehat{\beta}_J\right] = \sum_{j'=1}^{M}\widehat{r_{jj'}}\beta_{j'}$.

To quantify the strength of association without cancelling of effects in opposite directions, consider the distribution of GWAS χ^2 statistics of association. $W_j = \frac{\widehat{\beta_J}^2}{Var(\widehat{\beta_J})}$, which is distributed as χ^2 with 1 df under the global null hypothesis $\beta_j = 0$ for $j = 1, \dots, M$. If any SNPs have true non-zero effects, then the distribution is non-central χ^2 with non-centrality parameter related to that of the Z scores derived above: $W_j \sim NC\chi_1^2(\lambda_j)$. In Chapter 1, I described the LD Score Regression method [20], which depends on derivation of the expected value of GWAS χ^2 statistics as a function of stratified LD scores $l_{jk} = \sum_{j' \in k} r_{jj'}^2$ quantifying the extent to which each SNP j tags variation in annotation category k:

$$E[W_j] = N \sum_{k=1}^{K} l_{jk}\tau_k + 1$$

Because the expected value of a $NC\chi^2$ variable is equal to the sum of its non-centrality parameters and its degrees of freedom, define $\lambda_j = N \sum_{k=0}^{K} l_{jk}\tau_k$. To ensure non-negativity of the estimates, and to avoid the problem of variance component estimates on the boundary of the parameter space ($\tau_k = 0$), I reparametrize the model with the transformation $\theta_k^2 = N\tau_k$. The invariance principle of maximum likelihood estimation ensures that finding the MLE of $\boldsymbol{\theta}$ and inverting the transformation will yield the MLE of $\boldsymbol{\tau}$. The suggestion of transforming variance components to avoid the problems of estimation caused by the bounded support on non-negative values was proposed by Box and Tiao [6]. This estimator may be biased, especially if the true parameter value is on the boundary, i.e. $\tau_k = 0$, but the properties of the MLE guarantee asymptotic unbiasedness.

To calculate this estimator, we maximize the complete data log likelihood function with respect to $\boldsymbol{\theta}$:

$$\ell(W;\theta) = \frac{-1}{2}\sum_{j=1}^{M}\sum_{k=1}^{K} l_{jk}\theta_k^2 + \sum_{j=1}^{M}\log\left(e^{\sqrt{w_j\sum_k l_{jk}\theta_k^2}} + e^{-\sqrt{w_j\sum_k l_{jk}\theta_k^2}}\right)$$

3.2.2. Gradient and Hessian

To perform optimization by means of a quasi-Newton algorithm requires the gradient function

$$\nabla\ell(W;\theta) = \left[\frac{\partial}{\partial\theta_k}\ell(W;\theta)\right]$$

Using the notation $\lambda_j(\theta) = \sum_{k=1}^{K} l_{jk}\theta_k^2$ with $\frac{\partial}{\partial\theta_q}\lambda_j(\theta) = 2\theta_q l_{jq}$ and $u_j(\theta) = \sqrt{w_j\lambda_j(\theta)}$ with $\frac{\partial}{\partial\theta_q}u_j(\theta) = \theta_q l_{jq}\sqrt{\frac{w_j}{\lambda_j(\theta)}}$, the *pth* element of the gradient is:

$$\frac{\partial}{\partial\theta_p}\ell(W,\theta) = \frac{-1}{2}\sum_{j=1}^{M}\frac{\partial}{\partial\theta_p}\lambda_j(\theta) + \sum_{j=1}^{M}\frac{\partial}{\partial\theta_p}log(e^{u_j(\theta)} + e^{-u_j(\theta)})$$

$$= \frac{-1}{2}\sum_{j=1}^{M}2\theta_p l_{jp} + \sum_{j=1}^{M}\frac{\frac{\partial}{\partial u_j}(e^{u_j(\theta)} + e^{-u_j(\theta)})}{e^{u_j(\theta)} + e^{-u_j(\theta)}}\frac{\partial}{\partial\theta_p}u_j(\theta)$$

$$= -\theta_p\sum_{j=1}^{M} l_{jp} + \sum_{j=1}^{M}\frac{e^{u_j(\theta)} - e^{-u_j(\theta)}}{e^{u_j(\theta)} + e^{-u_j(\theta)}}\theta_p l_{jp}\sqrt{\frac{w_j}{\lambda_j(\theta)}}$$

$$= -\theta_p \sum_{j=1}^{M} l_{jp} \, [1 - \sqrt{\frac{w_j}{\lambda_j(\theta)}} \, (\frac{e^{u_j(\theta)} - e^{-u_j(\theta)}}{e^{u_j(\theta)} + e^{-u_j(\theta)}})]$$

Because the likelihood function is twice differentiable, and the parameter vector $\boldsymbol{\theta}$ (though not $\boldsymbol{\tau}$) is unbounded on $\mathbb{R}^K$, the regularity conditions are met and it is thus possible to use the asymptotic distribution of the MLE to derive standard errors of the estimators: $\widehat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, I(\theta)^{-1})$, from the information matrix *I(θ)*. [18]

The information matrix contains second derivatives of the log likelihood function. For this derivation, define $u_j = \sqrt{w_j \sum_k l_{jk} \theta_k^2}$, and $\delta(k,k') = 1$ if $k = k'$ and zero otherwise, so that the (k,k') element of $I(\theta)$ is given by:

$$\frac{\partial^2}{\partial\theta_q \, \partial\theta_p} \ell(W,\theta) = -\theta_p \sum_{j=1}^{M} l_{jp} \frac{\partial}{\partial\theta_q} [1 - \sqrt{\frac{w_j}{\lambda_j(\theta)}} \, (\frac{e^{u_j(\theta)} - e^{-u_j(\theta)}}{e^{u_j(\theta)} + e^{-u_j(\theta)}})]$$

$$= \theta_p \sum_{j=1}^{M} l_{jp} \left[\sqrt{\frac{w_j}{\lambda_j(\theta)}} \frac{\partial}{\partial u_j} \left(\frac{e^{u_j(\theta)} - e^{-u_j(\theta)}}{e^{u_j(\theta)} + e^{-u_j(\theta)}} \right) \frac{\partial u_j}{\partial\theta_q} + \frac{\partial}{\partial\lambda_j} \sqrt{\frac{w_j}{\lambda_j(\theta)}} \frac{\partial\lambda_j}{\partial\theta_q} \left(\frac{e^{u_j(\theta)} - e^{-u_j(\theta)}}{e^{u_j(\theta)} + e^{-u_j(\theta)}} \right) \right]$$

$$= \theta_p \theta_q \sum_{j=1}^{M} l_{jp} l_{jq} \left[\frac{w_j}{\lambda_j(\theta)} \left(1 - \left(\frac{e^{u_j(\theta)} - e^{-u_j(\theta)}}{e^{u_j(\theta)} + e^{-u_j(\theta)}} \right)^2 \right) - \sqrt{\frac{w_j}{\lambda_j(\theta)^3}} \left(\frac{e^{u_j(\theta)} - e^{-u_j(\theta)}}{e^{u_j(\theta)} + e^{-u_j(\theta)}} \right) \right]$$
$$- \delta(p,q) \sum_{j=1}^{M} l_{jp} \left[1 - \sqrt{\frac{w_j}{\lambda_j(\theta)}} \left(\frac{e^{u_j(\theta)} - e^{-u_j(\theta)}}{e^{u_j(\theta)} + e^{-u_j(\theta)}} \right) \right]$$

The last equality follows from the derivatives:

$$\frac{\partial}{\partial u_j} (\frac{e^{u_j(\theta)} - e^{-u_j(\theta)}}{e^{u_j(\theta)} + e^{-u_j(\theta)}}) = 1 - (\frac{e^{u_j(\theta)} - e^{-u_j(\theta)}}{e^{u_j(\theta)} + e^{-u_j(\theta)}})^2$$
$$\frac{\partial}{\partial\lambda_j} \sqrt{\frac{w_j}{\lambda_j(\theta)}} = \frac{-1}{2} \sqrt{\frac{w_j}{\lambda_j(\theta)^3}}$$

Inverting the observed Information Matrix, consisting of the Hessian evaluated at the MLEs gives the estimated variance-covariance matrix of the estimators. Standard errors for each parameter estimate are obtained from square roots of the diagonal elements of this inverse.

Confidence intervals of τ_k are obtained by transforming the endpoints of confidence intervals for θ_k, keeping in mind that if the endpoints have opposite sign, the interval is of the form $[0, \frac{1}{N}\max(\theta_{LL}^2, \theta_{UL}^2))$.

3.2.3. Estimation algorithms

Optimization was performed with the variable metric algorithm published independently in 1970 by Broyden, Fletcher, Goldfarb and Shanno (BFGS) [9, 21, 24, 48]. This algorithm belongs to the class of quasi-Newton methods, which approximate the quadratic Taylor expansion of the target function. If $f(x)$ is a continuous, twice-differentiable scalar-valued function with vector input *x*, which attains its maximum value at x_0, then the gradient function $\nabla f(x_0) = 0$. The optimization algorithm performs an iterative search for this optimum point by means of the update equation $x_{n+1} = x_n + \alpha p_n$ using the second order Taylor series, with $\mathbf{H}f(x_n)$ representing the Hessian matrix of second derivatives:

$$0 = \frac{\delta}{\delta p}(f(x_n) + \nabla f(x_n)p + \mathbf{H}f(x_n)p^2)$$

Solving this with respect to p gives $p_n = -[\mathbf{H}f(x_n)]^{-1}\nabla f(x_n)$, and the step size α is then found by maximizing the one-dimensional function $f(x_n + \alpha p_n)$. Newton's method calculates these updates by solving a system of linear equations $[\mathbf{H}f(x_n)]p = -\nabla f(x_n)$, while quasi-Newton methods use an approximation $B_n \approx [\mathbf{H}f(x_n)]$ which is updated at each iteration to

$$B_{n+1} = B_n + \frac{y_n y_n^T}{y_n^T s_n} - \frac{B_n s_n s_n^T B_n^T}{s_n^T B_n s_n},$$

where $y_n = \nabla f(x_{n+1}) - \nabla f(x_n)$ and $s_n = \alpha p_n$. Implementation of this algorithm is available in the function optim() in the stats package of the R Project for Statistical Computing.

The non-central χ^2 likelihood function is non-convex and multimodal, due to symmetry introduced by the reparametrization as well as the original functional form. In this setting, gradient-based optimization may identify a local maximum, rather than the desired global maximum. To address this issue, I ran the optim() function 100 times for each simulation replicate with initial values for all parameters sampled independently from a $Uniform(0,1)$ distribution. The optimization result with the maximum value of the likelihood function is selected as the estimator, to obtain point estimates $\widehat{\boldsymbol{\theta}}$ and thus $\hat{\boldsymbol{\tau}} = \frac{1}{N}\widehat{\boldsymbol{\theta}}^{\mathbf{2}}$**.**

3.2.4. Comparison of Methods

The primary comparison of interest is between maximum likelihood estimation of $\boldsymbol{\tau}$, and estimation via stratified LD score regression [20]. The set of annotation categories used in this simulation study were the same as those presented in the Finucane *et al.* paper, and published online at https://data.broadinstitute.org/alkesgroup/LDSCORE/ [1].

These annotations represent dichotomized results from several distinct genomic functional elements published as part of the ENCODE project [22]. These genomic elements exons, introns, and untranslated regions of genes, as well as promoter and enhancer regions, highly conserved regions, and specific regulatory elements including DNaseI hypersensitive sites, transcription factor binding motifs, and others (see Appendix B). Corresponding to each of the 24 core categories, an extended category was defined by enlarging the boundaries of the annotated regions by 500 base pairs on each side. These were by Finucane *et al.* to capture possible measurement error in the process of dichotomizing the annotations [20]. For four of the core categories, derived from raw annotations that proved difficult to dichotomize neatly, additional peak annotation categories were defined. Finally, a baseline category was included containing all SNPs, for a total of $K = 53$ binary annotation categories considered in the analysis.

The stratified LD scores for these annotation categories, calculated from 1000 Genomes Project sample of European ancestry were also downloaded from the LD Score Regression project website [1]. As recommended by the developers, only common (MAF>0.05) SNPs in the HapMap3 panel were used as LD score observations, though all SNPs with annotation were used in calculation of the scores. All simulations presented in this chapter are based on chromosome 21 only.

I also compared performance with the Method of Moments approach implemented in GEMMA, which can only analyze disjoint annotation categories [12, 60]. To make this possible, I defined disjoint categories based on the K=28 core and peak annotations by assigning each SNP to the smallest category in which it is located. The intention of this approach is to facilitate estimation of per-SNP heritability in all categories by avoiding any disjoint category to contain too few SNPs.

3.2.5. Simulation Study

To assess performance of these methods, I simulated GWAS summary statistics. These were based on a synthetic genotype sample of $N = 10{,}000$ individuals generated from the European ancestry reference panel in the 1000 Genomes Project Phase 1 [14], with HAPGEN2 software [50].

The number of reported SNPs on chromosome 21 was 409,331 of which 129,150 were diallelic with annotation available in the dichotomized ENCODE data. The LD scores for 15,379 HapMap3 variants on chromosome 21 with minor allele frequency at least 0.05 are used in LD score regression and maximum likelihood estimation.

