THE UNIVERSITY OF CHICAGO

ON ESTIMATING THE GENETIC BASIS OF COMPLEX TRAITS AND THEIR MOLECULAR INTERMEDIATES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS AND SYSTEMS BIOLOGY

BY
NICHOLAS WILSON KNOBLAUCH

CHICAGO, ILLINOIS
MARCH 2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like first to acknowledge all the people in my life who have been patient with me. The academic world doesn't run on money so much as it runs on people's time. As a consequence, with patience comes the risk of profligacy. There is no one in my academic life who has been as patient with me as my advisor, Assistant Professor Xin He. In the early days of my PhD, we would meet for several hours at a time and he would tutor me in Bayesian statistics, or we would review foundational papers in computational biology. As I developed as a student and as the group grew, we met less frequently, but he was always available when I needed his help. I have never met anyone with such an insatiable scientific curiosity. Xin's group leadership style is to lead his group by example. Anyone who has attended a He Lab meeting or taken a class from him is familiar with his refrain, after introducing a topic and sensing his audience is unfamiliar: "you should know this". At first hearing this, it lands as an expression of disappointment (which I'm sure it sometimes is), but I've come to appreciate that there's a lot more to it than that. It is a call to do better. It is an affirmation that Xin believes in you, that you are someone who might have known, and who can learn. We come to the academy to learn, Xin's attitude is a reminder to all those around him of that. I am Xin's first PhD student; as challenging as this PhD has been for me, I'm sure it wasn't easy for him either, and so for that I will be eternally grateful.

I would also like to express deep gratitude to my committee, Marcelo Nobrega, Carole Ober, and Matthew Stephens. Carole has taught me so much about the practice of human genetics, and about what it takes to be a leader of a large, multi-institution scientific endeavor. Marcelo has given me excellent feedback as the preterm birth project, and despite several setbacks and delays, has always been a source of clear, honest scientific vision. Being on the fourth floor of Cummings, I have had the privilege of witnessing scientists from near and far, young and old, present their work at the weekly fourth floor meetings. I feel as though I have learned as much having Matthew as a fellow audience member as I have from the presenta-

tions themselves. Matthew has an incredible ability to rapidly get to the heart of a scientific problem, and to re-frame it in a clear and simple way.

I would like to thank my parents, who have held academic achievement in high regard since as long as I can remember. As I began my PhD journey, I feel like I got to know them all over again in a new light, as fellow former PhD students. While neither of my parents are biologists, I attribute my fascination with the field to the manner in which my parents each in their own way raised me to seek out, to investigate, and to nurture, the sacred and the profound. I have had the immense fortune of having had incredible office mates in the course of my PhD. I would have been lucky to have one of Wei Wang, Xiang Zhu, and Abhishek Sarkar as office mates, but I managed to share an office with all three of them. It is no exaggeration when I say that I truly consider each of them world class thinkers — the ability to swivel around in my chair and pose a question that popped into my head to one or several of them at once, and to have them stop whatever they were doing and patiently explain it to me was the most wonderful parts of graduate school for me.

I have only met my daughter Lucía in the final weeks of my dissertation writing process, but she has been a quite incredible motivating influence on me, so I would like to thank her for that. Finally, I would like to thank my wife, Sofía, who has supported me throughout my PhD. There are many ways in which I have had to put my life on hold while pursuing my PhD. In sharing her life with me, my wife has had to put her life on hold as well. The hold took longer than either of us anticipated, and I am eternally grateful for her patience, and for the privilege of sharing my life with her.

ABSTRACT

For a variant to have a causal effect on an organism-level trait, there must be a chain of causal events starting with the DNA-sequence, proceeding through one (or more often many) molecular intermediates in a functional pathway before it is observable at the organism level. The causal influence of a variant on a trait is, unfortunately, neither necessary nor sufficient for a variant to appear to be statistically associated with a trait in an association study, and spurious association of genotype and phenotype is not uncommon. It is for this reason that much of the hard work of a genetic association study begins after the association statistics have been generated. The true goal of the genetic association study is not simply to identify the genetic variation that most correlates with phenotype, but to try to identify the set of variants whose correlation with the trait of interest are driven by causal relationships, rather than by coincidence or confounding. In this dissertation I discuss three strategies for relating genotype to phenotype based on the results from genetic association studies. I first discuss the method FGEM, which combines the output from gene-based association tests with gene-level annotation data to both estimate the enrichment of the annotations and re-prioritize genes based on those enrichment estimates. I find that FGEM's joint modeling of gene-level association data with gene-level annotation data is a powerful approach for identifying enriched pathways. Furthermore, I find that identification of enriched pathways can be used to identify additional causal genes. Next I describe my method for heritability estimation from summary statistics, RSSp. RSSp uses GWAS summary statistics and an estimate of pairwise Linkage Disequilibrium (LD) to estimate narrow-sense heritability (h^2). I find that RSSp estimates heritability in polygenic traits from GWAS summary statistics and a reference LD panel with accuracy comparable to in-sample methods. Finally, I discuss my efforts in discovering risk genes for preterm birth via fine-mapping GWAS summary statistics. I find that disease-relevant functional genomic annotations are useful for improving statistical fine-mapping. Using this approach I identified new genes not (directly) implicated from GWAS alone.

CHAPTER 1

INTRODUCTION

Decades of population genetics theory and statistical genetics evidence have demonstrated that for any complex human trait, common sequence-level variation at the vast majority of genetic loci has a small effect on that trait. As a consequence, the statistical association signal at the majority of these loci, taken individually, is difficult to distinguish from random noise. At the same time, countless loci have been identified, through rigorous statistical genetics and increasingly, through functional validation, as containing variants causally-linked to hundreds if not thousands of traits[11]. For a variant to have a causal effect on an organism-level trait, there must be a chain of causal events starting with the DNA-sequence, level change, proceeding through one (or more often many) molecular intermediates in a functional pathway[1], before it is observable at the organismal level. The first link in this chain connects the causal variant to the gene. We can broadly categorize variant to gene causal relationships, coding and non-coding. By virtue of the mapping of the human genome and our knowledge of the central dogma of molecular biology [23], by knowing the position and sequence of a variant one can immediately know with the variant lies within an exon (or splice-site boudary) of a characterized gene. If so, it is a simple bioinformatic exercise to ascertain the consequence of the change in DNA sequence on primary (and increasingly secondary and tertiary) amino acid structure[2]. The functional consequence of the majority of human genetic variation falls into the second category of variant to molecular-intermediate causal relationships[58]. This second category consists of all other mechanisms of causality, perhaps the best understood of which are the ”regulatory variants”. Unlike the relationship between protein coding DNA, RNA and protein, which is essentially universal among living things[22], the relationship between a variant outside of a coding region and the gene through which the variant acts can be extremely variable[30]. Indeed, the ability for the cell to regulate gene expression relies on the dynamic nature of gene regulatory elements[82], and in multi-cellular organisms, the ac-

tivity of gene-regulatory elements varies not just across time, but also across tissue[30]. Constructing a putative path from genotype to a phenotype with a cell-type specific etiology for a non-coding variant (or set of variants) requires accounting for this complexity.

The causal influence of a variant on a trait is, unfortunately, neither necessary nor sufficient for a variant to be associated with a trait. It is not uncommon for variants that appear highly correlated with a trait to be spuriously associated[48]. It is for this reason that much of the hard work of a genetic association study begins after the association statistics have been generated. The true goal of the genetic association study is not simply to identify the genetic variation that most correlates with phenotype, but to try to identify the set of variants whose correlation with the trait of interest are driven by causal relationships, rather than by coincidence or confounding. A common first step in the identification of the genetic variants that contribute to variation in a trait is a Genome-wide association study (GWAS). In a GWAS, tens of thousands, or increasingly, hundreds of thousands of individuals are genotyped and phenotyped, and the millions of loci that commonly vary in the sample are queried for their association with a trait of interest. To identify variants causally related to a trait of interest it is necessary that such variants exist, which is to say that the trait must be heritable. GWAS, and the summary statistics they generate, provide a genome-wide view of individual variant-gene association, but as a consequence of the correlation between variants induced by linkage disequilibrium (LD), it can be difficult to pinpoint the individual causal variants, the genes these causal variants act through, as well as the total genetic contribution of causal variants to variation in the trait of interest.

In this dissertation I will discuss three strategies for relating genotype to phenotype. In the second chapter I outline the method FGEM, which combines the output from gene-based association tests with gene-level annotation data to both estimate the enrichment of the annotations and re-prioritize genes based on those enrichment estimates. I then employ FGEM to the task of identifying mutational cancer driver genes, using gene-level Bayes factors from

the recently published `driverMAPS` method[96], combined with gene-level annotation from the Gene Ontology[83]. Taken together, the recurrently enriched biological processes identified by FGEM recapitulate the hallmarks of cancer[37]. FGEM further implicates several biological processes as being relevant in a subtype-specific manner. Using these enrichment estimates, FGEM identifies cancer genes that are either known cancer genes from the literature, but missed by `driverMAPS`, known cancer genes in other cancer types but implicated in a new cancer type, and a few genes not previously known to be cancer genes in any cancer type.

In the third chapter I describe my method for heritability estimation from summary statistics, RSSp. RSSp uses GWAS summary statistics and an estimate of pairwise Linkage Disequilibrium (LD) to estimate narrow-sense heritability (h^2). RSSp is based on the previously published Regression with Summary Statistics (RSS) likelihood[98] — by using an infinitesimal prior and by modeling z-scores rather than regression coefficients, a form of the marginalized likelihood is revealed that is very fast to compute, and thus, to optimize. To evaluate the performance of RSSp compared to existing methods for heritability estimation I employ a large-scale GWAS simulation. In simulations across a variety of genetic architecture, based on real genotypes from the UK biobank[81], I show that RSSp estimates h^2 better than the widely used LD score regression[10] — outperforming in terms of both bias and variance — and I show how RSSp performs comparably to a method for heritability estimation based on individual-level data (i.e `GCTA`). I additionally discuss considerations in matching a reference LD panel to a GWAS.

In the fourth chapter I describe my efforts at incorporating functional annotations with GWAS data to identify causal genes in the context of preterm birth (PTB). PTB is believed to be responsible for approximately 1 million deaths globally, and is believed to be the leading cause of death for children under 5[50]. While several loci have been associated with the risk of PTB, the causal variants in these loci remain unknown. I first assessed enrichment of GWAS

signals of gestational duration in functional annotations of variants in pregnancy related cell types. I then took advantage of this enrichment to perform Bayesian statistical fine-mapping using susie[85]. Using this strategy we are able to identify new gestational duration associated variants that would not have been identified without the functional information. The fine-mapped causal variants were then linked to genes using a combination of promoter-capture HiC from the cell type(s) of interest, variant locations relative to genes (e.g. inside coding or UTR sequences), and distance information. The genes we identified were significantly enriched both for genes differentially expressed between endometrial mesenchymal stem/stromal cells and differentiated decidual stromal cells.