I generated GWAS analysis based on simulated individual level data for a continuous trait with heritability of 0.25 on chromosome 21, under two scenarios of category-specific heritability enrichment. In the first scenario, all 28 core and peak categories were assigned non-zero values of τ_k of random magnitude, while the baseline and extended categories were not enriched for heritability. I drew 28 values from a uniform (0,1) distribution, and linearly rescaled them to attain chromosome-wide heritability of $h_G^2 = 0.25$ by the relationship $h_G^2 = \sum_{k=1}^{K} M_k \tau_k$, where M_k is the number of SNPs in annotation category *k*. In the second scenario, only two categories were given true non-zero values of τ_k. The Enhancer_Andersson ($M_{k_1} = 822$) and H3K4me1_Trynka methylation ($M_{k_2} = 55{,}649$) categories were selected from the extremes of the category sizes to investigate the influence of category size on the estimates of heritability enrichment. These categories were assigned the same true value of $\tau_k = 4.43 \times 10^{-6}$ selected to generate a phenotype with heritability of $h_G^2 = 0.25$. These two categories are highly overlapping, with 736/822 of the SNPs annotated as Enhancers also included in H3K4me1 marks. These categories were chosen to assess performance at both a large annotation category, and a subset with additional enrichment. This mimics the conditions of, for example, including all promoter regions in one annotation, when a subset of these regions are directly involved in regulating trait-related processes. In order to implement the GEMMA analysis, which requires non-overlapping annotations, modified disjoint sets were defined by removing all Enhancer SNPs from the H3K4me1 category.

For each replicate, given a fixed vector of variance components $\boldsymbol{\tau}$**,** I simulated GWAS summary statistics by the following procedure:

1. Define random effect variance $Var(\beta_j) = \sum_{k=1}^{K} a_{jk}\tau_k$ for SNP *j=1, ... , M.*
2. Draw length-M vector $\boldsymbol{\beta}$ of SNP effects by sampling each β_j independently from $N(0, \sum_{k=1}^{K} a_{jk}\tau_k)$distributions with variances defined in step 1.
3. Draw length-N vector ϵ of independent subject-level residuals from $N(0, 0.75)$, to maintain heritability of $h^2 = 0.25$ and unit phenotype variance.
4. Calculate length-N phenotype vector as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$**.** The simulated genotype sample of $N = 10{,}000$ individuals described in the first paragraph of this section was used for all simulation replicates. Because simulated genotypes are on the (0, 1, 2) scale, I multiply each β_j by a factor of $\left(2p_j(1-p_j)\right)^{-1/2}$ to obtain per-allele effect size γ_j.

5. Perform the univariate test of association between each SNP *j=1, ... , M* and phenotype $\boldsymbol{y}$**.** Obtain estimates $\widehat{\gamma_j}$, standard errors, and χ^2 test statistics $W_j = \frac{\widehat{\gamma_j}^2}{Var(\widehat{\gamma_j})}$ using the GWASTools R package [23].

6. Estimate $\hat{\boldsymbol{\tau}}$ by 5 methods:

   a. GEMMA method-of-moments with 28 disjoint categories

   b. LD score regression (LDSC) with *K*=29 strictly defined categories;

   c. LDSC with *K*=53 extended categories;

   d. LD-MLE with *K*=29 core categories;

   e. LD-MLE, with *K*=53 extended categories.

For each scenario, I simulated 500 random phenotype replicates, and used the GWAS results from these replicates as input for the comparison methods. After encountering issues with identifiability of the MLE estimator at saddle points of the likelihood function, I performed an additional 2,000 replicates to increase the effective sample size for the MLE methods.

3.3. Results

To assess the sensitivity of maximum likelihood estimation to the initial values used by the BFGS algorithm, I plotted the maximized value of the likelihood function across all 100 repetitions of optimization. Representative results from one of the simulation replicates of scenario 1 are shown in **Figure 3.1**. The maximum likelihood estimates are largely independent of the initial values provided to the optimization algorithm, with the symmetry in $\pm\theta_k$ for the transformed parameters (defined in section 3.2.1) visible for several of the categories. In other cases, such as the Enhancer_Andersson category, all estimates were positive, and in other annotations the estimates are indistinguishable from zero.

For legibility of presentation in describing the point estimates of the extremely small quantities τ_k, I report results for the estimation of $N * \tau_k = \theta_k^2$ (*k=1, ..., K*), with $N = 10{,}000$ in the simulations. **Figure 3.2** (top panel) plots the estimates of $N * \tau_k$ across replicates of simulation scenario 1, from the model with only the 28 non-zero categories and the base annotation containing all SNPs included. The bias and root mean squared error (RMSE) of LDSC and MLE estimators of the 28 non-zero variance components from both scenarios are reported in Appendix B.

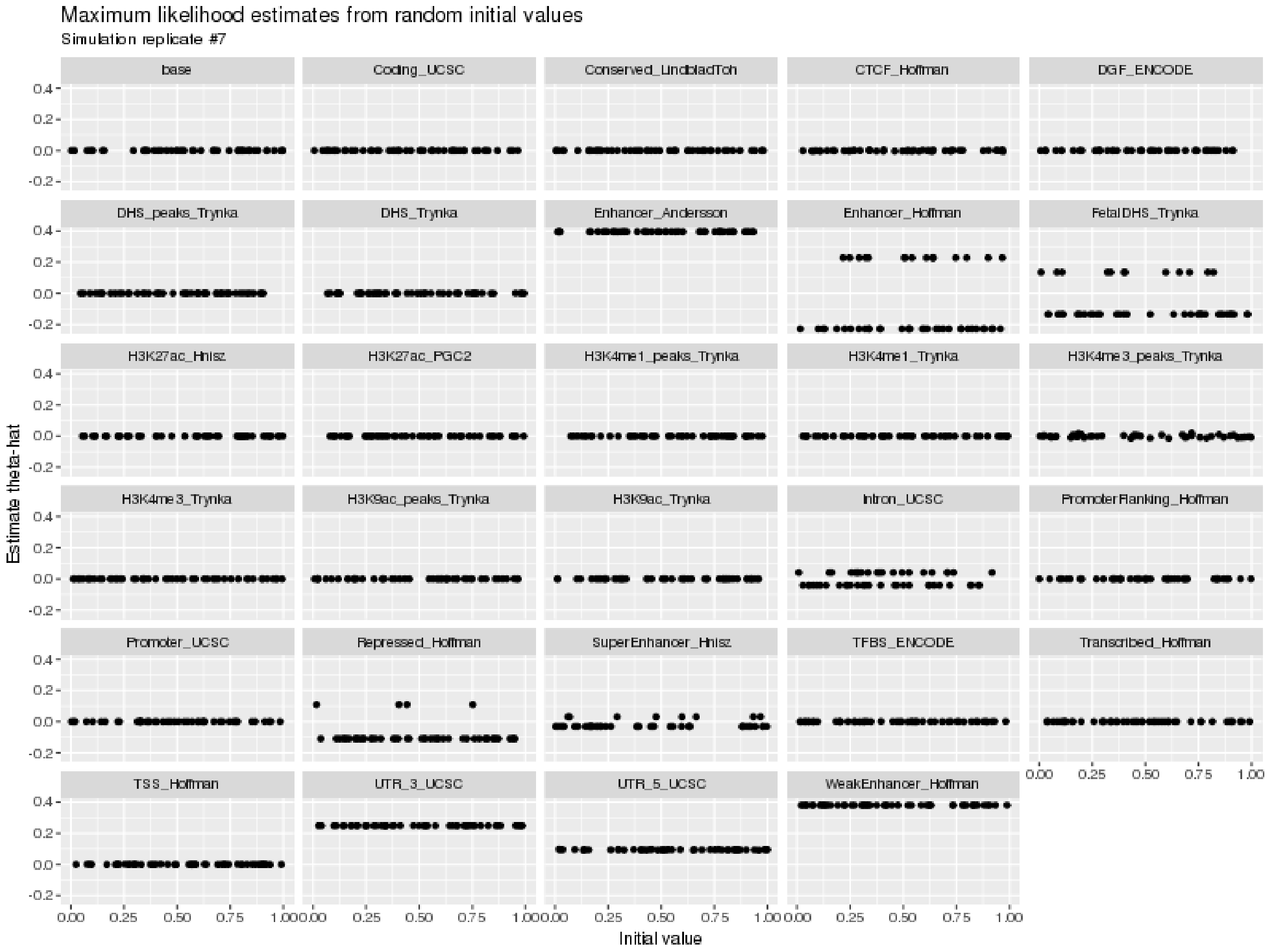

Figure 3.1. Maximum likelihood estimates of all simulated non-zero θ_k across 100 random initializations of the likelihood maximization algorithm for replicate number 7 under simulation scenario 1.

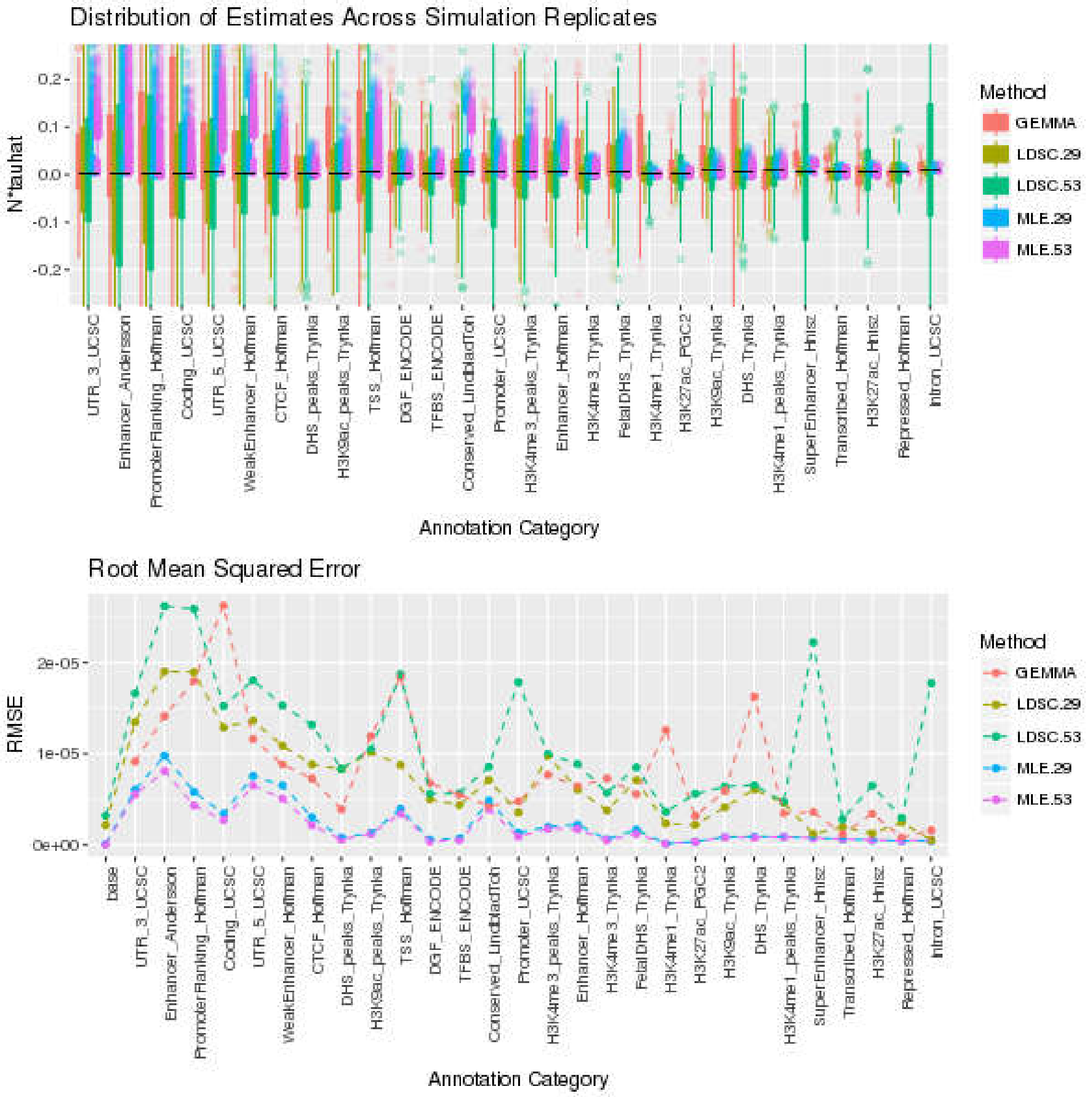

Figure 3.2. Distribution plots of estimates, bias, and root mean squared error (RMSE) of estimates of $N * \tau_k$ for categories with non-zero simulated parameter value, and base category including all SNPs. Compared estimators are method of moments implemented in GEMMA, LD score regression with 29 annotations (LDSC.29) and with 53 annotations (LDSC.53), and maximum likelihood estimation with 29 annotations (MLE.29) and 53 annotations (MLE.53). The horizontal black lines in the top panel indicate true parameter values. Categories are ordered left-right from least to greatest heritability.

The bias of the MLE method was similar to that of the LDSC method (**Figure 3.2**, middle panel, and **Table B.1.** in **Appendix B**). When the model was correctly specified, with only the 28 truly enriched annotations and the base annotation including all SNPs, the average bias of MLE estimates was less than that from LDSC for 16 of the 28 enriched categories, but when the redundant extended-500 categories were included in the model the average bias of LDSC estimates was lower for 16 of the 28 annotations. The GEMMA estimators had higher average bias than the correctly specified MLE in 26 of the 28 categories. This may be partly due to the fact that I recoded the annotations for GEMMA to be non-overlapping, as that method requires.

Figure 3.2, bottom panel, summarizes the RMSE of the compared estimators across categories. There is a consistent trend of lower bias and RMSE for larger categories with more SNPs on the right side of the figure, across all estimation methods. The point estimates from MLE are less variable than LD score regression or GEMMA, regardless of whether the extraneous extend-500 categories are included in the model. All methods showed a trend of greater estimation error (measured as RMSE) for annotation categories containing fewer SNPs. Though the bias of MLE estimates is almost entirely positive, this asymmetry is compensated by the lower variability.

3.3.1. Standard Errors and Confidence Intervals

The confidence intervals derived from permutation-based standard errors provided by the GEMMA and LDSC software packages cover the true parameter value at close to the nominal level.

However, the confidence intervals of the maximum likelihood estimates derived from the observed Fisher Information matrix $I\big(\hat{\theta}\big) = -H(\hat{\theta})$ show dramatically deflated coverage proportions. Upon closer inspection, I found that a substantial proportion of simulation replicates had negative eigenvectors of the Fisher Information matrix. This suggests that the maximum likelihood estimate $\hat{\theta}$ is located at a *saddle point* of the likelihood function, rather than a maximum value. Figure 3.3 shows the proportion out of 2,500 simulation replicates for which each annotation category obtained a negative estimate of $I(\hat{\theta})$. These results are included in the summary of point estimates in Figure 3.2, but the parameters corresponding to negative entries in $I(\hat{\theta})$ are excluded from the assessment of confidence intervals in Figure 3.4 because no standard errors exist from which to calculate confidence intervals.

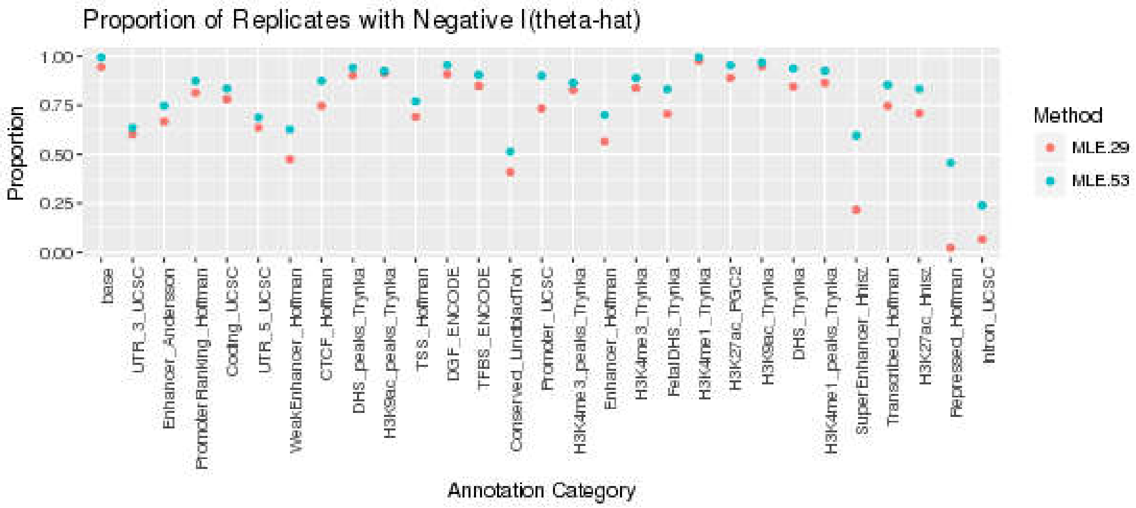

Figure 3.3 Proportion of MLE with negative diagonal elements of $I(\hat{\boldsymbol{\theta}})$, from least to greatest proportion of heritability per annotation category.

Saddle points may be characterized as positions in parameter space with zero gradient of the likelihood function, which are neither maxima nor minima. Inspecting the gradient function, derived in section 3.2.2, we observe that $\frac{\partial}{\partial \theta_p} \ell(W, \theta) = 0$ whenever $\theta_p = 0$, as a consequence of the transformation $\theta_p^2 = N\tau_p$ taken to enforce non-negativity of the variance component estimates. Because of symmetry about the axes of the parameter space, these saddle points are not isolated but orthogonal hyperplanes through the origin.

The asymptotic theory of maximum likelihood estimation is not applicable at these points. Therefore, we consider standard errors in those parameter estimates with positive values along the diagonal of the Fisher Information matrix. These results are shown in **Figure 3.4.**

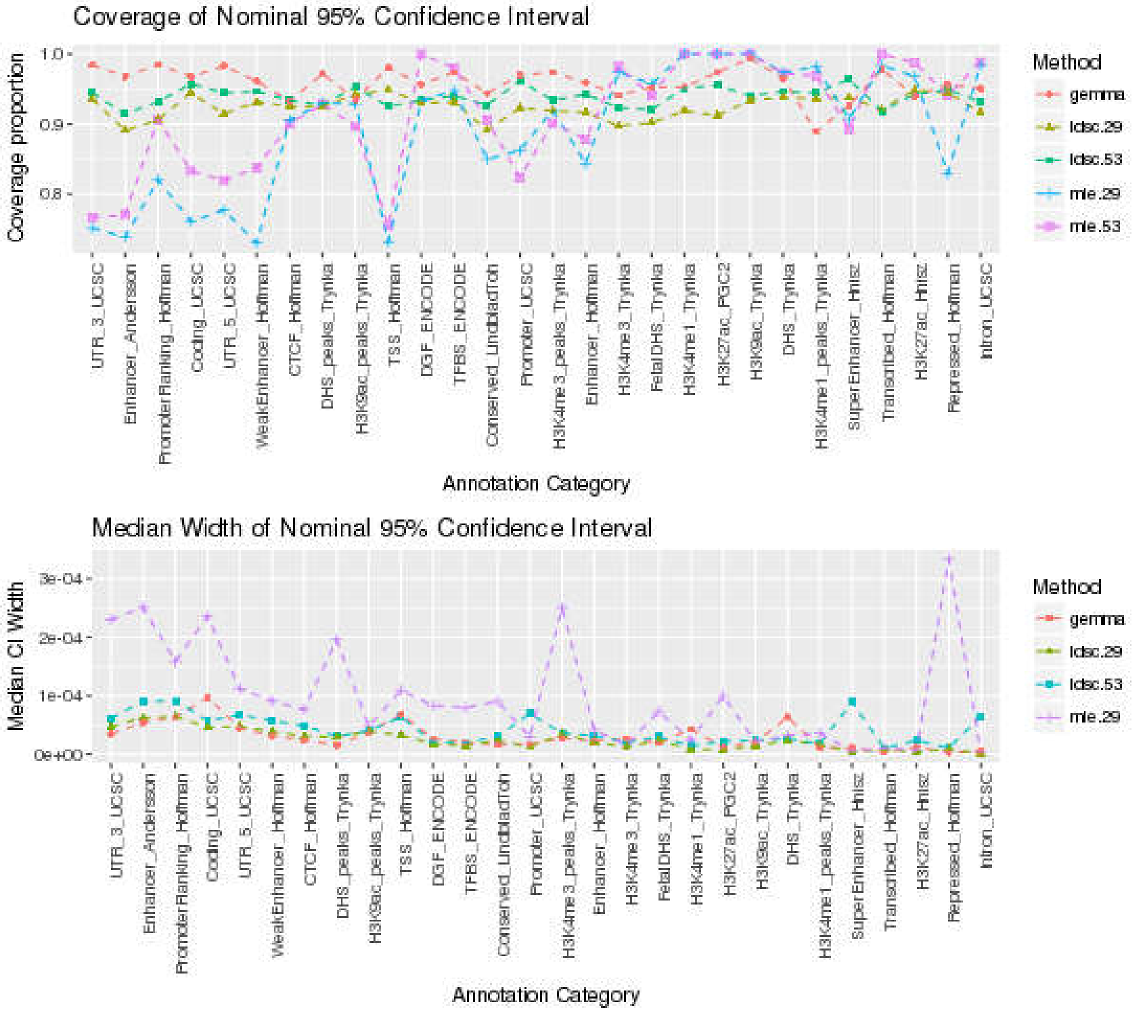

Figure 3.4. Coverage proportions and median width of confidence intervals, for estimates with positive diagonal entries in the Fisher Information matrix.

3.4. Discussion and Conclusions

I have presented an approach for estimating partitioned heritability of complex traits from GWAS summary statistics via a maximum likelihood approach. This approach depends upon modeling the distribution of univariate Wald test statistics as a function of stratified LD, which quantifies the extent to which a given SNP tags variation in annotation categories under consideration. In contrast to the LD score regression method, I derive the joint likelihood of GWAS test statistics for common variants, and maximize this function to obtain estimates of the heritability enrichment parameters.

Both of these methods rely upon simplifying assumptions about the underlying genetic model. In particular, the LD score framework assumes that all true causal variants are used for estimation of the LD scores, which are then used as independent variables in the regression or MLE analysis. The distribution of summary test statistics also depends upon an additive genetic model with no gene-gene or gene-environment interactions. The presented likelihood function will be misspecified in the presence of such more complex genetic effects. While it would theoretically be possible to model non-additive or interaction effects, the computational burden would increase, and the number of possible alternative models is so large that selecting interactions, or recessive or dominant effects for consideration would require stronger hypotheses about the underlying genetic model.

I simulated two scenarios: with 29 and 2 enriched functional categories, and analyzed each with LD score regression and MLE using both 29 core categories, and 53 categories with overlapping extended categories added. When the redundant categories were included in the model, a higher proportion of the core annotations with non-zero simulated enrichment were set to zero at saddle points of the likelihood function. The asymptotic normal distribution of the maximum likelihood estimator does not apply in this case, so we cannot obtain theoretical standard errors for those parameter estimates that were optimized at (or within computational tolerance of) zero.

A practical limitation for the applicability of the MLE approach to partitioning heritability lies in the method of optimizing the likelihood function. Reparametrizing to obtain non-negative estimates exchanged the problem of maximum likelihood estimates on the boundary of the parameter space for the problem of symmetry in the new parameter θ. In fact, the K-fold symmetry (where K is the number of annotation categories) induced by reparametrization causes the model not to be uniquely identifiable, in the sense that different parameter values generate the same likelihood function: $\ell(W;\theta) = \ell(W;-\theta)$. This violates one of the most basic conditions necessary for the consistency of the maximum likelihood estimator, because there is not a single true parameter value in the transformed space, but as many as 2^K equivalent parameter vectors, corresponding to the positive and negative square roots of the original parameter $\boldsymbol{\tau}$.

However, if we consider this property of the proposed estimator to be removing from the model those parameters with negative entries in $I(\hat{\theta})$, this may be interpreted as an approach for model selection, where confidence intervals are less relevant.

The burden of optimizing the reparametrized likelihood increases with the number of SNPs in the analysis, as it requires calculating the gradient of the joint likelihood at every iteration of the algorithm. Extension of this approach to genome-wide summary statistics may be facilitated by stochastic optimization algorithms which evaluate the gradient at a subsample of observations at each iteration [16, 3, 38]. This is a potential direction for further work.

CHAPTER 4. FEATURE SELECTION FOR TISSUE-SPECIFIC ANNOTATION

4.1. Introduction

Biological mechanisms regulating gene expression vary across tissues and cell types. For example, histone modifications are epigenetic variation that affect the accessibility of DNA for transcription, as well as the recruitment of RNA and other molecules involved in gene expression. The exact genomic locations of these histone modifications are known to vary across cell types, with evidence suggesting a role in cellular differentiation across tissues and organ systems within the body [51]. In Chapter 3 I investigated estimation of partitioned heritability by genomic annotations defined by aggregation across multiple tissue and cell types. For regulatory elements such as histone modifications that are activated in a tissue-specific manner, this aggregation was part of the process of dichotomizing the annotation categories published with the stratified LD score regression software and used in Chapter 3 [20]. However, this aggregation may obscure regulatory mechanisms that affect a trait of interest, but are only observed in certain types of cells.

Typically, there is little or no prior knowledge about the genetic architecture of the trait of interest; so we expect that the true mechanism is unknown. Because the large number of functional annotation categories exceeds the capability of existing methods for integrative analysis, a principled approach to selecting relevant annotation categories would improve the performance of the procedures for the partitioning of heritability described in Chapter 3, and integrative fine mapping as presented in Chapter 2. Problems where the desired outcome is an optimal set of predictor variables for a given outcome are known as feature selection, and various methods have been developed to address these problems.

In this chapter, I present a novel approach to the selection of functional annotation categories for post-GWAS integrative analysis, and compare three specific implementations of the general approach. These are each based on criteria defined in terms mutual information (MI) and conditional mutual information (CMI), two quantities developed from the field of information theory to represent the degree of dependence between variables in a non-parametric manner. I will briefly compare several feature selection methods that make use of these quantities in various ways. I then compare results from the minimum redundancy, maximum relevance (mRMR) and conditional feature selection algorithms in tissue-specific histone mark annotations in 100 tissues, classified into 10 tissue groups. I apply these methods to selection of these tissue-specific annotations in relation to GWAS summary statistics from the Genetic Investigation of ANthropometric Traits (GIANT) consortium meta-analysis of BMI in populations of European ancestry [37]. These methods aim to facilitate more focused examination of tissue-specific annotations to enable a better path to understanding the biology of GWAS results.

4.2. Methods

In this chapter, I present the application of three methods based on MI and CMI criteria to the problem of selecting tissue-specific histone mark annotations for subsequent integration with GWAS results. These methods seek a balance between comprehensive inclusion of features that are informative for the outcome, and parsimonious selection of non-redundant features. They are the minimum redundancy, maximum relevance (mRMR) method of Peng *et al* [17, 44], the joint mutual information (JMI) method of Yang and Moody [55], and the general form of conditional mutual information method ("cond") described by Brown et al [8]. To my knowledge, these methods have not been used in the context of genomic functional annotation or GWAS, though the mRMR and JMI approaches have been applied to studies of gene expression for classification of cancer subtypes [17]. The goal in that example was to identify a subset of genes whose expression levels are highly informative of cancer subtype, but which are as independent of each other as possible.