CHAPTER 2

FGEM: A BAYESIAN METHOD FOR GENE DISCOVERY THAT INTEGRATES FUNCTIONAL INFORMATION ABOUT GENES

2.1 Introduction

Geneticists often aggregate genetic evidence of variants within a gene to test if the gene is related to a trait of interest. These gene-level tests are among the most powerful and commonly used tools in the geneticists' toolkit for relating genotype to phenotype. Gene-based tests have been used in many contexts, including Genome-wide association studies[26], Transcriptome-wide association studies[33], rare variant analysis from exome sequencing studies[44, 90], rare variant analysis in family data[39], and cancer driver gene discovery using somatic mutations[96]. In fact, almost all methods for genetic analysis that use very rare mutations, including *de novo* germline and somatic mutations, are gene-based, as it is almost impossible to identify individual causative mutations[51, 59].

There are several advantages to performing gene-level tests using a Bayesian framework. Variants inside a gene often have very different functionalities, e.g. nonsense mutations can be highly deleterious while missense mutations can have very different effects depending on where they are located. Similarly, for tests that combine different types of variants in a gene, such as common and rare variants, it is important to consider different effects of these variants, with rare variants generally having more deleterious effects. While frequentist methods in theory can use different weights for different groups of variants in the test[44], it is difficult in practice to know what weights should use. Bayesian statistical framework allows researchers to effectively combine evidence of different sets of variants, by using different effect size distributions. Importantly, these distributions can be estimated from data using Empirical Bayes methods or by fully Bayesian Inference with MCMC[63]. The power of a Bayesian approach to gene-based analysis has been demonstrated in very different contexts. Transmis-

sion and De novo Association test (TADA) and its extensions[39] are widely used to analyze variants from sequencing studies in parent-child trios, by effectively combining de novo and inherited variants to improve the power. DriverMAPS is a recently developed method for identifying signatures of positive selection in cancer driven genes, with the strength of selection varying across positions in a gene depending on functional information of the positions[96]. The results of Bayesian gene-based analysis are often summarized as Bayes factors, which compares the model where the gene has an effect on phenotype (causal model) vs. non-causal model.

A common follow-up after identifying a set of putative causal genes is pathway analysis, which tests if certain biological pathways are enriched in these genes[99][15][49]. This approach is supported by the evidence that the functions of genes underlying complex traits, including cancer, often converge on certain biological processes[86]. In the simplest form, pathway analyses apply a cutoff on trait-associated genes, and then use Fisher's enrichment test for pathways overrepresented in the genes that pass the cutoff. More sophisticated analyses compare the distribution of trait associations in genes in a pathway to those outside the pathway. Pathway analysis is an important tool for geneticists to learn possible biological mechanisms by which the putative causal genes may act to influence the trait of interest. This information can provide guidance in deciding on which genes are the most likely to replicate, and most worthy of follow-up[42].

Gene-level tests and pathway analysis are generally treated as separate problems. There are compelling reasons to combine the two, using the pathway enrichment results to set informative priors in Bayesian gene-level tests. Conceptually, if a gene belongs to a disease-related pathway, then *a priori*, the gene is more likely to be a disease gene. Incorporating this prior would thus improve our power to identify disease risk genes. This possibility has been demonstrated in earlier work on GWAS variant-level analysis. In `fgwas`, for example, a variant is annotated by functional information such as conservation and enhancer marks, and the method

learns the enrichment of these annotations in putative causal variants and uses these results to set prior probabilities of association of variants[66]. Similar ideas have also been used for gene-based analysis in GWAS, where gene annotations are based on pathways[15][100]. All these methods, however, are specifically designed for GWAS, and cannot be used for other gene-based analysis, e.g. those based on rare variants or somatic mutations. Additionally, for the gene-based GWAS analysis, they can only incorporate a prior derived from a single annotation, while in reality, multiple pathways/annotations may be informative.

We propose a method for combining gene-level evidence, as summarized by Bayes factors, with one or more gene-level annotations to jointly estimate the global enrichment of the annotations, and to re-estimate a gene-level posterior conditional on the estimated enrichment. We call this method `FGEM`. The method is easy to use, requiring only a set of Bayes factors from gene-level analysis and functional annotations of the genes. This generality makes the method, and our software, useful for a wide variety of settings. FGEM is related to frequentist methods that control false discovery rate or family-wise error rate while weighing different hypothesis using external information[91][92][94]. However, these methods cannot be applied to the Bayesian setting, which, as we argued, has advantages in gene-based tests. We demonstrate the power of the `FGEM` model by applying it to the problem of identifying mutational cancer driver genes. We use the gene-based Bayes factors generated by the `driverMAPS` method, as applied to 18 cancer types from The Cancer Genome Atlas (TCGA) data[8][96], and use the Gene Ontology Biological Processes as gene-level annotations.

2.2 Method

2.2.1 FGEM Model

FGEM uses an empirical Bayes approach to construct the *a priori* probability that a gene is causal, based on the annotations of genes. The method uses genetic data (summarized as

gene-level Bayes factors or likelihood ratios) and a set of gene-level annotations to inform which annotations are relevant and to what extent.

For each gene $g \in \{1 \dots G\}$, let the indicator variable $z_g = 1$ denote that gene g is causally related to the trait or disease of interest. The evidence for and against the hypothesis that $z_g = 1$ can be summarized using a Bayes factor:

$$B_g = \frac{P(x_g|z_g = 1)}{P(x_g|z_g = 0)}$$

where x_g is the subset of a length G vector of genetic data corresponding to the g-th gene.

Let F be the number of features for which functional annotations are available for each of our G genes. Let $\mathbf{a}_g$ denote the length F vector of annotations for gene g, and $\mathbf{A}$ denote the matrix with F rows and G columns consisting of $\mathbf{a}_1 ... \mathbf{a}_G$ We define the prior probability of z_g as a logistic function of the annotations of g, and we include a parameter we will refer to as the "intercept", β_0:

$$\pi(\beta, \mathbf{a}_g) = P(z_g = 1|\mathbf{a}_g, \beta) = \frac{1}{1 + e^{-(\beta_0 + \sum_{f=1}^{F} A_{f,g}\beta_f)}} \tag{2.1}$$

The likelihood of β is computed by treating the data from each gene as coming from a two-component mixture model (where $z_g = 1$ and where $z_g = 0$) and marginalizing over the two components:

$$P(\mathbf{x}|\beta, \mathbf{A}) = \prod_{g=1}^{G} P(x_g|\beta) = \prod_{g=1}^{G} [\pi(\beta, \mathbf{a}_g)P(x_g|z_g = 1) + (1 - \pi(\beta, \mathbf{a}_g))P(x_g|z_g = 0)]$$

By factorizing out the term $\prod_{g=1}^{G} P(x_g|z_g = 0)$ (which does not depend on β), the likelihood for β (up to a constant of proportionality) can be expressed in terms of $\mathbf{B}$:

$$P(\mathbf{x}|\beta,\mathbf{A}) \propto \prod_{g=1}^{G} [\pi(\beta,\mathbf{a}_g)B_g + (1-\pi(\beta,\mathbf{a}_g))] \tag{2.2}$$

Given a particular value of β, and a bayes factor B_g, the posterior probability that $z_g = 1$ is given by:

$$P(z_g = 1|B_g,\beta,\mathbf{a}_g) = \frac{\pi(\beta,\mathbf{a}_g)B_g}{\pi(\beta,\mathbf{a}_g)B_g + 1 - \pi(\beta,\mathbf{a}_g)} \tag{2.3}$$

The goal of the FGEM method is to simultaneously estimate the enrichment β for a relevant set of features and the gene-level posterior probability $P(Z_g = 1|\mathbf{a}_g,\beta,x_g)$ that each gene is causally related to the trait of interest. FGEM estimates β by maximizing the log-likelihood, given by Equation 2.2. Given an estimate of β it is straightforward to compute $P(Z_g = 1|\mathbf{a}_g,\beta,x_g)$ based on Equation 2.3.

When the number of annotations is large, it is impossible, from both a computationally and interpretability standpoint, to include all features in the model. Furthermore, as the number of features in the model increases, the probability that some subset of features will be collinear with one-another increases, which can complicate model-fitting, as β becomes unidentifiable. This is especially important when a binary, hierarchical feature set like the Gene Ontology. To avoid these issues, FGEM maximizes the penalized log-likelihood, in a multi-stage feature-selection and model fitting procedure. In the first step, all single-feature-plus-intercept models are fit, and a p-value is obtained for each model by comparing the single-feature-plus-intercept model to the intercept-only model via the likelihood ratio test. From this set of single-feature models, all of the nonsignificant (i.e. features with Benjamini-Hochberg adjusted \$p\$-values greater than 0.05) features were removed from the analysis.

In the second step, significant features passing the filter were combined in a joint model and fit by maximizing the log-likelihood, penalized with an elastic-net penalty. The Limited Memory Broyden-Fletcher-Goldfarb-Shanno algorithm (LM-BFGS)[13] is among the most pop-

ular algorithms for unconstrained optimization over scalar, differentiable functions, and while suitable for the un-penalized single-feature plus intercept models, cannot be used in the penalized setting without modification. One limitation of LM-BFGS, is that the function that is being optimized must be differentiable. Unfortunately, sparsity-inducing l_1-regularized models of the form:

$$f(\theta) = p(\theta|\mathbf{x}) + C\|\theta\|_1$$

are not differentiable when any of the elements of the parameter vector (θ) are 0[35]. The Orthant-wise limited-memory quasi-Newton method is a variant of LM-BFGS which is designed precisely for fitting L_1-regularized, sparsity inducing models. FGEM utilizes the Orthant-wise LM-BFGS algorithm to maximize the marginalized likelihood in the presence of a non-zero l_1 penalty.

One well-known problem with L_1 penalty is that it encourages both sparse models and shrinks the parameters. The two goals may conflict with each other. When the number of features/parameters is large, a large L_1 penalty may be needed to remove noisy features, but then it may shrink the parameters of the true features too much, leading to sub-optimal parameter estimates and prediction performance. This has motivated the relaxed lasso procedure[38], and has been adopted by the widely used R `glmnet` package. The relaxed Lasso procedure typically starts with standard Lasso to select feature, but in the next step, refit the model using only selected features with a smaller (or no) penalty to shrink parameters. Our procedure here is a relaxed version of elastic net. Specifically, in the first step, the objective function corresponding to the negative of the elastic-net penalized log-likelihood:

$$-\mathscr{L}(\beta;\mathbf{A},\mathbf{x}) + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{F}\beta_j^2 + \alpha\sum_{j=1}^{F}|\beta_j|\right)$$

is minimized, where $\mathscr{L}(\beta;\mathbf{A},\mathbf{x}) = \sum_{g=1}^{G}[\log\Big(\pi(\beta,\mathbf{a}_g)B_g + (1-\pi(\beta,\mathbf{a}_g))\Big)]$.