Each of these methods follows a similar strategy for selecting an optimal set of features by means of an iterative stepwise process that selects one additional feature at each iteration by maximizing a scoring function over the set of features not yet selected. Formally, if Y is the outcome of interest, and $\Xi = \{X_1, \dots, X_T\}$ is the set of all available annotation categories (or, more generally, candidate features), then the first step selects $S_1 = \{X_{(1)} \in \Xi\}$ as the feature that is most informative for the outcome (by the definition of MI in section 4.2.1). Then, at each subsequent step of the iteration, given a set of selected features $S_t = \{X_{(1)}, X_{(2)}, \dots, X_{(t)}\}$, select the next feature $X_{(t+1)} = X_j \in \Xi \backslash S_t$ that maximizes the scoring function $J_*(X_j, S_t, Y)$. The specific form of the scoring functions is what distinguishes the three methods. All three scoring functions are defined in terms of the MI and CMI between the outcome, the previously selected features, and the remaining features which are considered as candidates for selection.

4.2.1. Definitions

Mutual information (MI) is a measure of dependence between two random variables, defined from the joint and marginal probability densities of two variables (which may be scalar or vector-valued) as:

$$I(X;Y) = \int_y \int_x p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy$$

This quantity will be zero when *X* and *Y* are independent, so that their joint and marginal pdfs will be related by the equation $p(x,y) = p(x)p(y)$. The MI is more general than a correlation coefficient, requiring no distributional assumptions about either variable or the form of their relationship, as long as the joint and marginal distributions are known or estimated empirically. The MI may be interpreted as an answer to the question: "how much does knowing the value of *X* reduce the uncertainty regarding *Y?"*

The conditional MI (CMI), also useful for feature selection using iterative search procedures, is defined as

$$I(X_1;X_2|Y) = \int_Y p(y)\int_{x_2}\int_{x_1} p(x_1,x_2|y)\log\left(\frac{p(x_1,x_2|y)}{p(x_1|y)p(x_2|y)}\right)dx_1dx_2dy$$

This quantity answers the question: "If we already know the value of *Y,* how much does knowing X_1 reduce uncertainty regarding X_2?"

4.2.2. Feature Scoring Functions

The score functions for each of the three methods I compared in this project are as follows, with $|S_m|$ representing the number of features in the previously selected set:

1. The Minimum Redundancy Maximum Relevance (mRMR) criterion [44] was developed as a greedy, iterative algorithm to optimize the Max-Dependency criterion over all subsets S of the set of candidate features:

$$D(S,y) = I(\{x_1, \dots, x_m\};y)$$

Peng et al show that the greatest possible increase at each iteration, conditional on the previously selected set, is obtained by maximizing the scoring function:

$$J_{mrmr}(X_k) = I(X_k; Y) - \frac{1}{|S_m|} \sum_{j \in S_m} I(X_k; X_j)$$

The first term, $I(X_k; Y)$ represents the *relevance* of feature X_k for the outcome Y, and the penalty term $\frac{1}{|S_m|} \sum_{j \in S_m} I(X_k; X_j)$ represents the *redundancy* of X_k given the previously selected features $X_j \in S_m$. This penalty term will be minimized for the candidate feature that is most independent of those already selected.

2. The Joint Mutual Information (JMI) criterion is based on the intuition that the desired property of a feature set is that all selected features are collectively predictive of the outcome [55]. To quantify this property, the definition of the scoring function is:

$$J_{jmi}(X_k) = \sum_{j \in S_m} I\big((X_k, X_j); Y\big)$$

By the identity $I\big((A, B); C\big) = I(A; C\,|B) + (B; C)$ [5] this is equivalent to

$$J_{jmi}(X_k) = I(X_k; Y) - \frac{1}{|S_m|} \sum_{j \in S_m} \big[I\big(X_k; X_j\big) - I(X_k; X_j|Y)\big]$$

The additional term $I(X_k; X_j|Y)$ is interpreted as measuring the extent to which the information X_k contains regarding Y is *complementary* to the information contained in the previously selected features.

3. The conditional mutual information criterion ("cond", not to be confused with the conditional mutual information statistic "CMI") allows for higher-order interactions.

$$J_{cond}(X_k) = I(X_k; Y) - I(X_k; S_m) + I(X_k; S_m|Y)$$

The last 2 terms are multi-dimensional integrals, which are more computationally demanding. These also depend on the full previously selected set S_m, and therefore must be computed at each step of the iterative selection process. In theory, this criterion considers the possibility of higher-order interactions between the features, as would occur with pairwise mutually independent variables that are not jointly independent.

4.2.3. Estimation Algorithm

The definition of MI as a function of the joint and marginal densities makes no distributional assumptions on the outcome or feature variables. However, this generality contributes to the challenge of estimation. For multivariate normal distributions, MI is an exact function of the covariance.

In this project, I use a nonparametric estimator, which approximates the empirical probability density in a neighborhood around each observed point as a function of the distance to its *k*th-nearest neighbor, given a well-defined metric on the space of $(X_1, \ldots, X_p, Y)$ [35]. Both the order k and the choice of metric norm on the joint feature-outcome space are tuning parameters of the algorithm. This method builds upon the estimation of the Shannon entropy of a (possibly multidimensional) random variable *X*, defined as $H(X) = -\int_X p(x) \log p(x) dx$ [49]. The *k*th-nearest neighbor algorithm is based on the probability distribution $P_k(\epsilon)$ of the distance ϵ from a given data point to its *k*th-nearest neighbor. This is calculated as

$$P_k(\epsilon)d\epsilon = \frac{(N-1)!}{(k-1)!\,(N-k-1)!} \frac{dq_i(\epsilon)}{d\epsilon} d\epsilon \times q_i^{k-1} \times (1-q_i)^{N-k-1}$$

Here, $q_i(\epsilon)$ is the probability mass of a ball of radius ϵ centered at the i^{th} data point. Integrating the density above yields

$$E(\log q_i) = \int_0^{\infty} \log q_i\,(\epsilon) P_k(\epsilon) d\epsilon = \psi(k) - \psi(N)$$

with $\psi(x) = \frac{d}{dx}\log(\Gamma(x))$ representing the digamma function. Then, by assuming constant probability density of *X* within the ϵ ball, so that $q_i(\epsilon) \approx c_d \epsilon^d p(x_i)$ where *d* is the dimensionality of *X,* and c_d is the volume of the ball given the choice of metric used, and combining these equations to obtain

$$\widehat{H}(X) = \psi(N) - \psi(k) + \frac{d}{N}\sum_{i=1}^{N} \log \epsilon(i)$$

Extending this concept to the estimation of mutual information hinges upon the identity $I(X, Y) = H(X) + H(Y) - H(X, Y)$. However, in order to avoid accumulation of bias due to using different length scales ϵ for each of the three components of this sum, the marginal entropy estimations are conditioned upon the distance to the kth-nearest neighbor in the complete (X, Y) space. Specifically, the quantity $n_x(i)$ is defined as the number of points x_j whose distance from x_i is less than $\epsilon(i)/2$, regardless of their location in the Y dimensions, and $n_y(i)$ defined analogously. Then the MI estimate is given by

$$I_{kNN}(X, Y) = \psi(N) + \psi(k) - \frac{1}{N}\sum_{i=1}^{N}\big(\psi(n_x(i) + 1) + \psi(n_y(i) + 1)\big)$$

This algorithm is implemented in the Non-parametric Entropy Estimation Tooolbox (NPEET) Python library [52]. When the number of observations (here, SNPs) is large, the computational bottleneck of these algorithms is identifying nearest neighbors in the feature space. In this situation, the software documentation recommends a subsampling strategy of estimating MI as the average of estimates from small random samples. I performed estimation within each chromosome separately, and then calculated a weighted average of these estimates with weights given by the number of SNPs per chromosome.

4.2.4. Materials

To assess relative performance of the feature selection methods described in section 4.2.2, I analyzed the simulated GWAS summary statistics described in section 3.2.6 of the previous chapter. The independent variables of the simulation are the stratified LD scores for 53 non-tissue-specific functional annotations analyzed in the paper on partitioning heritability with LD score regression, and available for download from the investigators' website [20, 1]. To briefly summarize the simulations, I simulated a sample of GWAS genotypes on chromosome 19 from the 1000 Genomes Project reference panel of European ancestry using HAPGEN2 software [14, 50]. Individual SNP effects were drawn from mean-zero Gaussian random effect distributions with variance dependent upon the SNP-level functional annotation and specified parameter values fixed for each simulation scenario. Both scenarios are specified so that there is total trait heritability of $h^2 = 0.25$ on chromosome 21. In the first scenario, heritability is enriched in 28 annotation categories, the sizes and simulated heritability of which are reported in Table 4.2. An additional 24 categories are defined with 500bp extension beyond 24 of the truly enriched categories, to examine the performance of the selection method when features have large degree of overlap. Four of the annotations, encoding DNAse hypersensitive sites, H1K4me1, H1K5me3 and H3K9ac1 histone mark data included annotations of even more strictly defined “peaks”. These were considered as distinct annotations with additional heritability enrichment on top of that of the non-peak annotations, however these peak annotations do not have corresponding “extend.500” annotations. In the second scenario, only two features are truly enriched. One of the enriched categories (H3K4me1_Trynka) contains 55,649 SNPs out of 129,155 with observed annotation. The other truly enriched category (Enhancer_Andersson) is much smaller with only 822 SNPs, and substantial overlap with the larger, enriched H3K4me1_Trynka annotation.

The tissue-specific histone marks used in the real data analysis of BMI contain dichotomized peaks from assays measuring H3K4me1 and H3K4me3 methylation, and H3K9ac and H3K27ac acetylation in 100 diverse tissue samples, though not all histone marks are observed in all tissues. These data were generated as part of the Roadmap Epigenomics project [15], and post-processed by Trynka et al [51] and Hnisz et al [27]. LD scores for each of these annotations in populations of European ancestry were calculated by Finucane et al [20]. These category-specific LD scores are the features of interest. The tissues are organized into ten groups corresponding to functional organ systems as adrenal, cardiovascular, central nervous system (CNS), bone, gastrointestinal (GI), immune, kidney, liver, muscle, and other. The number of tissues and specific annotations in each group are shown in **Table 4.1** and a list of the specific tissues and histone marks are located in **Appendix C**. Genome-wide summary statistics of SNP-BMI association are the dependent variable in the real data analysis. These are obtained from the GIANT consortium meta-analysis results in the stratum of European ancestry only [37]. A total of 968,740 SNPs have both annotation and GWAS statistics available.

Group	**Adrenal**	**Cardio**	**CNS**	**Bone**	**GI**
Tissues	3	7	10	3	16
Histone Marks	10	15	34	4	44
Group	**Immune**	**Kidney**	**Liver**	**Muscle**	**Other**
Tissues	41	2	3	11	4
Histone Marks	67	5	6	10	25

Table 4.1. Number of tissues and tissue-specific histone mark annotations per group.

4.3. Results

4.3.1. Feature Selection in Simulated Data

The application of these methods to simulated data considers two distinct questions regarding the performance for selecting sets of functional annotations with enriched heritability for a trait of interest. First, to see how well each of the methods was able to distinguish between annotations with a high degree of overlap, I compared the selection probabilities for the strict core annotation categories to their corresponding extended annotations. Second, to assess the impact of the magnitude of heritability enrichment, I compared the simulated enrichment coefficients to the order of feature selection in the stepwise procedure.

Feature selection results in the simulated data are reported in **Table 4.2.** The proportion of simulated replicates in which each feature was selected are shown for the mRMR and JMI selection criteria. Neither selection criterion was able to distinguish between the signal in the truly enriched, narrowly defined core annotation categories and the corresponding extended categories. In simulation scenario 1, I considered 24 pairs of overlapping strict and extended annotation categories, eleven of which (45.8%) had higher selection proportion for the extended category than the strict one. When I examined the effect of simulated trait heritability within each annotation, counting the extended categories as tagging the heritability in their corresponding restricted annotation, it became apparent that categories responsible for a greater proportion of (simulated) trait heritability were selected more often, regardless of whether they contained the unnecessary extension regions. This may be seen in Table 4.2, where the JMI selection rate of the extended annotations, reported in the rightmost column, is higher in rows where the strict annotation it contains has higher simulated total heritability (h_k^2, third column).

			Strict annotations			Extended annotations		
	τ_k	h_k^2	# SNPs	p_{select} mRMR	p_{select} JMI	# SNPs	p_{select} mRMR	p_{select} JMI
Scenario 1								
Intron_UCSC	8.3E-07	0.0399	47763	0.006	0.869	49052	0.006	0.976
Repressed_Hoffman	5.3E-07	0.0317	59924	0.002	1	93092	0.008	1
H3K27ac_Hnisz	5.5E-07	0.0302	55210	0.143	1	59040	0.217	1
Transcribed_Hoffman	6.6E-07	0.0262	39649	0.002	1	92792	0.033	1
SuperEnhancer_Hnisz	7.7E-07	0.0205	26690	0	0.181	27124	0	0.334
H3K4me1_peaks_Trn	8.1E-07	0.0183	22418	0	0.002	--	--	--
DHS_Trynka	7.3E-07	0.0161	22099	0	0.002	64211	0.017	1
H3K9ac_Trynka	8.2E-07	0.0145	17766	0	0	31900	0.064	0.785
H3K27ac_PGC2	3.2E-07	0.0117	37171	0.01	0.874	46211	0.041	1
H3K4me1_Trynka	1.0E-07	0.0057	55649	0.105	1	77594	0.194	1
FetalDHS_Trynka	4.8E-07	0.0055	11513	0	0	37854	0	0.838
H3K4me3_Trynka	2.9E-07	0.0053	18168	0	0	34945	0.002	0.867
Enhancer_Hoffman	4.2E-07	0.0036	8623	0	0	20664	0	0.002
H3K4me3_peaks_Trn	6.1E-07	0.0035	5711	0	0	--	--	--
Promoter_UCSC	6.7E-07	0.0033	4849	0	0	6000	0	0
Conserved_Lindblad	7.3E-07	0.0026	3568	0	0	46664	0.002	1
TFBS_ENCODE	1.4E-07	0.0024	17935	0	0	44867	0.004	0.993
DGF_ENCODE	1.2E-07	0.0023	18790	0	0	69172	0.058	1
TSS_Hoffman	6.4E-07	0.0015	2336	0	0	4509	0	0

H3K9ac_peaks_Trynk	2.1E-07	0.0011	5277	0	0	--	--	--
DHS_peaks_Trynka	7.0E-08	0.001	14724	0	0	--	--	--
CTCF_Hoffman	3.1E-07	0.0009	2910	0	0	8789	0	0
WeakEnhancer_Hoff	2.8E-07	0.0007	2603	0	0	10578	0	0
UTR_5_UCSC	5.4E-07	0.0005	829	0	0	4175	0	0
Coding_UCSC	1.6E-07	0.0003	1938	0	0	8630	0	0
PromoterFlanking_Hof	2.5E-07	0.0002	856	0	0	3772	0	0
Enhancer_Andersson	2.2E-07	0.0002	822	0	0	3352	0	0
UTR_3_UCSC	1.0E-07	0.0001	1419	0	0	3740	0	0
base	0	0	129150	0.085	1	--	--	--
Scenario 2								
Enhancer_Andersson1	4.4E-06	0.0036	822	0	0	3352	0	0
H3K4me1_Trynka1	4.4E-06	0.2464	55649	0.174	1	77594	0.171	1

Table 4.2. Feature selection results for the annotation categories with non-zero simulated enrichment in simulations scenarios 1 and 2. In scenario 1, all strict features are enriched with varying enrichment coefficients τ_k while the extended annotations have no additional enrichment ($\tau_k = 0$), but they have effective enrichment due to tagging of the strict annotation as a subset. Four annotations, defined as peaks of the DHS, H3K4me1, H3K4me3, and H3K9ac reads, do not have corresponding extended categories. Selection rates p_{select} are calculated out of 452 simulated replicates of scenario 1 and 484 replicates of scenario 2.

The extended annotation was selected in substantially more replicates than the strict annotation in the DHS_Trynka, H3K4me3_Trynka, Conserved_LindbladToh, and DGF_ENCODE categories. In all of these, the number of SNPs in the extended category was more than double the number in the strict category, suggesting the possibility of substantial overlap with other enriched annotations.

To address the second question regarding the impact of the magnitude of heritability enrichment per category, **Figure 4.1** shows the mean selection ranking (i.e. order in which it was selected) of each annotation category across 500 simulation replicates plotted against the true simulated heritability per category. The Y axis of these plots shows the mean selection rank, with annotations selected earlier appearing towards the top. The upwards diagonal trend visible in the plots for JMI and cond indicates that these methods select the most strongly enriched annotations early in their stepwise process, while the plot for mRMR shows a substantially weaker relationship.

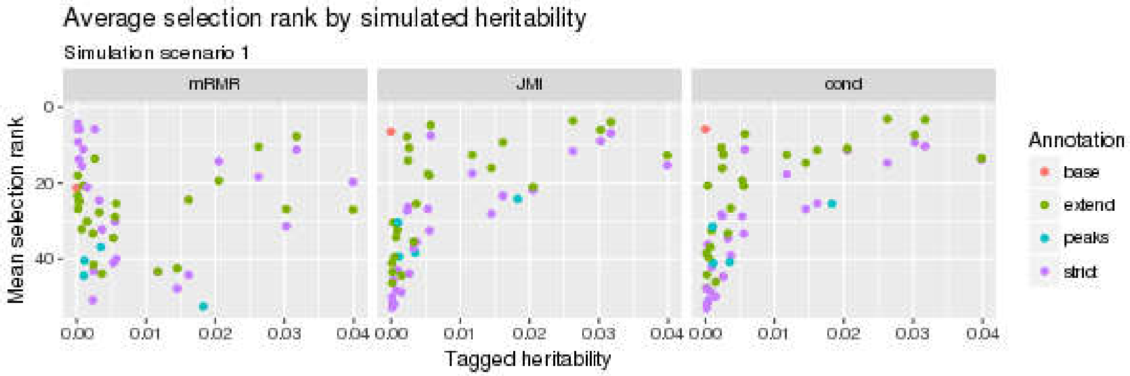

Figure 4.1. Selection rates of 53 strict and extended features based on 500 replicates of simulation scenario 1. Tagged heritability assigns the same true enrichment to the extended categories as the strict core categories inside of them.

I assessed the specificity, or true negative rate of the selection criteria from the results of the simulation scenario 2 analysis. The annotations defined by extending the truly enriched categories were excluded from calculation of the true negative rate. The mRMR criterion led to specificity of 98.7%, while the specificity of the JMI criterion was 63.3%. The JMI method selected more features than the mRMR method, as expected from the fact that the penalty term is reduced by a factor of $I\left(X_k; X_j \middle| Y\right)$ relative to the mRMR scoring function.

For the purpose of selecting features for inclusion in further statistical modeling, it is more important to include all relevant features than to avoid selecting uninformative features. The JMI criterion shows similar performance to the cond method, and is much more computationally efficient. Of these three methods, JMI is the most suitable for this reason.

4.3.2. Distribution of Mutual Information for Tissue-Specific Histone Features

The distribution of MI and CMI in the tissue-specific histone mark annotation and real data BMI summary statistics showed distinct structure due to the correlations in the underlying experiments, where the same histone marks were measured in many related tissue types. **Figure 4.2** shows the relative magnitude of $I(X_k;Y), I(X_k;\ X_j), I(X_k;X_j|Y)$ for histone marks in the 34 histone marks annotations of the central nervous system (CNS) tissue group. The mutual information with the outcome, $I(X_k;Y)$, is shown in the top row and right-most column of the upper heatmap panel. The magnitude of $I(X_k;Y)$ is substantially smaller than $I(X_k;\ X_j)$, represented by the darker-colored center of the plot. The CMI shows a similar pattern as the MI, with smaller magnitude. The block diagonal structure visible in both plots suggests that histone marks of the same type across different tissues have higher similarity than marks of different types in the same tissue.

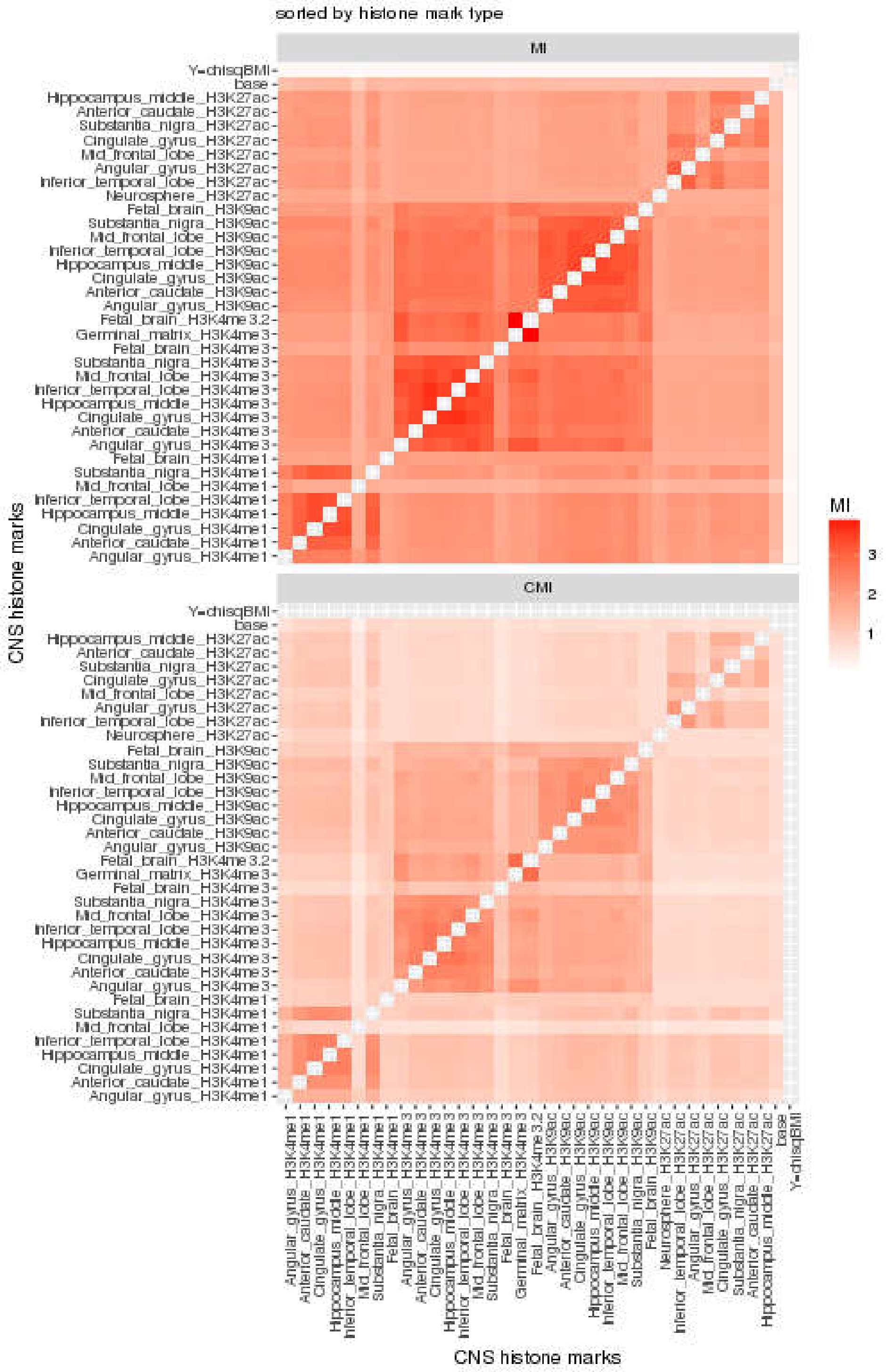

Figure 4.2. Heatmaps of pairwise Mutual Information $I(X_k;X_j)$ and Conditional Mutual Information $I(X_k;X_j|Y)$ among tissue-specific annotations in the central nervous system group sorted by type of histone mark. MI between annotations and Y (GWAS χ^2 statistics for BMI) are shown in the top row and rightmost column of the upper panel.

4.3.3. Comparison of Feature Selection Criteria

Real histone mark data was analyzed by tissue group, as described in **Table 4.1.** **Figure 4.3** shows the sequence of selection scores returned by the mRMR, JMI, and conditional scoring criteria. These plots show the cumulative sum of the maximized selection scoring functions defined in section 4.2.2. The X axis indicates the number of selected features at each step of the iterative search, and as no termination rule is defined, a complete ordering of the candidate features is returned.

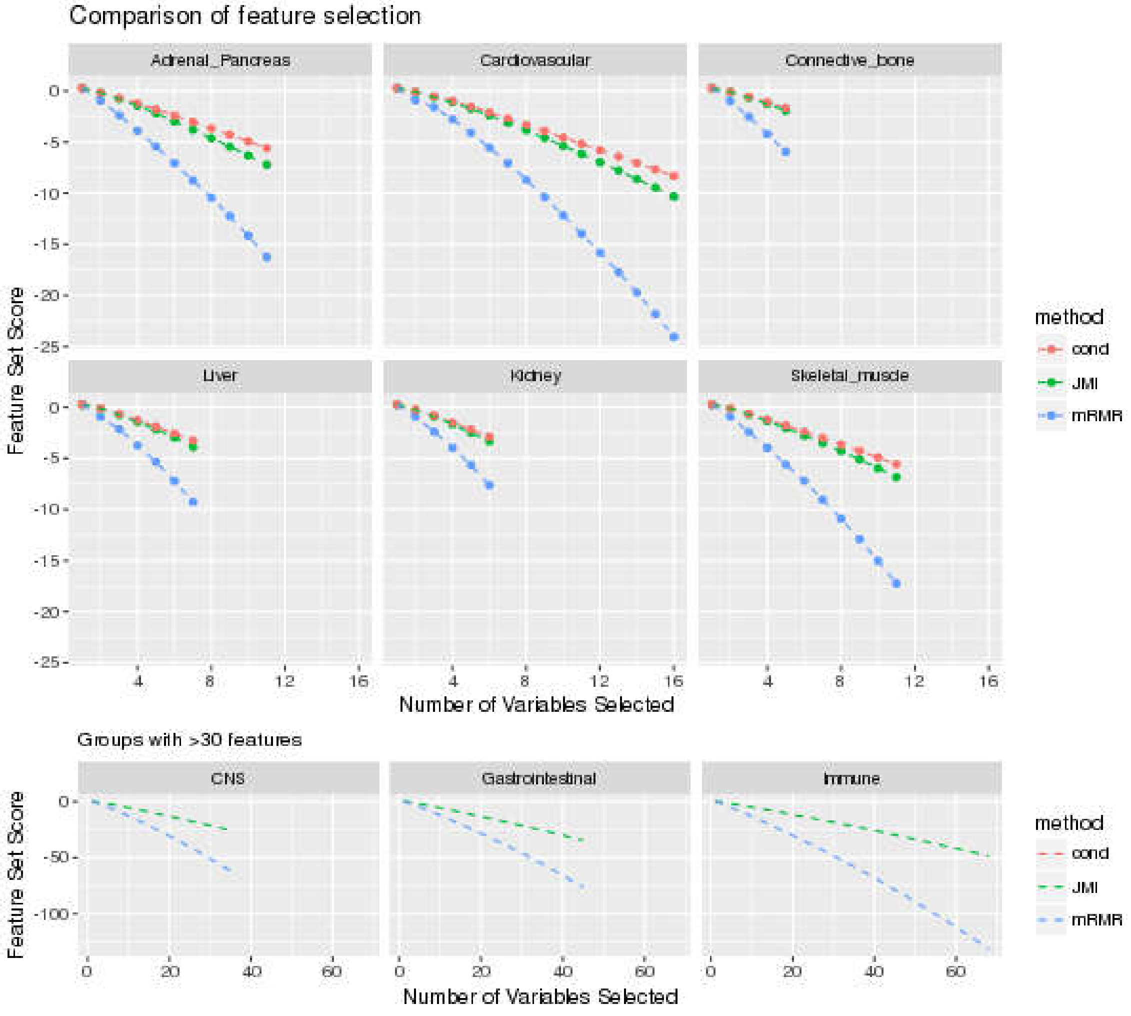

Figure 4.3. Comparison of feature selection criteria in real data from tissue-specific histone marks and BMI summary statistics.