The overall level of sparsity in the model is controlled by the parameter λ, while the proportion of l_1 vs l_2 penalty is determined by α. For this first step FGEM uses a default value of $\alpha = 0.95$, corresponding to a higher l_1 penalty relative to the l_2 penalty, which has the effect of encouraging sparsity in the model.

In the second step, features with $\beta_f = 0$ are removed from the analysis, and the model is refit with an l_2 penalty only.

$$-\mathscr{L}(\beta; \mathbf{A}, \mathbf{x}) + \lambda \left(\gamma \sum_{j=1}^{F} \beta_j^2 \right)$$

The hyperparameter γ allows for a stronger or weaker l_2 penalty in the second step as compared to the first step, but by default is set to $1 - \alpha$, yielding the same l_2 penalty as in the first step.

The optimal value of λ is chosen using 10-fold cross-validation over 100 values of λ, starting at zero, and ending at $\lambda_{\max}$, equally spaced on a log scale. $\lambda_{\max}$ is defined as the smallest value of λ which yields a $\beta = 0$ (excluding the intercept) from the elastic-net fit. For each of the 10 training-testing cross-validation splits, the two stage FGEM model is fit for each of 100 values of λ on the training set. The β from the two-stage fit is used to calculate the (unpenalized) log-likelihood on the testing set. The optimal λ is the λ with the highest testing-set log-likelihood summed over all 10 cross-validation folds.

Comparison with Fisher's Exact test

In the case of a single binary feature, one can apply a Bayes Factor cutoff to obtain a contingency table and assess the enrichment of the feature using Fisher's Exact test. We compared FGEM with Fisher's exact test, using an FDR cutoff of 0.1, and compared the p-values to those obtained from the single-feature, likelihood ratio test FGEM p-values.

2.2.2 Genetic Data from The Cancer Genome Atlas and `driverMAPS`

The Cancer Genome Atlas (TCGA) is a resource consisting of data on over 20 cancer types, gathered from thousands of individuals[8]. For several cancer types TCGA data include high coverage, whole-exome sequencing data for both the patient's solid tumor and matched adjacent normal tissue. By aggregating the somatic mutation data across multiple individuals with a particular cancer type, one can identify a set of genes that undergo somatic mutation at a frequency higher than expected by chance. For each of 18 TCGA cancer types, we used 20,848 gene-level Bayes factors obtained from running the statistical method `driverMAPS`, a recently developed Bayesian method for identifying driver genes, as input to the FGEM model. After obtaining the total set of gene-level Bayes factors, we eliminated "blacklisted" genes known to have mapping problems[74], as well as Olfactory Receptors, which are known to have mappability problems — one survey assessing the mappability of genes in the human genome found that the fraction of mappable reads for genes in the olfactory receptor family is at least 10% lower than the average-protein coding gene.[27].

2.2.3 Gene-Level Annotations using Gene Ontology

The "Biological Process" gene sets from the Gene Ontology were obtained using the Bioconductor package `GO.db`[16]. Of the 10,930 possible biological process gene ontology terms, the 2,198 terms that include 10 or more genes were deemed eligible for incorporation in this analysis, so as to reduce the multiple testing burden, and to ensure that we were well powered to accurately estimate the enrichment of each term included in the analysis. Each gene ontology term was encoded as a binary, gene-level feature using an indicator variable to indicate a gene's association with the corresponding term.

2.2.4 *Validation of predicted cancer genes using external resources*

IntOGen is a database of cancer driver genes[36]. It is populated by an ensemble method that incorporates seven different methods for identifying cancer driver genes. It weights each of the 7 methods according to their ability to predict membership in the The Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC)[75], making it a useful resource for benchmarking methods for identifying cancer genes. The database provides the dataset in which the driver gene was identified. After removing IntOGen driver genes that were identified in TCGA, we evaluated the performance of FGEM by computing the average increase in the gene-level posterior of the FGEM functionally informed model as compared to the uniform model for IntOGen driver genes and compared it to the average increase in genes not in the IntOGen database.

2.2.5 *FGEM Software Package*

Our method is distributed as a freely available R package[68] `FGEM`, which is available at the GitHub repository `https://github.com/CreRecombinase/FGEM`. In addition to the implementing an optimized version of the FGEM likelihood itself, the package also has functionality for both single-feature (single-annotation) and multi-feature (multiple annotation) model fitting. FGEM relies on both the `RcppEigen`[3] and `StanHeaders` R packages for efficient computation of the likelihood and its gradients, which are passed to the optimizer routine.

2.3 Results

2.3.1 A probabilistic framework for gene-set enrichment and gene prioritization applied to cancer gene discovery

Our approach is outlined in Figure 2.1. In brief, we combine gene-level Bayes factors summarizing the hypothesis that a gene is causally related to a trait of interest with gene-level annotations about that gene to identify the properties causal genes are likely to have, and simultaneously re-estimate the gene-level posterior probability that the gene is causal. Let Z_g be and indicator variable with $Z_g = 1$ indicating that gene g is causally related to the trait of interest, and $Z_g = 0$ indicating that it is not. While Z_g is unobserved, the evidence for and against $Z_g = 1$, as calculated by a gene-based test on some body of genetic evidence, is summarized by the Bayes factor for that gene B_g. FGEM incorporates gene-level annotation, represented as an F (the total number of gene-level features) by G (the total number of genes) matrix $\mathbf{A}$. FGEM relates Z_g to $\mathbf{A}$ through a length F enrichment parameter β. For a particular gene g, $P(z_g = 1|\mathbf{a}_g, \beta) = \frac{1}{1+e^{-(\beta_0+\sum_{f=1}^{F} A_{f,g}\beta_f)}}$. This relationship between feature and and response is analogous to a logistic regression on a latent variable ($\mathbf{Z}$). The procedure for model fitting is described in the Methods section 2.2.

Under the single-feature FGEM model, in which each feature is considered one at at time, if the value of β for a binary gene-level annotation is greater than 0, this indicates that genes with this annotation have a higher probability of being causal to the trait of interest than background genes. It is also possible for the estimate of β to be less than 0, indicating that the genes with the annotation have a lower probability of causal association than random genes. The statistical significance of a single feature is assessed using the likelihood ratio test, testing if $\beta = 0$, from which p-values were calculated.

For a particular value of β (and $\mathbf{A}$), we can compute a new value for $P(Z_g = 1|\mathbf{A}, \beta)$. We refer to the value of β_a as the enrichment of feature a. For each of the 18 TCGA cancer types, we fit

a single-feature model for each of the 2,657 Biological Process related GO terms. We refer to the value of β for each feature when fit one at a time as single-feature enrichment. Significant features for each cancer type were then jointly fit for each cancer type. The estimated value of β for each feature under this model is referred to as the multi-feature enrichment, and the procedure for obtaining multi-feature enrichment estimates is described in the Methods section 2.2.1.

2.3.2 Recurrent enriched annotations reflect the hallmarks of cancer

After removing Gene Ontology Biological Process features with a small number of annotated genes, there were 2,657 features. Evaluating the enrichment of each of these features in a single-feature fashion with the 18 TCGA cancer types resulted 47,826 single-feature enrichment estimates. We first evaluated the number of significantly enriched features, stratified by cancer type. With a false discovery rate (FDR) of 0.01, all cancer types but KIRP and UCS had at least one (i.e. 16 out of 18) significant association, with HNSC having the most, at 38 (Table 2.1). With a relaxed FDR of 0.15, all 18 cancer types had at least one significantly associated feature (Table 2.1). In all cancer types analyzed, all features with enrichments significantly different from 0 (at all tested FDR) were positively enriched.

We next compare our single feature analysis with a simple Fisher's exact test, which uses a cutoff to define candidate genes, and tests enrichment of a pathway in all genes passing the cutoff. Overall, we see clear correlation of the results of these two tests (Figure 2.3), however, one notable difference is that a number of features with small p-values by FGEM single feature model have p-values equal to 1 under FET - see the data points close to the vertical y-axis in Figure 2.3. This counter-intuitive results can be explained by the fact that with a hard cutoff, many gene sets will have no genes passing the cutoff, so will be missed by FET, having $p = 1$ under FET. Since FGEM treats gene status z_g as latent variable to be marginalized out, avoiding the hard cutoff, it is possible to identify these gene sets.

cancer	FDR= 0.01	FDR= 0.05	FDR=0.1	FDR=0.2
HNSC	38	95	122	209
LUAD	31	77	111	190
BLCA	26	51	67	110
UCEC	22	52	92	141
GBM	20	42	55	99
CESC	14	36	67	115
PAAD	11	22	32	54
BRCA	8	26	37	82
LIHC	8	33	55	95
LUSC	4	11	24	53
PRAD	3	15	50	131
SKCM	3	8	14	26
KIRC	2	2	14	53
ESCA	1	13	31	54
SARC	1	8	14	29
TGCT	1	8	15	27
KIRP	0	0	0	6
UCS	0	4	20	23

Table 2.1: Number of significantly enriched features in single-feature enrichment test at four False Discovery Rates.

We identified a set of recurrent features: features that were significantly enriched in more than one cancer type. We characterized 161 Gene Ontology features as significantly enriched in more than one cancer type, and 50 features were significantly enriched in 5 or more cancer types. The top features ranked by the number cancer types in which the feature was enriched 2.2 recapitulates almost all of the 10 "Hallmarks of cancer" [37]. The feature with the highest mean (and median) enrichment is GO:2000774, positive regulation of cellular senescence, with a median enrichment estimate of 5.458, and a mean enrichment estimate of 7.91. This feature is important for tumorigenesis: mutations in genes that prevent cells from entering oncogene induced cellular senescence (especially those related to the ARF/TP53 pathway) are essential for progression of almost all cancers[18]. Two most recurrent features, those enriched in the largest number of cancers, include: GO:0007265 (Ras protein signal transduction) and GO:0008285, negative regulation of cell population proliferation (Table 2.2). Onco-

genic Ras, and members of the Ras signaling pathway, have been implicated in several other cancer hallmarks[67]. Negative regulation of cell population proliferation, like positive regulation of cellular senescence, is key to preventing uncontrolled cell growth, a defining feature of cancer[37].

GO Term	Average β	No. significant	Description
GO:0007265	3.28	12	Ras protein signal transduction
GO:0008285	2.29	12	negative regulation of cell population proliferation
GO:0019221	2.63	11	cytokine-mediated signaling pathway
GO:0010628	2.34	11	positive regulation of gene expression
GO:0032228	5.60	10	regulation of synaptic transmission, GABAergic
GO:0010666	4.22	10	positive regulation of cardiac muscle cell apoptotic process
GO:0051402	3.17	9	neuron apoptotic process
GO:2000134	2.94	9	negative regulation of G1/S transition of mitotic cell cycle
GO:0007050	2.51	9	cell cycle arrest
GO:0045893	2.02	9	positive regulation of transcription, DNA-templated
GO:0043276	5.25	8	anoikis
GO:2000379	4.22	8	positive regulation of reactive oxygen species metabolic process
GO:0043491	3.57	8	protein kinase B signaling
GO:0043542	3.07	8	endothelial cell migration
GO:0000165	2.66	8	MAPK cascade

Table 2.2: Top features of mutational cancer driver genes from single-feature enrichment analysis. Features are ranked by the number of cancer types in which the feature was significant at (FDR-adjusted) $p \leq 0.1$, and then by the average enrichment estimate among all cancer types.

2.3.3 FGEM integrates multiple gene-level annotations to re-prioritize mutational driver genes

We applied the full FGEM model to 18 cancer types, and used the estimated parameters to compute posterior probabilities of all genes. To evaluate the results, we take advantage of IntOGen, a database of cancer driver genes that is populated by an ensemble method that incorporates seven different methods for identifying mutational driver cancer genes. To check whether FGEM re-prioritization improved prediction of cancer driver genes, we compared the posterior probabilities of IntOGen validated genes vs. the posterior probabilities under the uniform model (intercept only). In every cancer type, validated cancer genes have higher posterior under the functional model as compared to the uniform model, and genes that were not previously identified as cancer genes in IntOGen had on average lower functional posterior compared to uniform (Table 2.2).

After validating that FGEM improves the power to detect mutational driver genes we turned our attention to which genes increased the most as a consequence of the functional prior. The gene with the largest average increase in posterior probability between the functional and uniform posterior, across cancer types, was Transforming Growth Factor Beta 1, or TGFB1. The posterior for TGFB1 increased by an average of 0.503. While TGFB1 is not characterized by IntOGen as a mutational cancer driver (due to the relatively lower number of somatic point mutations observed in tumors), the role of TGFB1 in metastasis, and the role of the TGF-β signaling pathway more generally in cancer progression is widely known and studied[95]. After TGFB1, the gene with the largest increase in posterior probability that is not a known mutational cancer driver is SMAD3, a crucial regulator of the TGF-β signalling pathway[60].

We highlighted the results of two cancer types, breast cancer (BRCA) and uterine cancer (UCEC). We compared the posterior probabilities of genes for these two cancer types, under functional (FGEM) vs. uniform prior (Figure 2.4). In both genes, we see a similar overall trend, with a number of genes showing higher posterior under the FGEM prior model comparing

with the uniform prior (the genes above the diagonal line). Many of these genes are known cancer driver genes, according to IntOGen, as labeled in Figure 2.4. To better understand how these genes are prioritized by FGEM, we plot the top 22 enriched GO processes in Figure 2.5 and Figure2.6 for both cancer types. The majority of these pathways are supported by earlier research. For example, the highest enriched pathways of UCEC includes RAS signaling and cell migration, the processes well known to be important for cancer.

2.4 Discussion

We have developed a statistical model for integrating gene-level Bayes factors with gene-level annotations to simultaneously re-prioritize the genes and estimate the enrichment of the features. In our analysis of gene-level Bayes factors generated from `driverMAPS` run on TCGA data, we find that the addition of gene-level features pushes 208 genes previously below a statistical significance threshold be novel driver genes across 18 cancer types. One of the most salient features of FGEM as compared to other enrichment methods like Fisher's Exact test is that FGEM does not binarize data into significant vs insignificant. This may have some advantage in identifying gene sets where the number of genes passing stringent statistical cutoff is small.

That the previously identified oncogenes JUN and TGFB1 showed a much high posterior probability under the functional model than the uniform models, and that these genes were not previously identified in IntOGen demonstrates the value that a method capable of comprehensive integration of all forms of somatic mutation might provide in identifying cancer genes. Both JUN and TGFB1 rely on mechanisms other than accumulation of point mutation to operate as oncogenes. As a consequence, they will be missed by methods that rely exclusively on somatic point mutation. It should be noted that although the FGEM analysis employed for this paper used binary gene-level features, there is nothing inherent in the method precluding inclusion of categorical or even continuous annotations. For categorical

variables (e.g encoding which, of several possible tissues, a gene is known to be expressed in) this would be trivial: by using a reference level[17] and a treatment encoding (an additional indicator variable for $k-1$ of the categorical variable's k levels, with the k-th level being an implicit reference level), the enrichment estimates would have the same interpretation as log odds ratios over the "intercept" model.

One important caveat of the FGEM model is that it does not account for errors or uncertainty in the gene-level annotations. While the Gene Ontology has a formal process for gene annotation, as well as a controlled vocabulary for describing the evidence underlying every gene-annotation pair, this is not true of most gene-level annotations, and even if it were, it is not clear how one might incorporate this evidence. It is also worth considering the extent to which publication bias, or the "file-drawer effect" might contribute to systematic errors in gene-level annotation. It is impossible to know the number of genes that have been tested for a particular biological process or molecular function. Gene Ontology maintains a blacklist of disallowed gene-feature relationships[43], but it captures only the most commonly misreported gene-feature relationships. Binary gene-level annotation of a GO term ablates the distinction between a gene that has not been assessed for a particular biological process and a gene for which there is strong evidence *against* it being involved in the process.

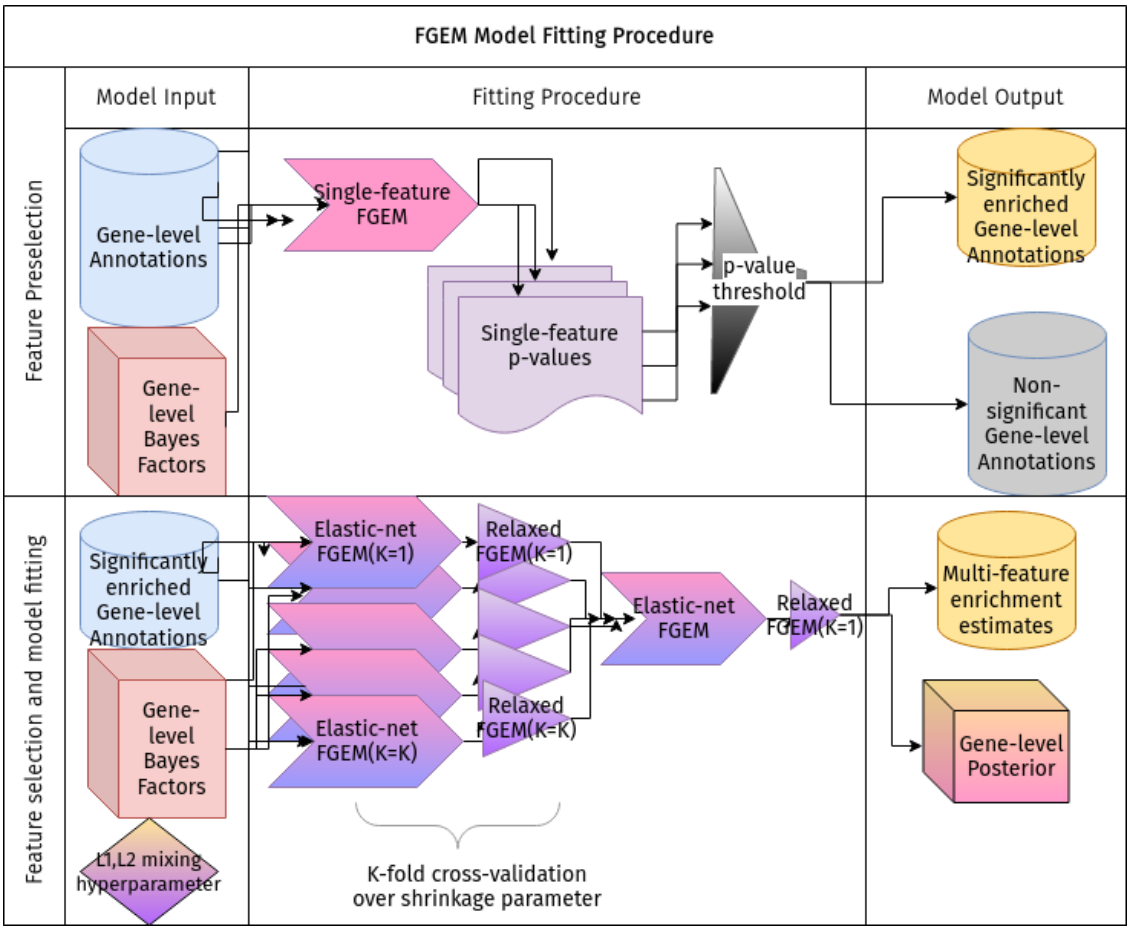

Figure 2.1: Overview of the FGEM procedure for gene-set enrichment and gene mapping. In the preliminary feature pre-selection phase, all single-feature models are fit, and p-values are obtained for each model. Features with FDR-adjusted p-value less than a significance cut-off (0.1) are incorporated in the multi-feature model. The multi-feature model is fit with an elastic-net penalty with a user-specified proportion of l_1 to l_2 penalty (α) (Elastic-net FGEM), then the subset of features with non-zero enrichment are refit with the l_1 penalty set to 0(Relaxed FGEM). K-fold cross-validation is used to determine the optimal penalty parameter (λ_{opt}). The multi-feature enrichment estimates are then used to generate gene-level posteriors.

	$\text{Posterior}_{(\text{Functional})} - \text{Posterior}_{(\text{Uniform})}$	
Cancer	**IntOGen Genes**	**Non-IntOGen Genes**
PAAD	0.0480582	-0.0077148
SARC	0.0356362	-0.0038287
CESC	0.0354395	-0.0031413
ESCA	0.0305561	-0.0049400
PRAD	0.0291863	-0.0060090
BLCA	0.0270332	-0.0026782
LIHC	0.0278445	-0.0017952
GBM	0.0285302	0.0002752
KIRC	0.0249977	-0.0032255
UCEC	0.0245470	-0.0030566
HNSC	0.0194032	-0.0028055
LUAD	0.0196972	-0.0004134
UCS	0.0073904	-0.0039279
KIRP	0.0080442	-0.0024497
BRCA	0.0074382	-0.0010199
TGCT	0.0041605	-0.0036802
SKCM	0.0067283	-0.0008418
LUSC	0.0018588	-0.0023163

Figure 2.2: Average increase in gene-level posterior of the functionally informed posterior as compared to posterior computed from uniform model, computed in validated IntOGen cancer genes and genes not in the IntOGen database. For every cancer type, IntOGen cancer genes on had on average posterior a higher posterior under the functional model than under the uniform. Genes not in the IntOGen database had on average lower posterior under the functional model.