In every group, all three criteria selected only the baseline annotation, containing all SNPs. If we consider that the baseline annotation will represent all residual heritability not attributable to any of these categories, it would make sense for this to be selected first. The downward slope of all lines in **Figure 4.3** indicates that subsequent features considered by the MI selection criteria is due to negative values of the selection scoring functions, indicating that the penalty for redundancy with the previously selected features was greater than the positive mutual information between the annotation-stratified LD scores and the outcome GWAS summary statistics. The conditional method exceeded the available computational resources for the groups containing more than 30 features, so only mRMR and JMI are shown for these (bottom row of **Figure 4.3**).

4.4. Discussion and Conclusions

Based on the analysis of simulated GWAS data, we found the JMI feature selection method to be preferable among the three methods compared. None of the methods was capable of distinguishing the truly enriched core annotations from the corresponding extended annotations containing the same enriched regions with the addition of a surrounding buffer without additional enrichment. However, both the JMI and cond methods showed a pattern of selecting the annotation categories with strongest enrichment early in the stepwise selection process. Of these, JMI is substantially more computationally efficient, because it only depends upon pairwise MI and CMI statistics, whereas the cond criterion requires calculation of the MI and CMI between candidate features and the vector of all previously selected features at each iteration.

In all groups of the real data set of tissue specific histone modifications, all criteria were negative for all features after the first was selected. This result may be interpreted as indicating that these tissue-specific annotation features are more strongly related to each other than they are to the outcome GWAS statistics. This could be a consequence of histone marks in the same genomic locations across tissue types, combined with weak enrichment of heritability within the sets of SNPs defined by these annotations. One potential approach to addressing this issue would be to down-weight the penalty terms in the scoring function, essentially allowing more redundancy in the selected feature set. However, it is unclear how this weighting parameter should be estimated from the data. Alternatively, if there is an *a priori* reason to select a certain number of features, the results obtained by MI-based feature selection may be useful in ranking the features in order to select the most salient for statistical modeling. Recoding or compressing these annotations to isolate the independent signatures of each one may be a promising direction of future work.

Despite the difficulty in determining a useful stopping rule for the stepwise selection procedure in the presence of high overlap among categories, MI-based approaches to feature selection provide a promising avenue for researchers overwhelmed by the numerous tissue-specific functional annotations available. When the purpose of annotation feature selection is to choose annotations for inclusion in a method such as PAINTOR with a fixed upper bound on the number of annotations that can be included, this approach provides ranked ordering of the available annotations in terms of the strength of evidence for their relevance to the trait of interest.

CHAPTER 5. DISCUSSION

This thesis examines three approaches to translate the results of genome-wide association studies into practicably applicable insights into the biological processes underlying complex traits. In all of these approaches, genomic functional annotation provides an additional source of information about the associated genetic variants. Functional annotation classifies SNPs according to their locations in protein coding genes and known regulatory elements, including tissue-specific regulatory elements that are active only in certain cell types, providing opportunities and challenges for the aggregation, attribution, and characterization of genetic effects and trait heritability. Each project presented in this thesis addresses a specific question in genetic epidemiology by synthesizing data from GWAS and functional annotation. The fine mapping project in Chapter 2 regionally investigates genetic effects in terms of identifying individual causal variants with plausible functionality, while the problems of heritability enrichment in Chapters 3 and 4 evaluate genetic effects on the level of annotated SNP sets. In Chapter 3, I proposed a novel method to estimate the coefficients of enrichment for genomic annotation, based on the derived distribution of GWAS summary statistics as a function of annotation and LD structure. In Chapter 4, I explore approaches to select relevant annotations in a model-free, non-parametric framework based on mutual information statistics. These projects provide new tools and methods for researchers to hone in on those genetic variants that are most likely functional for the phenotype of interest.

In this summary chapter, I discuss a few themes that recur in the three projects, how each project demonstrates those themes, and possible directions for future work. One commonality is that all three projects use SNP-trait association statistics as the observed outcome, and SNP-level annotation and LD structure as predictors for the estimation of true genetic effects. The derivation in Chapter 2 of the AnnoRE fine mapping estimator as best linear unbiased predictor for SNP effects shows that the estimate constructed from individual-level genotype and phenotype data may be approximated by GWAS summary statistics and LD structure from an ancestry-matched reference panel. In the LD score regression model framework of Chapters 3 and 4, the units of observation are SNPs, rather than people. This is advantageous for practical reasons, to maximize sample size by facilitating the application of these methods in consortia of study samples seeking to pool their results. The logistical requirements of both informed consent and privacy protection for the study participants, as well as the increased computational burden of storing and analyzing all individuals together has led to an interest in statistical methods that use GWAS summary statistics as input. A second theme in the three projects is the trade-off between investigating focused hypotheses as opposed to a more exploratory approach. I have generally assumed that the investigator does not have specific biological question, such as “are regulatory regions in the Amygdala enriched for association with BMI?” But rather, I have supposed that there is no *a priori* evidence to prefer some specific subset of the available annotations. Both of these themes contribute to another—issues of computational complexity arising from the scale of these data. Even in the simulation studies, where I considered smaller sample regions rather than genome-wide data, the importance of considering algorithmic efficiency was apparent. For example, a major advantage of the AnnoRE fine mapping method over the multi-level modeling approach of

PAINTOR was in avoiding the issues of algorithmic convergence in PAINTOR's EM algorithm, by estimating the annotation-level enrichment separately from genome-wide level data. In Chapter 3, maximum likelihood estimation in the model of the complete distribution of GWAS χ^2 statistics of association introduced, rather than simply modeling their mean values as in stratified LD score regression. To the extent that "fitting" a model entails optimizing a function of some sort, the tractability of this optimization problem may be far from trivial in cases where classical methods break down or require strong distributional assumptions to be applicable. In many cases, these computational challenges have been studied in the computer science or physics literature, each of which has their own standards and conventions for acceptable methodology, which differ from those in statistics or epidemiology. Throughout this thesis, I attempt to describe the high-level strategy of the algorithms used in each project, but in some cases alternative estimation methods may be better suited to the functions of interest. Because algorithmic development is beyond the scope of these projects, I only consider algorithms with up-to-date implementations in R or Python.

The fine mapping project in Chapter 2 addresses these issues by using a highly specific model, conditional on estimates of the heritability enrichment in each annotation category. In the simulation analysis, I supposed that the annotation-level heritability enrichment coefficients were equal to the true parameter values used in simulation, while for the real data analysis I used heritability enrichment calculated with the LD score regression software from the GIANT consortium GWAS summary statistics for BMI in European ancestry populations. In light of the results of Chapter 3, which showed wide variation in the estimates of enriched heritability between simulation replicates with the same true parameter scenarios, incorporating the standard errors of the estimates of heritability enrichment into the fine mapping model may be a promising direction for future work. By treating these heritability estimates as fixed and known values, I was able to obtain closed form expressions for the conditional causal SNP effects, and avoided the necessity of setting an upper bound on the number of causal SNPs per locus. Because the enrichment coefficients are estimated separately from the fine mapping model, further research may be done to evaluate the robustness of the choice of predicted causal variants to varying estimates of these parameters.

The maximum likelihood approach to partitioned heritability estimation presented in Chapter 3 aimed to improve the heritability estimates used to prioritize SNPs in the fine mapping model by building upon the LD score regression method [20]. I derived the likelihood of the non-central χ^2 GWAS statistics of association as a function of stratified LD scores and reparametrized enrichment coefficients $\theta_k^2 = N\tau_k$, to ensure non-negativity of the estimated variance components. This likelihood function proved challenging to optimize, as the symmetry in $\pm\theta_k$ induced non-convexity of the target function and saddle points along each hypersurface with $\theta_k = 0$. Maxima of the likelihood function on those surfaces may be interpreted as removing the corresponding annotation category from the model, but the asymptotic theory of maximum likelihood estimation is not applicable for estimation of standard errors and hypothesis testing at those estimates. In retrospect, if this relatively simple trick for estimating variance components was easy and gave good results, it would be common practice by now. Other transformation that limit estimates to the acceptable parameter space need to be considered in future research. Additionally, scaling up this method from the simulation study on chromosome 21 presented in Chapter 3 to genome-wide applications would require rethinking the algorithmic implementation for maximizing the likelihood function. Stochastic gradient descent methods are a promising direction for further investigation, as this class of algorithms is well suited to high-dimensional non-convex optimization problems [16, 3, 38]. Alternative approaches to reparametrizing the model of observed GWAS statistics, or estimation of standard errors with a block jackknife approach such as that used in LDSC and GEMMA [10, 60] are possible solutions to the problems encountered with obtaining standard errors from the Fisher Information matrix under the asymptotic theory of maximum likelihood estimation.

The mutual information feature selection project in Chapter 4 aims to identify sets of tissue-specific functional annotations most relevant to a given trait of interest, using a model-free non-parametric framework to reduce the dependence upon distributional assumptions. The formulation of the problem is related to Chapter 3 where the independent variables are stratified LD scores for functional annotation and the outcome is GWAS marginal χ^2 statistics, with the distinct goal of identifying an optimal set of annotations for inclusion in more stringently specified parametric models such as those in Chapters 2 and 3. Brown, Pocock, Zhou and Luján decompose the conditional log likelihood as a sum of terms representing the model goodness of fit, the selection of relevant features for modeling, and the residual entropy of the outcome unrelated to the available features [8]. This decomposition suggests an interpretation of model goodness of fit as composed of three distinct components representing the choice of predictive variables, the functional form of model specification, and the true residual error not attributable to the observed candidate predictors. In particular, if an outcome y is related to the available features $\boldsymbol{x}$ with a true generative model given by the conditional probability density $p(y|\boldsymbol{x})$. If γ is a binary vector representing selection of a subset $\boldsymbol{x}_\gamma$ of the available features, then $p(y|\boldsymbol{x}_\gamma)$ represents the distribution of y given that subset, and $p(y|\boldsymbol{x}_\gamma, \theta)$ is a specific model relating these selected features to the outcome of interest. With these definitions, the conditional log likelihood can be decomposed as a finite sample estimator

$$-\ell(\gamma, \theta|\boldsymbol{x}, y) \approx E_{xy}\left\{log\frac{p(y|\boldsymbol{x}_\gamma)}{p(y|\boldsymbol{x}_\gamma, \theta)}\right\} + E_{xy}\left\{log\frac{p(y|\boldsymbol{x})}{p(y|\boldsymbol{x}_\gamma)}\right\} - E_{xy}\{p(y|\boldsymbol{x})\}$$

The middle term in this decomposition, $E_{xy}\left\{log\frac{p(y|\boldsymbol{x})}{p(y|\boldsymbol{x}_\gamma)}\right\}$ represents the divergence between the conditional distributions of y given the selected set of features, and given all features. This term is minimized by the selection of an optimal feature set, as investigated in Chapter 4, whereas the first term represents the estimation of a model relating the features to the outcome, as in Chapter 3. These may be considered as distinct components of statistical modeling, because the estimation of model parameters is performed conditional on a choice of included features, and the question of selecting an optimal feature set may be considered independent of a given model specification. Mutual information approaches have been successfully applied to feature selection problems in gene expression studies, and other situations where the number of available predictors would overwhelm the capability of a given model. When applied to the question of selecting tissue specific functional annotations in Chapter 4, the high degree of overlap between tissue-specific histone mark annotations, relative to the modest enrichment effects of association, led the penalties for redundancy among selected features to outweigh the improvement to the explained variability of the outcome, and no tissue-specific features were selected for relevance to BMI. However, the results of the MI selection algorithm did give an ordering of the candidate features in terms of strength of evidence of relevance to the trait. If we are willing to accept redundancy due to overlapping categories, this approach can estimate an optimal set of annotations when the investigator specifies the desired number of annotations.