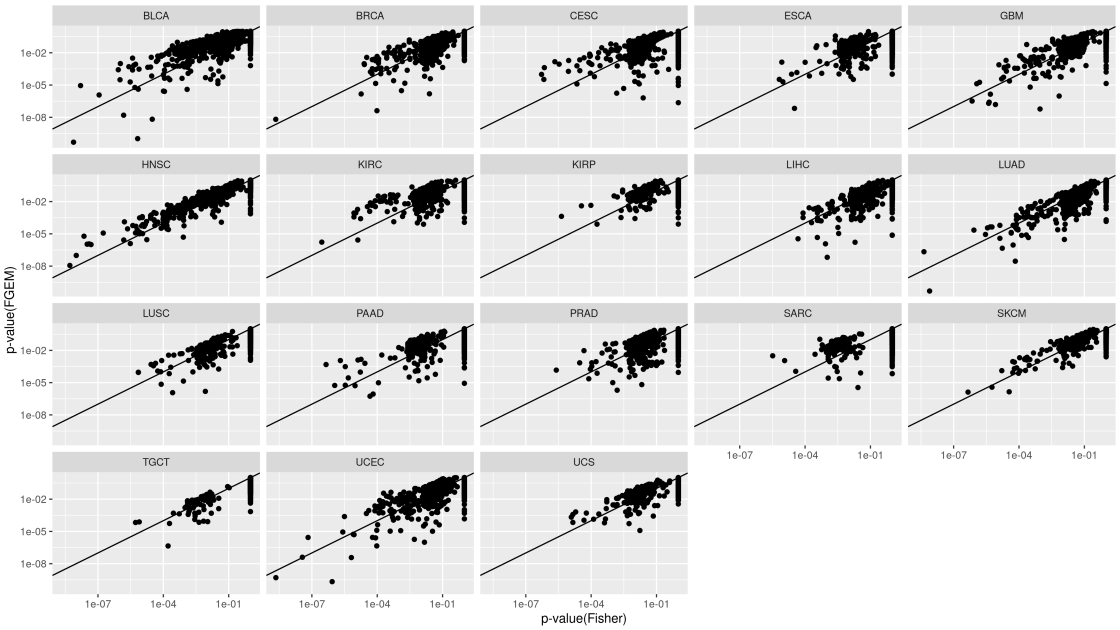

Figure 2.3: Comparison of single-feature FGEM and Fisher's exact test p-values for 18 TCGA cancer types.

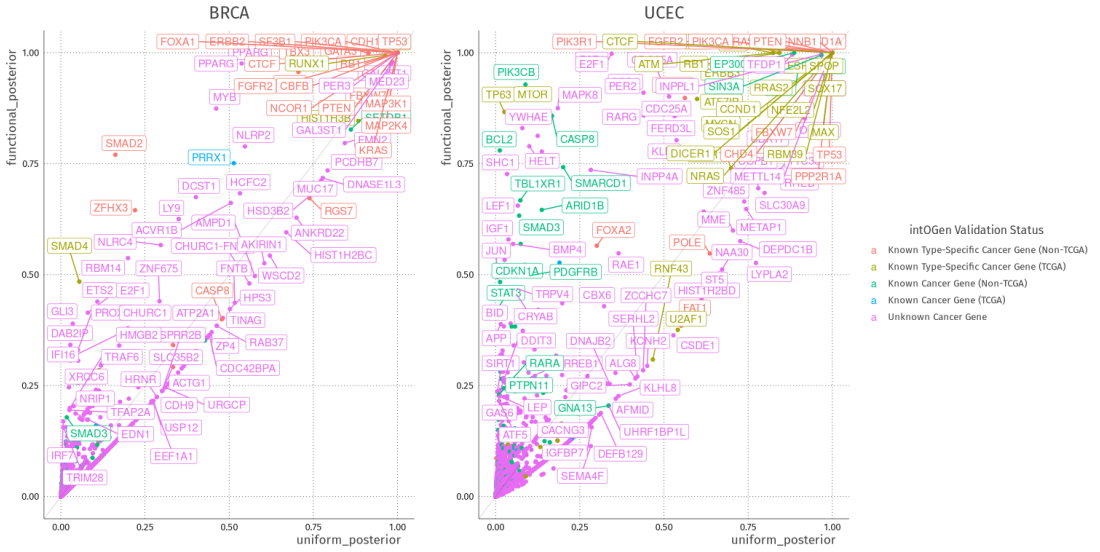

Figure 2.4: Comparison of gene-level posterior under uniform and functional models for Breast Invasive Carcinoma (BRCA) and Uterine Corpus Endometrial Carcinoma (UCEC).

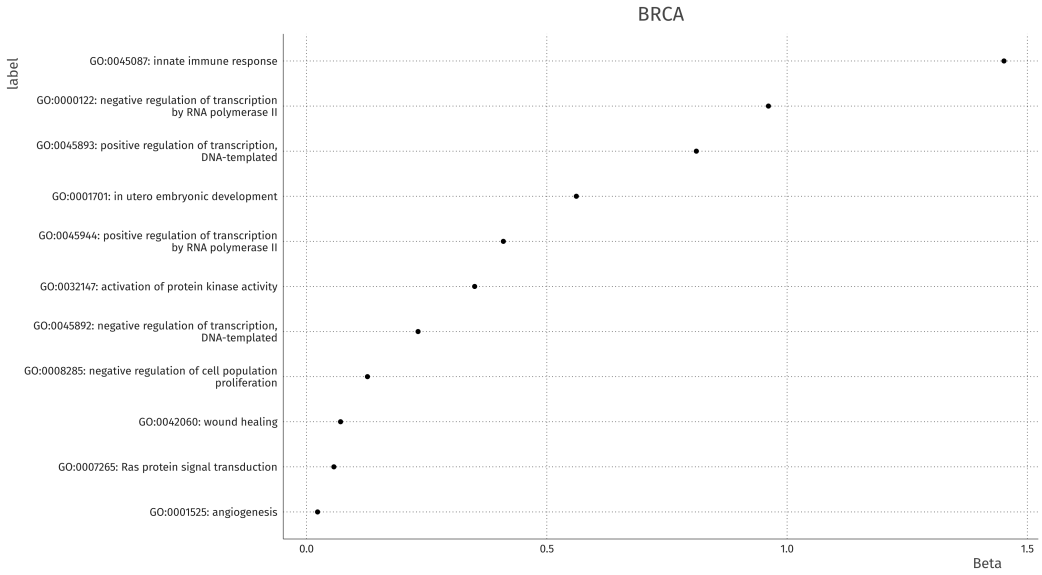

Figure 2.5: Joint enrichment estimate of the features ranked by enrichment for BRCA.

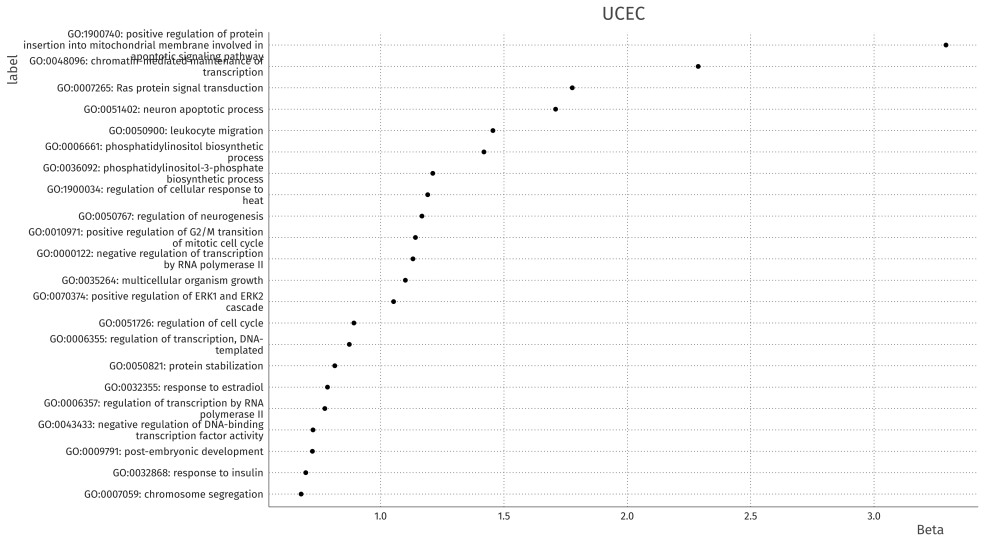

Figure 2.6: Joint enrichment estimate of the top 22 features (by absolute enrichment) for UCEC. There were in total 62 features with nonzero enrichment estimates in the final joint model for UCEC

CHAPTER 3

RSSP: COMPUTATIONALLY EFFICIENT, LIKELIHOOD-BASED ESTIMATION OF HERITABILITY FROM GWAS SUMMARY STATISTICS

3.1 Introduction

There is no concept more central to the study of genetics than that of heritability. In the simplest definition, a trait is heritable if some component of the phenotypic variation in the population is attributable to genetic variation[25]. Narrow-sense heritability, or h^2 is the proportion of a trait's variance that can be attributable to additive genetic variance. h^2 has traditionally been measured by studying twins or pedigrees, but can be biased when assumptions about the sources of phenotypic covariance are violated[47]. Over the last decade or so, SNP-based methods have been developed[89] and widely applied to the problem of h^2 estimation, based on large scale genome-wide association studies (GWAS). These methods leverage information of variants across the genome, rather than only strongly associated variants, which explain only a small fraction of heritability estimated from twin studies. The state of art method for estimating heritability using genotype and phenotype data of unrelated individuals is Genome-based restricted maximum likelihood (GREML)[89]. GREML has been used to great effect to explain a large proportion of heritability found by family studies [89]. These methods, however, requires individual-level data, which can be difficult to access, as well as the construction of an $n \times n$ relatedness matrix (with n being the number of individuals). The size of this matrix increases quadratically as sample size increases, raising substantial computational burden. The latest methods for heritability estimation often use GWAS summary statistics, in the form of the estimated effect sizes and standard errors, and p-values, of marginal association of each variant. Summary statistics have become the most common method for summarizing and storing estimated relationships between genotype and phenotype [56]. There are many advantages of working with summary statistics, comparing with in-

dividual level data, including: easier comparison of genetic-association signal across traits at a locus[12], across loci for a trait[64], and the comparison of patterns of genotype-phenotype associations across populations[69]. LD score regression is the most widely used method for estimating heritability from GWAS summary statistics. LD score regression uses GWAS summary statistics and an estimate of the linkage disequilibrium (LD) between those variants to estimate the heritability of the trait as well as the extent of confounding in the GWAS[10]. Other summary statistics based methods, e.g. LDAK-SumHer, allows more complex relationship of the effect sizes of variants and their minor allele frequencies (MAF) and LD structure.

All the summary statistics based methods for heritability estimation, however, are essentially method-of-moment estimators. These methods often use ad-hoc procedures to deal with dependency of nearby variants due to LD. A full likelihood based model for parameter estimation, while accounting for LD among variants, would be statistically optimal. Our work takes advantages of the likelihood model developed by our collaborators, known as Regression with Summary Statistics (RSS)[98], in modeling the relationship between marginal associations of single variants and the true effect sizes of all variants. Here we present a model with RSS likelihood with a normal prior distribution of effect sizes, which we call polygenic RSS, or RSSp, and we demonstrate a computationally efficient technique to make inference under RSSp. Additionally, we estimate heritability, not as a single unknown parameter, but using the sum of Percent of Variance Explained (PVE) of all individual SNPs, a strategy that makes the results less sensitive to the prior assumption. Using simulation, we show under truly polygenic genetic architectures, RSSp is able to better estimate heritability than LD score regression, even matching the performance of GREML that uses individual level data. We also explore the consequences of using an out-of-sample reference LD from external data to demonstrate the importance of accurate estimation of linkage disequilibrium and explore how shrinkage estimators of LD can improve estimates of heritability.

3.2 Methods

3.2.1 Background

The most common method for associating genotype with phenotype is through an additive model:

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

where $\mathbf{y}$ is the length n vector of phenotypes, X is an (n by p) matrix representing genotype, (which we will assume is centered for mathematical convenience, and without loss of generality), β is a vector (length p) of variant-level effects and ϵ is noise/error. In this model, the narrow sense heritability is defined as $h^2 = \frac{\text{Var}(\mathbf{X}\beta)}{\text{Var}(\epsilon)}$. With current GWAS sample sizes the number of variants is much larger than the number of samples $i.e. p >> n$, so it is difficult or even impossible to estimate β_j for each variant j, so β is often treated as a random variable, following some prior distribution. In a GWAS context $\mathbf{X}$ and $\mathbf{y}$ are fixed, which means that estimating h^2, *for the population from which the sample is drawn* can be reduced to estimating the distribution parameter(s) of β.

The most common means of summarizing the results of a GWAS is to use GWAS summary statistics. The summary statistics we will be interested in are the marginal effect size at a particular variant (j), which we will denote as $\hat{\beta}_j$, and the standard error of that estimate, which we will denote as $\hat{\sigma_j^2}$. The estimates and standard errors are usually from simple regression of Y against X_j, genotype of variant j:

$$\hat{\beta}_j := ({X_j}^T X_j)^{-1} {X_j}^T y$$

$$\hat{\sigma_j^2} := (n X_j^T X_j)^{-1} (y - X_j \hat{\beta}_j)^T (y - X_j \hat{\beta}_j)$$

3.2.2 Regression with Summary Statistics (RSS)

RSS relates marginal association statistics to effect sizes β by using the LD matrix:

$$\hat{\beta}|\beta \sim N(\hat{S}\hat{R}\hat{S}^{-1}, \hat{S}\hat{R}\hat{S})$$

where $\hat{\beta}$ is the length p vector of GWAS effect size estimates, $\hat{\mathbf{S}}$ is a diagonal matrix where $\hat{S_{j,j}} = \hat{\sigma_j^2}$, and $\hat{R}$ is the sample correlation matrix, (otherwise known as the LD matrix), assuming variants are normalized (i.e. mean 0, variance 1).

In the original RSS paper there were two priors on β that were discussed. The first is based on the Bayesian Sparse Linear Mixed Model (BSLMM)[97] where true effects (β) come from a mixture of sparse and polygenic components:

$$\beta_j \sim \pi N(0, \sigma_B^2 + \sigma_P^2) + (1-\pi)N(0, \sigma_P^2)$$

Here σ_B^2 represents the variance of the sparse component, while σ_P^2 represents the variance of the polygenic component. Fitting the RSS model with this prior is quite computationally demanding, as the MCMC requires computing the multivariate normal density function, which itself requires cholesky decomposition of a $p \times p$ matrix, an $O(p^3)$ operation. If one assumes that $\sigma_P^2 = 0$, i.e that there is no polygenic component, one arrives at the BVSR model:

$$\beta_j \sim \pi N(0, \sigma_B^2) + (1-\pi)\delta_0$$

The posterior for the BVSR model can be still difficult to compute, despite efforts using MCMC or Variational Bayes. If instead of assuming that $\sigma_P^2 = 0$, one assumes that $\sigma_B = 0$ (or equivalently that $\pi = 0$) we arrive at the following model:

$$\beta_j \sim N(0, \sigma_P^2)$$

With a normal prior (rather than a mixture of two normal distributions) and multivariate normal likelihood, we can write down the analytic form of the marginalized likelihood as shown below.

We start with some statistical background of multivariate normal distribution. If we know that the marginal Gaussian distribution for some variable $\mathbf{x}$ and a conditional Gaussian distribution for some $\mathbf{y}|\mathbf{x}$ of the forms:

$$p(\mathbf{x}) = N(\mathbf{x}|\mu, \Lambda^{-1})$$

$$p(\mathbf{y}|\mathbf{x}) = N(\mathbf{y}|A\mathbf{x} + \mathbf{b}, L^{-1})$$

then the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by [6]:

$$p(\mathbf{y}) = N(\mathbf{y}|A\mu + \mathbf{b}, L^{-1} + A\Lambda^{-1}A^T)$$

$$p(\mathbf{x}|\mathbf{y}) = N(\mathbf{x}|\Sigma\left\{A^T L(\mathbf{y} - \mathbf{b}) + \Lambda\mu\right\}, \Sigma)$$

where :

$$\Sigma = (\Lambda + A^T LA)^{-1}$$

Given this result, we can derive the marginal distribution of $\hat{\beta}$ and the posterior of β. Given the prior for β as $\beta \sim N(0, I_p\sigma_\beta^2)$, and that the RSS likelihood is $\hat{\beta}|\beta \sim N(\hat{S}\hat{R}\hat{S}^{-1}\beta, \hat{S}\hat{R}\hat{S})$, we can replace β with $\mathbf{x}$ and $\hat{\beta}$ with $\mathbf{y}$ by making the following substitutions:

Symbol	Replacement
μ	0
b	0
Λ^{-1}	$I_p\sigma_\beta^2$
A	$\hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}}^{-1}$
L^{-1}	$\hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}}$

We then see that the distribution of $\hat{\beta}$ after marginalizing over β is:

$$\hat{\beta}|\sigma_\beta^2 \sim N(0, \sigma_\beta^2\hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}}^{-1}\hat{\mathbf{S}}^{-1}\hat{\mathbf{R}}\hat{\mathbf{S}} + \hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}})$$

We can rewrite this as :

$$\hat{\beta}|\sigma_\beta^2 \sim N(0, \sigma_\beta^2\hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}}^{-2}\hat{\mathbf{R}}\hat{\mathbf{S}} + \hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}})$$

Computing the marginalized likelihood in this case, though involving only a single parameter, requires an expensive recalculation of the multivariate normal probability density function, in particular the re-computation of the determinant and inverse of $\sigma_\beta^2\hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}}^{-2}\hat{\mathbf{R}}\hat{\mathbf{S}} + \hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}}$ for each value of σ_β^2. A common computational trick for recomputing a multivariate normal density is to precompute a Cholesky decomposition of the covariance matrix, as the computationally expensive aspects of computing the multivariate normal density (in particular computing the determinant and inverse of the covariance matrix) have efficient implementations when the covariance matrix has been Cholesky-decomposed. This trick is unfortunately not applicable in this situation. Even if both $\hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}}^{-2}\hat{\mathbf{R}}\hat{\mathbf{S}}$ and $\hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}}^{-2}\hat{\mathbf{R}}\hat{\mathbf{S}}$ were factored separately, there is no way to add the two matrices to maintain the Cholesky form.

3.2.3 *Polygenic RSS (RSSp)*

If instead of modeling $\hat{\beta}$ and β, we model $\hat{\mathbf{u}}$ and $\mathbf{u}$, where $\hat{u}_j = \frac{\hat{\beta}_j}{\hat{\sigma_j^2}}$, and $u_j = \frac{\beta_j}{\hat{\sigma_j^2}}$. This is effectively the same as effect-size in terms of standardized genotypes, and leads to the same (implicit) prior as LD score regression[10] and GCTA[89]. With the prior $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^2)$, the likelihood in terms of $\mathbf{u}$ becomes:

$$\hat{\mathbf{u}}|\mathbf{u} \sim \mathcal{N}(\mathbf{Ru}, \mathbf{R}),$$

the marginal distribution of $\hat{\mathbf{u}}$ is:

$$\hat{\mathbf{u}}|\sigma_u^2 \sim N(0, \sigma_u^2\mathbf{R}^2 + \mathbf{R})$$

Now we can show that computation under this model can be made very efficient. If we take the eigenvalue decomposition of $\mathbf{R}$, $\mathbf{R} = \mathbf{QDQ}^T$, with $\mathbf{Q}$ being the p by p matrix of eigenvectors, and $\mathbf{D}$ being the diagonal matrix of eigenvalues (such that $D_{ii} = \lambda_i$, where λ_i is the i-th eigenvalue), we can rewrite the marginalized likelihood as $\hat{u} \sim N(0, \sigma_u^2\mathbf{QD}^2\mathbf{Q} + \mathbf{QDQ})$. Letting $\mathbf{v} = Q^T\mathbf{u}$, and $\hat{\mathbf{v}} = Q^T\hat{u}$, and exploiting the property of the multivariate normal distribution that an affine transformation of a multivariate normal random variable has a multivariate normal distribution, the likelihood for $\hat{\mathbf{v}}$ is $\hat{\mathbf{v}}|\sigma_u^2 \sim \mathcal{N}(0, \sigma_u^2\mathbf{D}^2 + \mathbf{D})$. As all the off-diagonal terms of the covariance matrix for $\mathbf{v}$ are 0, we can equivalently write the likelihood for $\hat{\mathbf{v}}$ as the joint probability of p independent univariate normal variables where:

$$\hat{v}_j \sim N(0, \sigma_u^2\lambda_j^2 + \lambda_j)$$

Computing the marginalized-likelihood estimate of σ_u^2 in terms of $\hat{\mathbf{v}}$ and $\mathbf{D}$ can be done without the costly covariance inverse or determinant calculations required in the original multivariate probability density function. We estimate σ_u^2 by the "Brent" 1D optimization routine in R[7]. While this procedure requires a costly eigendecomposition of R, it only needs to be

done once, without the need of updating it during optimization. Furthermore, R would be the same across all complex traits, so the eigendecomposition can be pre-computed and stored, further saving computational time.