These themes of estimation from marginal summary statistics, trade-offs of model complexity or generality, and computational challenges are quite general and certainly not unique to genetic association studies. Yet, any researcher attempting to integrate functional annotation into the interpretation of GWAS must in some way confront them. The methods I have proposed offer solutions to specific questions of causal SNP identification in fine mapping, estimation of partitioned heritability enrichment coefficients, and selection of relevant annotation features. These projects have been presented in the order in which they were performed, but an applied analysis of GWAS results would more naturally proceed in the opposite order: first selecting relevant annotations from the set of all available candidates, then estimating the enrichment coefficients from genome-wide summary statistics, and finally performing fine mapping in loci with genome-wide significant trait associations. This series of analysis steps would identify strong candidate SNPs for follow-up validation *in vitro* or model organism studies. Well validated causal variants explain the observed signal in significant GWAS loci in a way that is both more biologically interpretable in terms of potential mechanism of action (e.g. disruption of a transcription factor binding site) and more clinically relevant as for inclusion in genetic risk score models. Integrating functional annotation with GWAS summary statistics advances our understanding of the genetic foundations of human complex traits.

APPENDIX A. Derivation of Best Linear Unbiased Predictor for Random Effects

Suppose we have specified a linear random effects model for a mean-centered outcome *y* and *M* predictors observed in *N* subjects, where $\boldsymbol{X}$ is the $N \times M$ design matrix:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}\,, \qquad \boldsymbol{\epsilon} \sim N_N\big(0, R(\phi_R)\big)\ , \qquad \boldsymbol{\beta} \sim N_M\big(0, H(\phi_H)\big)$$

The quantities ϕ_R, ϕ_G represent hyperparameters for the residual covariance, and the random effects covariance. If the study sample is independent and identically distributed, then $\phi_R = \sigma_R^2$ and $R = \sigma_R^2 I_N$, but if there is a known correlation structure in the sample, for example in a GWAS of related individuals, then the covariance matrix *R* may be specified accordingly. For the fine mapping model defined in Chapter 2, I let the random effects covariance H be diagonal, with unequal entries defined by the estimated heritability enrichment coefficients (see Section 2.2.1). Conditional on these covariance parameters, the joint probability density of $\boldsymbol{y}, \boldsymbol{\beta}$ may be written as:

$$p(\boldsymbol{y}, \boldsymbol{\beta} | X, H, R) = p(\boldsymbol{y}|\boldsymbol{\beta}, R)p(\boldsymbol{\beta}|H)$$

$$= \frac{1}{\sqrt{2\pi|R|}} \exp\left(\frac{-1}{2}\ (\boldsymbol{y} - \boldsymbol{X\beta})^T R^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})\right) \frac{1}{\sqrt{2\pi|H|}} \exp\left(\frac{-1}{2}\ \boldsymbol{\beta}^T H^{-1} \boldsymbol{\beta}\right)$$

We maximize this function with respect to $\boldsymbol{\beta}$ by considering its logarithm:

$$\log p(\boldsymbol{y}, \boldsymbol{\beta}|X, H, R) = c\ - \frac{1}{2}[\log(|R|) + \log(|H|) + (\boldsymbol{y} - \boldsymbol{X\beta})^T R^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{\beta}^T H^{-1} \boldsymbol{\beta}]$$

To find the maximizing value, set the derivative with respect to $\boldsymbol{\beta}$ equal to zero:

$$0 = \nabla_\beta \log p(\boldsymbol{y}, \boldsymbol{\beta}|X, H, R) = \boldsymbol{X}^T R^{-1} \boldsymbol{y}\ - \frac{1}{2} \boldsymbol{X}^T R^{-1} \boldsymbol{X} - \frac{1}{2}\left(\boldsymbol{y}^T R^{-1} \boldsymbol{X}\right)^T + H^{-1} \boldsymbol{\beta}$$

Solving this expression yields:

$$\widetilde{\boldsymbol{\beta}}|\boldsymbol{X}, \boldsymbol{y}, R, H = [\boldsymbol{X}^T R^{-1} \boldsymbol{X} + H^{-1}]^{-1} \boldsymbol{X}^T R^{-1} \boldsymbol{y}$$

APPENDIX B. Partitioned Heritability Estimates

	category size	True	MLE	LDSC	MLE	LDSC	disjoint size	GEMMA
Categories in model			K=53	K=53	K=29	K=29		K=28
Scenario 1								
Enhancer_Andersson	822	2.15E-07	12.80	-11.43	17.85	-17.32	822	19.59
UTR_5_UCSC	829	5.43E-07	4.24	-1.72	5.88	-1.35	829	7.23
PromoterFlanking	856	2.45E-07	4.14	-6.69	6.69	-11.2	813	34.32
UTR_3_UCSC	1419	1.02E-07	24.55	6.69	26.75	12.05	1344	30.33
Coding	1938	1.64E-07	3.76	6.37	6.08	-0.06	705	51.35
TSS	2336	6.40E-07	1.28	0.64	2.04	2.11	1801	9.23
WeakEnhancer	2603	2.79E-07	6.75	3.75	11.96	3.61	2165	12.67
CTCF	2910	3.12E-07	0.92	-0.99	2.93	-2.59	2540	11.12
Conserved	3568	7.26E-07	2.06	-2.27	3.24	-2.81	2574	-0.18
Promoter_UCSC	4849	6.70E-07	-0.69	-0.36	-0.25	0.39	2940	1.18
H3K9ac_peaks	5277	2.13E-07	-0.09	-1.99	-0.04	-4.45	2799	36.29
H3K4me3_peaks	5711	6.13E-07	-0.07	1.77	0.21	2.37	2193	3.82
Enhancer_Hoffman	8623	4.22E-07	0.92	1.07	1.89	0.85	3161	7.85
FetalDHS	11513	4.80E-07	-0.11	2.99	0.59	2.31	4837	5.33
DHS_peaks	14724	6.98E-08	0.46	-25.07	1.10	-26.5	3389	17.67
H3K9ac	17766	8.16E-07	-0.95	0.53	-0.92	0.75	4948	3.1
TFBS_ENCODE	17935	1.36E-07	-0.01	-2.59	0.62	-1.74	4579	8.87
H3K4me3	18168	2.92E-07	-0.53	-0.43	-0.32	-2.87	2741	10.3
DGF_ENCODE	18790	1.23E-07	-0.47	10.79	0.30	8.46	3624	5.05
DHS_Trynka	22099	7.30E-07	-0.86	-0.08	-0.61	1.21	1447	6.56
H3K4me1_peaks	22418	8.15E-07	-0.89	-0.16	-0.79	-0.1	4163	-1.89
SuperEnhancer	26690	7.68E-07	-0.38	-1.40	0.07	-0.04	7493	3.26
H3K27ac_PGC2	37171	3.15E-07	-0.93	0.25	-0.81	-1.84	6285	2.44
Transcribed	39649	6.62E-07	-0.90	-0.01	-0.81	-0.67	15326	0.49
Intron_UCSC	47763	8.35E-07	-0.43	2.84	-0.23	-0.06	9578	0.83
H3K27ac_Hnisz	55210	5.48E-07	-0.84	0.83	-0.73	0.26	3779	-0.95
H3K4me1	55649	1.02E-07	-0.99	-0.52	-0.91	0.59	2345	64.92
Repressed	59924	5.29E-07	-0.33	-0.26	0.51	0.23	26444	-0.55
All	129150	0					NA	
Scenario 2								
Enhancer_Andersson	822	4.43E-06	-0.27	-0.55	0.11	-0.77	822	1.31
H3K4me1_Trynka	55649	4.43E-06	-0.99	-0.03	-0.95	-0.07	2345	2.14

Table B1. Proportional bias of estimates of SNP effect variance components, defined as ratio of bias to true simulated value, for annotation categories with non-zero parameter . Results of LD score regression and MLE analysis are shown from models including only the 28 enriched annotations plus baseline (K=29), and with the addition of 24 "extend-500" annotations (K=53).

	category size	True	MLE	LDSC	MLE	LDSC	disjoint size	GEMMA
Categories in model			K=53	K=53	K=29	K=29		K=28
Scenario 1								
Enhancer_Andersson	822	2.15E-07	37.47	121.68	41.75	88.37	822	65.61
UTR_5_UCSC	829	5.43E-07	11.73	33.26	10.64	25.08	829	21.39
PromoterFlanking	856	2.45E-07	17.99	105.59	21.58	77.32	813	73.4
UTR_3_UCSC	1419	1.02E-07	52.92	162.81	40.6	131.84	1344	89.4
Coding	1938	1.64E-07	16.18	93.11	14.93	78.88	705	160.58
TSS	2336	6.40E-07	5.72	29.26	4.45	13.71	1801	28.76
WeakEnhancer	2603	2.79E-07	17.13	54.86	18.85	39.04	2165	31.66
CTCF	2910	3.12E-07	7.26	42.19	9.69	28.27	2540	23.31
Conserved	3568	7.26E-07	5.26	11.82	6.15	9.77	2574	5.85
Promoter_UCSC	4849	6.70E-07	1.43	26.64	1.99	5.29	2940	7.11
H3K9ac_peaks	5277	2.13E-07	5.03	49.14	3.2	47.82	2799	56.08
H3K4me3_peaks	5711	6.13E-07	3.06	16.25	2.08	15.99	2193	12.55
Enhancer_Hoffman	8623	4.22E-07	4.14	21.05	2.73	14.43	3161	15.22
FetalDHS	11513	4.80E-07	2.64	17.72	2.08	14.8	4837	11.56
DHS_peaks	14724	6.98E-08	7.92	119.57	8.61	119.06	3389	55.78
H3K9ac	17766	8.16E-07	1.00	7.83	0.97	5.07	4948	7.23
TFBS_ENCODE	17935	1.36E-07	3.87	41.66	8.16	32.17	4579	40.31
H3K4me3	18168	2.92E-07	1.73	19.51	1.85	12.93	2741	25.02
DGF_ENCODE	18790	1.23E-07	3.30	45.56	3.83	40.49	3624	55.02
DHS_Trynka	22099	7.30E-07	1.07	8.98	1.01	8.35	1447	22.27
H3K4me1_peaks	22418	8.15E-07	1.06	5.78	0.99	5.62	4163	4.25
SuperEnhancer	26690	7.68E-07	0.98	28.97	0.83	1.55	7493	4.69
H3K27ac_PGC2	37171	3.15E-07	1.02	17.78	1.18	6.96	6285	10.08
Transcribed	39649	6.62E-07	0.96	4.26	0.92	2.95	15326	1.74
Intron_UCSC	47763	8.35E-07	0.65	21.28	0.57	0.62	9578	1.89
H3K27ac_Hnisz	55210	5.48E-07	0.96	11.87	0.89	2.28	3779	6.24
H3K4me1	55649	1.02E-07	1.01	35.31	1.63	22.7	2345	123.14
Repressed	59924	5.29E-07	0.85	5.61	0.76	4.64	26444	1.5
All	129150	0					0	
Scenario 2								
Enhancer_Andersson	822	4.43E-06	1.67	6.02	1.84	4.45	822	3.67
H3K4me1_Trynka	55649	4.43E-06	0.99	0.86	0.92	0.52	2345	3.29

Table B2. Proportional root mean squared error (RMSE) of estimates of SNP effect variance components, defined as ratio of RMSE to true simulated value, for annotation categories with non-zero parameter values. Results of LD score regression and MLE analysis are shown from models including only the 28 enriched annotations plus baseline (K=29), and with the addition of 24 “extend-500” annotations (K=53).

APPENDIX C. Tissue-specific Histone Mark Data

Tissue group	Cell type	H3K4me1	H3k4me3	H3K27ac	H3K9ac
Adrenal/ Pancreas	Fetal adrenal	X	X		
	Pancreas	X	X		
	Pancreatic islets	X	X	X	X
CNS	Angular gyrus	X	X	X	X
	Anterior caudate	X	X	X	X
	Cingulate gyrus	X	X	X	X
	Fetal brain	X	X	X	X
	Germinal matrix		X		
	Hippocampus middle	X	X	X	X
	Inferior temporal lobe	X	X	X	X
	Mid frontal lobe	X	X	X	X
	Neurosphere			X	
	Substantia nigra	X	X	X	X
Cardiovascular	Aorta		X		
	Fetal heart	X	X		X
	Fetal lung	X	X		X
	Left ventricle	X	X		
	Lung	X	X		
	Right atrium	X	X		
	Right ventricle	X	X		
Connective/ Bone	Breast fibroblast primary	X	X		
	Chondrogenic dif			X	
	Osteoblast			X	
	Penis foreskin fibroblast primary	X	X		
Gastrointestinal	Colon smooth muscle	X	X	X	X
	Colonic mucosa	X	X	X	X
	Duodenum mucosa	X	X	X	X
	Duodenum smooth muscle	X	X	X	
	Esophagus	X			
	Fetal large intestine	X	X		
	Fetal stomach	X	X		
	Gastric	X	X		
	Rectal mucosa	X	X	X	X

	Rectal smooth muscle	X	X	X	X
	Sigmoid colon	X	X		
	Small intestine	X	X		
	Stomach mucosa	X	X		X
	Stomach smooth muscle	X	X	X	X
Immune	CD14			X	
	CD14 primary	X	X		
	CD15 primary	X	X		
	CD19			X	
	CD19 primary (BI)	X	X		
	CD19 primary (UW)	X	X		
	CD20			X	
	CD25+ CD127-			X	
	CD25- CD45RA+			X	
	CD25- IL17+ Th17			X	
	CD25- IL17- Th stim			X	
	CD25int CD127+			X	
	CD3 primary			X	
	CD3 primary (BI)	X	X		
	CD3 primary (UW)	X	X		
	CD34 primary	X	X		
	CD4 memory primary	X	X		
	CD4 naïve primary	X	X		
	CD4 primary		X		
	CD4+ CD25+ CD127-	X	X		
	CD4+ CD25- CD45RO+	X	X		
	CD4+ CD25- CD45RA+	X	X		
	CD4+ CD25- IL17+ Th17	X	X		
	CD4+ CD25- IL17- PMA	X	X		
	CD4+ CD25+ Th	X	X		
	CD4+ CD25int CD127+ Tmem	X	X		
	CD56 primary	X	X		
	CD8 memory primary	X	X		