Heritability is closely related to the parameter σ_u^2, which is interpreted as average Percent of Variance Explained (PVE) by a single SNP. We obtain the point estimate by maximizing $P(\hat{v}|\sigma_u^2)$. Instead of using this point estimate, we estimate h^2 by summing over the estimated PVE of each SNP. This estimate, introduced by Zhu and Stephens [98], was known as "summary PVE" (SPVE). Specifically, given true effect sizes of all SNPs, it is defined as:

$$\text{SPVE}(v) = \frac{\text{Var}(X\beta)}{n\text{Var}(y)} = \frac{1}{n}\text{Var}(\tilde{X}u) = \frac{1}{n}u^T\tilde{X}^T\tilde{X}u = \frac{1}{n}u^TRu = \frac{1}{n}v^TDv \tag{3.1}$$

where n is sample size and $\tilde{X}$ stands for normalized genotypes. Furthermore, we assume the variance of y is 1. Of course, we cannot use this formula since v is unknown. But we can compute its posterior expectation. Suppose we know σ_u^2 (using point estimate by maximizing marginal likelihood), we can derive the posterior distribution of v using properties of multivariate normal distribution:

$$v|\hat{v},\sigma_u^2 \sim N(\mu,\Sigma) \text{ where } \mu = \Sigma\hat{v}, \Sigma = \left(D + \frac{1}{\sigma_u^2}I\right)^{-1} \tag{3.2}$$

The posterior mean of SPVE is then given by:

$$\text{E}(\text{SPVE}(v)|\hat{v},\sigma_u^2) = \frac{1}{n}\text{tr}(D\Sigma) + \frac{1}{n}\mu^TD\mu \tag{3.3}$$

We can now plug in μ and Σ, we have:

$$\text{E}(\text{SPVE}(v)|\hat{v},\sigma_u^2) = \frac{1}{n}\left[\sum_{j=1}^{p}\frac{1}{1+\frac{1}{\sigma_u^2\lambda_j}} + \sum_{j=1}^{p}\frac{\hat{v}_j^2}{\lambda_j\left(1+\frac{1}{\sigma_u^2\lambda_j}\right)^2}\right] \tag{3.4}$$

3.2.4 Linkage disequilibrium

RSSp requires the LD matrix of variants. When individual level data is available, we can compute the "in-sample" LD using individual genotype data. When this is not available, it is common to compute the out-of-sample LD from the "reference-panel", which is the set of individuals not used in generating GWAS summary statistics, but from the same population as the samples where GWAS summary statistics were derived. Rather than computing and storing 8 million by 8 million genome-wide LD matrix, a blockwise diagonal approximation to the LD matrix was made. LD between variants was only estimated for variants within 1,703 approximately independent LD blocks. The boundaries of independent LD blocks were previously identified in the 1000 genomes EUR population using the method ldetect[5].

3.2.5 LD shrinkage estimators

To attempt to improve the estimate of LD, I used the LD shrinkage estimator developed by Wen and Stephens[88], which uses an estimate of the recombination rate, as well an estimate of the effective population size to improve the estimate of correlation between variants. I implemented a standalone version of this shrinkage estimator in an R package called `LDshrink`[80].

If X is a $n \times p$ matrix of genotype dosages, such that $X_{i,j}$ represents the number of effect alleles at the \$j\$th variant in the \$i\$th individual, and $\hat{\Sigma}$ is the $p \times p$ the estimate of covariance between variants, where:

$$\hat{\Sigma} = (1-\theta)^2\mathbf{S} + \frac{\theta}{2}\left(1-\frac{\theta}{2}\right)\mathbf{I}$$

, where

$$S_{jk} = \begin{cases} \text{Cov}(X_{.j}, X_{.k}), \text{if } j = k \\ e^{-\frac{\rho_{jk}}{2n}}\text{Cov}(X_{.j}, X_{.k}), \text{otherwise} \end{cases}$$

and ρ_{jk} is an estimate of the population-scaled recombination rate between variants j and k.

After shrinkage, the covariance matrix is converted to a correlation matrix, and values below a given threshold (by default 0.001), are rounded down to 0 to induce sparsity in the resulting LD matrix.

Populating the ρ parameter requires an estimate of the population-scaled recombination rate at the the sites in the simulation. The method `pyrho`[78] is a fast, demography-aware method for inference of fine-scale recombination rates, and is based on the fused-LASSO. The `pyrho` method applied to the British in England and Scotland (GBR) individuals from the 1000 genomes project[19] were used in the estimation of the local recombination rate for the UK biobank simulations.

3.3 Results

3.3.1 Assessing heritability estimate using GWAS simulations under infinitesimal model

To estimate the effectiveness of RSSp at estimating heritability, we simulated phenotypes and estimated GWAS summary statistics. We then employed state of the art individual-level data-based (`GCTA`[89]) and GWAS summary statistics-based (`ldsc`[10]) heritability estimation methods. To make the simulations as realistic as possible, real individual-level genotypes were used.

Genotype data from individuals from the UK biobank were used as the basis of simulations. Two random non-overlapping subsets of 10,000 unrelated individuals were randomly drawn (without replacement) from the 487,409 total individuals in the UK biobank dataset. Both datasets were subset to include only the variants with allele frequency > 1% in both subsets, resulting in 8,327,757 variants in total. Causal variant effects and phenotypes traits were simulated using a modified version of the `simu` software[21], a tool for simulating GWAS phenotypes based on real genotype data. `simu` simulates causal effects from a normal distribution

and uses the GCTA model of scaled genotypes. For the infinitesimal simulation, 80 traits were simulated for 8 h^2 values from 0.1 to 0.8 in increments of 0.1, with 10 trait replicates being simulated at each level of heritability. After simulating the phenotype, GWAS summary statistics for each simulated phenotype were then generated using GCTA's implementation of the `fastGWA` mixed linear model-based GWAS method[45], with the first 10 principle components used as quantitative covariates.

Heritability was estimated using GCTA's GCTA-GREML analysis, using a GRM constructed using the 8,327,757 variants and using 10 principal components as quantitative covariates. To estimate heritability using LD score regression, LD scores were generated on the 8,327,757 variants. This step either uses the genotypes of individuals used for the GWAS simulation ("in-sample"), or the second, equally-sized, samples (the LD reference panel) not used in simulations. LD scores were estimated using `ldsc`, using a 1 centimorgan sliding window, as per the `ldsc` tutorial on the `ldsc` website[9]. The `pyrho` method applied to the British in England and Scotland (GBR) individuals from the 1000 genomes project[19] was used in the estimation of the local recombination rate. As we also wished to test for the effect of out-of-sample LD vs in-sample LD, and test LD score regression with a varying intercept and without a varying intercept, we reran LD score regression with and without a varying intercept, and with in-sample LD scores and with reference panel LD scores. Finally, RSSp was fit on the data. One LD matrix per independent block of the genome was estimated(in-sample and out-of-sample) and diagonalized as per described in Section 3.2.4.

Estimates of h^2 from the individual-level data method GCTA were on average the closest to the true value of h^2, and is largely unbiased across every simulated value of h^2 (Figure 3.1). RSSp estimates are close to GCTA estimates with only slightly larger standard errors (Figure 3.1). In contrast, `ldsc` estimates, with fixed or free intercept term, show significant bias across all settings, and are much worse than both GCTA and RSSp (Figure 3.1). We evaluated the results of summary statistics using out-of-sample LD. The results of RSSp remain largely

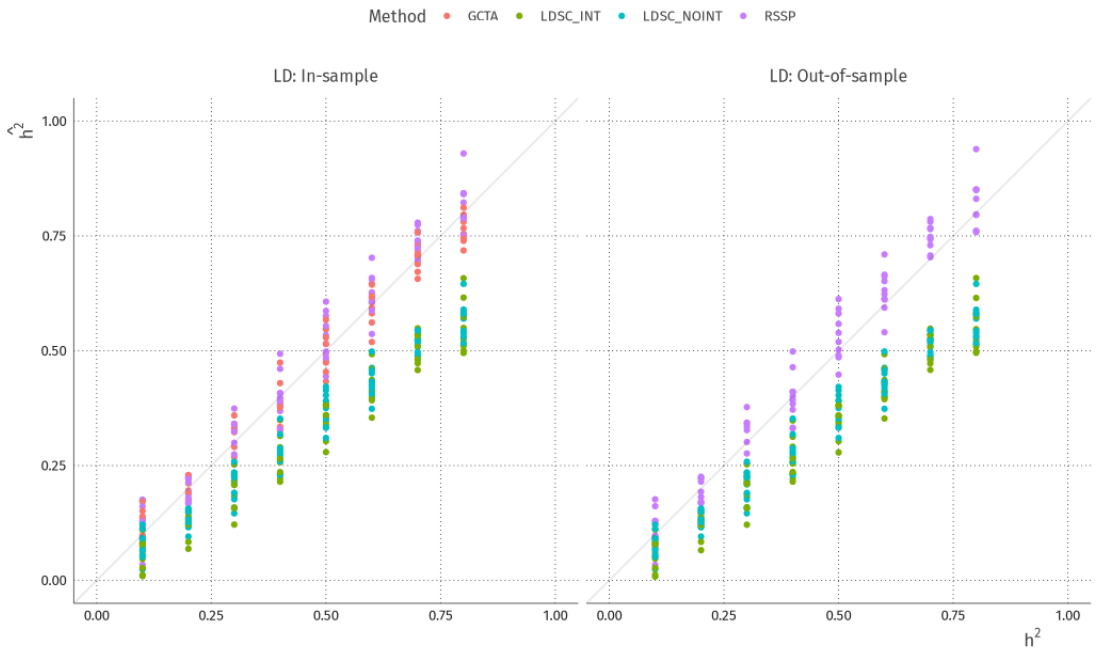

Figure 3.1: $\hat{h^2}$ across 10 replicates across 8 values of h^2 for 4 methods of heritability estimation, estimated using either in-sample LD (left panel), or an external reference LD (right). The light grey line along the diagonal represents a perfect 1 to 1 relationship between $\hat{h^2}$ and h^2. The heritability estimation methods are as follows: `GCTA` is the individual-level data single-component GREML, with 10 principal components used as continuous covariates (As GCTA requires individual level data, there can be no out-of-sample result); `LDSC_INT` is LD score regression [10] run with the default setting where the intercept term is allowed to vary; `LDSC_NOINT` is LD score regression run with a fixed intercept of 1; `RSSP` is RSSp (fit without a shrinkage estimator of LD)

unbiased and similar to the results using in-sample LD, while `ldsc` performs more poorly (Figure 3.1, right). We further quantified the performance of each method under each setting by both bias and mean squared error (MSE) of h^2 estimates. These results show clearly that RSSp performance is comparable to GCTA and significantly better than `ldsc`. (Table 3.1).