	CD8 naïve primary (BI)	X	X		
	CD8 naïve primary (UCSF-UBC)	X	X		X
	CD8 primary		X		
	Fetal thymus	X	X		
	Mobilized CD34			X	
	Mobilized CD34 primary	X	X		
	Peripheral blood mononuclear primary	X	X		X
	Spleen	X	X		
	Th0			X	
	Th1			X	
	Th2			X	
	Thymus	X			
	Treg primary		X		
Kidney	Kidney	X	X	X	X
Liver	Liver			X	
	Liver (BI)	X	X		X
	Liver (UCSD)	X	X		
Other	Adipose nuclei	X	X	X	X
	Breast luminal epithelial	X			
	Breast myopithelial	X	X		X
	Breast vHMEC	X	X		
	Fetal placenta	X	X		
	Ovary	X	X		
	Penis foreskin keratinocyte	X	X		X
	Penis foreskin melanocyte	X	X		
	Placenta amnion	X	X		
	Placental chorion	X	X		
Skeletal muscle	Fetal leg muscle	X	X		
	Fetal trunk muscle	X	X		
	Psoas muscle	X	X		
	Skeletal muscle	X	X	X	X

APPENDIX D. Comparison of Mutual Information Estimators

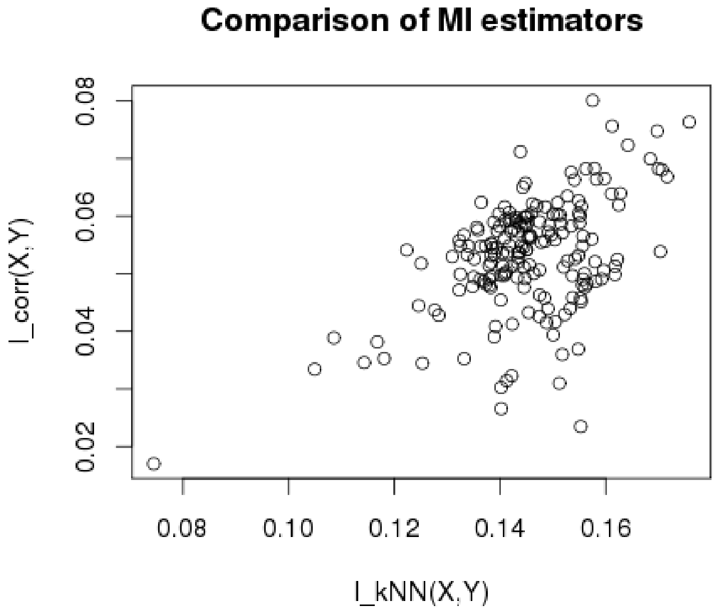

Figure D. Comparison of Mutual Information Estimators

BIBLIOGRAPHY

[1] LD Score online data supplement. https://data.broadinstitute.org/alkesgroup/-LDSCORE/. Accessed: 2016-04-12.

[2] Tamara Aid-Pavlidis, Pavlos Pavlidis, and Tõnis Timmusk. Meta-coexpression conservation analysis of microarray data: a "subset" approach provides insight into brain-derived neurotrophic factor regulation. *BMC Genomics*, 10:420, September 2009.

[3] Edoardo M Airoldi and Panos Toulis. Scalable estimation strategies based on stochastic approximations: Classical results and new insights. *Statistics and Computing*, 25:781–795, July 2015.

[4] N H Barton, A M Etheridge, and A Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, July 2017.

[5] Mohamed Bennasar, Yulia Hicks, and Rossitza Setchi. Feature selection using joint mutual information maximisation. *Expert Systems with Applications*, 42(22):8520 – 8532, 2015.

[6] G. E. P. Box and G. C. Tiao. Bayesian estimation of means for the random effect model. *Journal of the American Statistical Association*, 63(321):174–181, March 1968.

[7] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169:1177–1186, June 2017.

[8] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66, 2012.

[9] C. G. Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[10] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47:291–295, March 2015.

[11] A Cattaneo, N Cattane, V Begni, C M Pariante, and M A Riva. The human BDNF gene: peripheral gene expression and protein levels as biomarkers for psychiatric disorders. *Translational Psychiatry*, 6:e958, November 2016.

[12] Guo-Bo Chen. Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Frontiers in Genetics*, 5:107, 2014.

[13] Wenan Chen, Beth R Larrabee, Inna G Ovsyannikova, Richard B Kennedy, Iana H Haralambieva, Gregory A Poland, and Daniel J Schaid. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics*, 200:719–736, July 2015.

[14] 1000 Genomes Project Consortium, Gonçalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, November 2012.

[15] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518:317–330, February 2015.

[16] Christopher De Sa, Kunle Olukotun, and Christopher Ré. Global convergence of stochastic gradient descent for some nonconvex matrix problems. *CoRR*, abs/1411.1134, 2014.

[17] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3:185–205, April 2005.

[18] Bradley Efron and David V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65(3):457–482, December 1978.

[19] Cathy E. Elks, Marcel den Hoed, Jing Hua Zhao, Stephen J. Sharp, Nicholas J. Wareham, Ruth J. F. Loos, and Ken K. Ong. Variability in the heritability of body mass index: A systematic review and meta-regression. *Frontiers in Endocrinology*, 3, 2012.

[20] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R Day, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R B Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J Daly, Nick Patterson, Benjamin M Neale, and Alkes L Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47:1228–1235, November 2015.

[21] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, mar 1970.

[22] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, Renqiang Min, Pedro Alves, Alexej Abyzov, Nick Addleman, Nitin Bhardwaj, Alan P Boyle, Philip Cayting, Alexandra Charos, David Z Chen, Yong Cheng, Declan Clarke, Catharine Eastman, Ghia Euskirchen, Seth Frietze, Yao Fu, Jason Gertz, Fabian Grubert, Arif Harmanci, Preti Jain, Maya Kasowski, Phil Lacroute, Jing Leng, Jin Lian, Hannah Monahan, Henriette O'Geen, Zhengqing Ouyang, E Christopher Partridge, Dorrelyn Patacsil, Florencia Pauli, Debasish Raha, Lucia Ramirez, Timothy E Reddy, Brian Reed, Minyi Shi, Teri Slifer, Jing Wang, Linfeng Wu, Xinqiong Yang, Kevin Y Yip, Gili Zilberman-Schapira, Serafim Batzoglou, Arend Sidow, Peggy J Farnham, Richard M Myers, Sherman M Weissman, and Michael Snyder. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489:91–100, September 2012.

[23] Stephanie M Gogarten, Tushar Bhangale, Matthew P Conomos, Cecelia A Laurie, Caitlin P McHugh, Ian Painter, Xiuwen Zheng, David R Crosslin, David Levine, Thomas Lumley, Sarah C Nelson, Kenneth Rice, Jess Shen, Rohit Swarnkar, Bruce S Weir, and Cathy C Laurie. GWASTools: an R/bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics (Oxford, England)*, 28:3329–3331, December 2012.

[24] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–23, Jan 1970.

[25] Juan R González, Marta González-Carpio, Rosario Hernández-Sáez, Victoria Serrano Vargas, Guadalupe Torres Hidalgo, Marta Rubio-Rodrigo, Ana García-Nogales, Manuela Núñez Estévez, Luis M Luengo Pérez, and Raquel Rodríguez-López. FTO risk haplotype among early onset and severe obesity cases in a population of western Spain. *Obesity*, 20:909–915, April 2012.

[26] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, Schizophrenia Working Group of the Psychiatric Genomics Consortium, SWE-SCZ Consortium, Anna K Kähler, Christina M Hultman, Shaun M Purcell, Steven A McCarroll, Mark Daly, Bogdan Pasaniuc, Patrick F Sullivan, Benjamin M Neale, Naomi R Wray, Soumya Raychaudhuri, Alkes L Price, Schizophrenia Working Group of the Psychiatric Genomics Consortium, and SWE-SCZ Consortium. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics*, 95:535–552, November 2014.

[27] Denes Hnisz, Brian J Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A Sigova, Heather A Hoke, and Richard A Young. Super-enhancers in the control of cell identity and disease. *Cell*, 155:934–947, November 2013.

[28] J.S. Hodges. *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[29] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, feb 1970.

[30] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198:497–508, October 2014.

[31] Manolis Kellis, Barbara Wold, Michael P Snyder, Bradley E Bernstein, Anshul Kundaje, Georgi K Marinov, Lucas D Ward, Ewan Birney, Gregory E Crawford, Job Dekker, Ian Dunham, Laura L Elnitski, Peggy J Farnham, Elise A Feingold, Mark Gerstein, Morgan C Giddings, David M Gilbert, Thomas R Gingeras, Eric D Green, Roderic Guigo, Tim Hubbard, Jim Kent, Jason D Lieb, Richard M Myers, Michael J Pazin, Bing Ren, John A Stamatoyannopoulos, Zhiping Weng, Kevin P White, and Ross C Hardison. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 111:6131–6138, April 2014.

[32] Gleb Kichaev and Bogdan Pasaniuc. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *American Journal of Human Genetics*, 97:260–271, August 2015.

[33] Gleb Kichaev, Megan Roytman, Ruth Johnson, Eleazar Eskin, Sara Lindström, Peter Kraft, and Bogdan Pasaniuc. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics (Oxford, England)*, 33:248–255, January 2017.

[34] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, 10:e1004722, October 2014.

[35] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review. E*, 69:066138, Jun 2004.

[36] Danny Leung, Inkyung Jung, Nisha Rajagopal, Anthony Schmitt, Siddarth Selvaraj, Ah Young Lee, Chia-An Yen, Shin Lin, Yiing Lin, Yunjiang Qiu, Wei Xie, Feng Yue, Manoj Hariharan, Pradipta Ray, Samantha Kuan, Lee Edsall, Hongbo Yang, Neil C Chi, Michael Q Zhang, Joseph R Ecker, and Bing Ren. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, 518:350–354, February 2015.

[37] Adam E. Locke, Bratati Kahali, Sonja I. Berndt, Anne E. Justice, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 2015.

[38] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. arXiv:1068.03983

[39] Qiongshi Lu, Yiming Hu, Jiehuan Sun, Yuwei Cheng, Kei-Hoi Cheung, and Hongyu Zhao. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Scientific Reports*, 5:10576, May 2015.

[40] Qiongshi Lu, Xinwei Yao, Yiming Hu, and Hongyu Zhao. GenoWap: GWAS signal prioritization through integrated analysis of genomic functional annotation. *Bioinformatics (Oxford, England)*, 32:542–548, February 2016.

[41] Shane Neph, Andrew B Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A Stamatoyannopoulos. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150:1274–1286, September 2012.

[42] Paul J Newcombe, David V Conti, and Sylvia Richardson. JAM: A scalable Bayesian framework for joint analysis of marginal SNP effects. *Genetic Epidemiology*, 40:188–201, April 2016.

[43] Bogdan Pasaniuc and Alkes L Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews. Genetics*, 18:117–127, February 2017.

[44] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238, 2005.

[45] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, 94:559–573, April 2014.

[46] Evadnie Rampersaud, Braxton D Mitchell, Toni I Pollin, Mao Fu, Haiqing Shen, Jeffery R O'Connell, Julie L Ducharme, Scott Hines, Paul Sack, Rosalie Naglieri, Alan R Shuldiner, and Soren Snitker. Physical activity and the association of common FTO gene variants with body mass index and obesity. *Archives of Internal Medicine*, 168:1791–1797, September 2008.

[47] G. K. Robinson. That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32, 1991.

[48] D. F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(111):647–647, Sep 1970.

[49] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, Jul 1948.

[50] Zhan Su, Jonathan Marchini, and Peter Donnelly. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics (Oxford, England)*, 27:2304–2305, August 2011.

[51] Gosia Trynka, Cynthia Sandor, Buhm Han, Han Xu, Barbara E Stranger, X Shirley Liu, and Soumya Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45:124–130, February 2013.

[52] Greg Ver Steeg. Non-parametric entropy estimation toolbox (NPEET), 2017.

[53] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of GWAS discovery: Biology, function, and translation. *American Journal of Human Genetics*, 101:5–22, July 2017.

[54] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89:82–93, July 2011.

[55] Howard H. Yang and John Moody. Feature selection based on joint mutual information. In *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, pages 22–25, 1999.

[56] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication, Meta analysis (DIAGRAM) Consortium, Pamela A F Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael N Weedon, Ruth J Loos, Timothy M Frayling, Mark I McCarthy, Joel N Hirschhorn, Michael E Goddard, and Peter M Visscher. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, 44:369–75, S1–3, March 2012.

[57] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88:76–82, January 2011.

[58] Jingjing Yang, Lars G Fritsche, Xiang Zhou, Gonçalo Abecasis, and International Age-Related Macular Degeneration Genomics Consortium. A scalable Bayesian method for integrating functional information in genome-wide association studies. *American Journal of Human Genetics*, 101:404–416, September 2017.

[59] Qingyun Yang, Tiancun Xiao, Jiao Guo, and Zhengquan Su. Complex relationship between obesity and the fat mass and obesity locus. *International Journal of Biological Sciences*, 13:615–629, 2017.

[60] Xiang Zhou. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *Annals of Applied Statistics*, 11(4):2027-2051, December 2017.

CURRICULUM VITAE

[redacted]

[redacted]

[redacted]