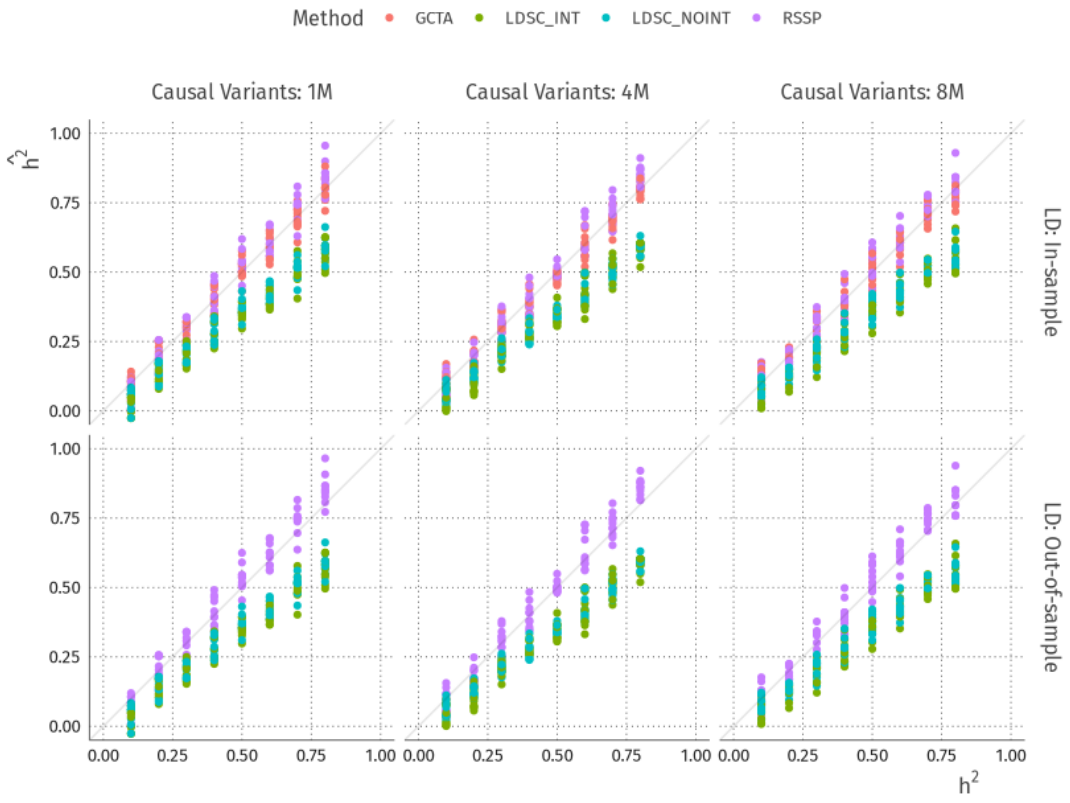

Figure 3.2: $\hat{h^2}$ across 8 values of h^2 with 10 replicates each for for methods of heritability estimation. $\hat{h^2}$ was estimated using either in-sample information (top panels), or an external reference LD panel (bottom panels). Two sparse genetic architectures are represented along side the original infinitesimal model. From left to right, causal variants contributing to h^2 numbered 1 million — for a sparsity of 87.5%, 4 million — for a sparsity of 50%, or 8 million, for a sparsity of 0%.

LD	Method	Bias	RMSE
In-sample	GCTA	-0.0027508	0.0388235
In-sample	RSSP_NOSHRINK	0.0122161	0.0483802
Out-of-sample	RSSP_NOSHRINK	0.0161776	0.0506190
In-sample	RSSP_LDSHRINK	0.0240206	0.0553949
Out-of-sample	RSSP_LDSHRINK	0.0260170	0.0568414
In-sample	LDSC_NOINT	-0.1280450	0.1461858
Out-of-sample	LDSC_NOINT	-0.1280562	0.1461914
In-sample	LDSC_INT	-0.1413375	0.1587786
Out-of-sample	LDSC_INT	-0.1416675	0.1591236

Table 3.1: Comparison of methods for heritability estimation on 8M variant simulation. Bias was calculated over all simulated values of h^2 Bias$(h^2, \hat{h}^2) = \frac{1}{80}\sum_{i=1}^{10}\sum_{j=1}^{8}(\hat{h^2}_{i,j} - {h^2}_j)$.

3.3.2 Assessing heritability estimation under varying levels of causal variant sparsity

To assess the robustness of RSSp to model misspecification, we performed simulations which relaxed the infinitesimal assumption. Using the 8,327,757 variant, 10,000 sample UK biobank dataset, we expanded our simulations to include sparse genetic architectures. Starting with the original 8,327,757 variants, we created two randomly selected causal variant subsets, containing 4,163,878 and 1,040,970 variants. Using these causal variant subsets, we simulated phenotypes under values of h^2 from 0.1 to 0.8 in increments of 0.1, from which we generated GWAS summary statistics on the original 8,327,757 variants. For each value of h^2, and each of the two causal variant subsets we ran 10 simulations. We then estimated heritability with each of the methods described previously.

We found RSSp, and indeed all three methods to be quite robust to varying levels of sparsity at the three levels simulated (Figure 3.2). These results are consistent with earlier papers, which reported that models based on infinitesimal assumption can often produce unbiased estimates despite the violation of this assumption [10].

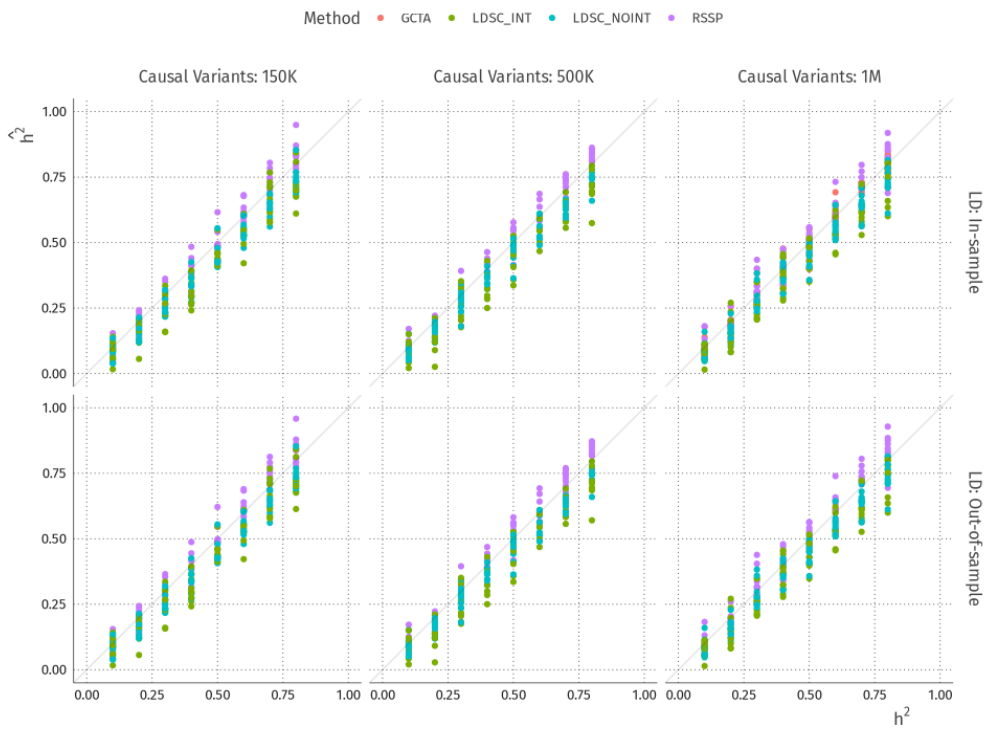

Figure 3.3: Replicate of original simulation scheme using the approximately 1 million HapMap 3 variants. From left to right, simulations with decreasing levels of sparsity with the proportion of variants contributing to h^2 at a sparsity of 87.5%, 50% and 0%. Note that for this simulation GCTA results are only shown for 0% sparsity.

Assessing heritability under simulations with a smaller number of total variants

We hypothesized that the relatively poor performance of LD score regression was due to the simulation involving a very large number of variants in the GWAS relative to the sample size. It is recommended that users of `ldsc` subset their GWAS summary statistics to the variants in the HapMap 3 reference panel, which contains only approximately 1 million common (in Europeans) variants[20][9]. To test this hypothesis we repeated our simulation scheme on a reduced total number of variants; an approximately 1 million variant subset of our original 8 million variant subset that overlapped with the HapMap3 variants. We simulated GWAS, recomputed LD, LD scores, and recomputed the GRM using only the HapMap3 variants, and

used the same range of h^2 for our phenotype simulations. We simulated GWAS where 150,000 of the ~ 1M variants were causal, and GWAS where 500K of the ~ 1M variants were causal, in addition to our infinitesimal simulation where all 1M variants were causal. We found that while the performance of LD score regression improved under the 1M total variant simulation, it still showed a slight downard bias, and underperformed both GCTA and RSSp in terms of bias and variance (Figure 3.3). Similar to the previous simulation with 8 million total variants, none of the three methods showed a marked decrease in performance when the proportion of causal variants decreased.

3.4 Discussion

In this study we show how SNP-heritability can be estimated using GWAS summary statistics and an out-of-sample reference LD panel with accuracy comparable to individual-level data methods, even when the number of variants exceeds the number of samples by several orders of magnitude.

While we have shown that accurate estimation of SNP-heritability from GWAS summary statistics using an out-of-sample reference LD panel, there are some important caveats and future directions. First, though there was no sample overlap between the GWAS cohort and the LD reference panel, both the cohort and the panel were randomly sampled from the same set of individuals. The set of variants used to simulate the GWAS were selected so that they were above 1 percent frequency in both the GWAS sample and in the LD reference panel, and the size of the reference panel was the same size as the GWAS panel. If, for example, the 503 individuals in the European subset of the 1000 genomes dataset were used as a reference LD panel, the match between the datasets would have likely been poorer.

In our simulations, we found that LD score regression is biased under truly infinitesimal genetic architectures. Even with the intercept fixed at 1, we found in our simulations that LD score regression consistently underestimates h^2, and the amount LD score regression under-

estimates h^2 increases as h^2 increases. This is consistent with a previous study that showed through simulation a downward bias in LD score regression heritability estimates[29]. Other studies[76], including the original LDSC paper, however, have not reported such downward bias. It is unclear what may explain such inconsistent findings, and this may depend on specific simulation settings. Our study has a large number of variants (8M) comparing with earlier studies. This reduces per SNP heritability/effect and may pose challenge to LDSC. We did find that the bias tends to be smaller with fewer variants at a given h^2 (data not shown). Alternatively, the downward bias may be related to estimation of LD scores. One possibility is that by using only variants within 1cM to calculate LD (the default setting in LDSC), we may miss some longer-range LD. However, we found that changing 1cM to 10cM makes no difference (data not shown). Another possible explanation is that the estimation errors, instead of bias, of LD scores may lead to reduced h^2 estimates. It is well known from statistical literature that measurement errors in explanatory variables, LD scores in the case of LDSC, lead to downward bias, known as attenuation bias, in the estimate of regression coefficient. While estimation of LD involving common variants (AF > 5%) is generally accurate, the errors may be higher for rarer variants (AF from 1% to 5%), which constitute the majority of variants in our simulation data. The small sample size (10K) in our simulation likely makes the problem worse. Additional simulations would be needed to test these possible explanations.

Previously studies had found that chunking the genome into chunks can introduce upward bias in heritability estimates, even when using in-sample LD estimates[41]. In our analysis, the sub-chromosome, block-wise approximation to the full LD matrix did not introduce significant upward bias in RSSp's heritability estimates. In Hou et al, significant upward bias of heritability estimates were observed with block sizes as large as 4.3 megabases (when analyzing simulations based on a single 34 megabase chromosome). The median block size of the ldetect LD blocks is much smaller, at approximately 1.5 megabases, and yet RSSp did not show appreciable inflation. The fact that RSSp performs well under the block-diagonal approxima-

tion to the genome-wide LD matrix allows for scalable estimate of SNP heritability on very large numbers of variants, as the computational complexity of diagonalizing the LD matrix is determined by the size of the largest LD block, not on the total number of variants. In addition, as diagonalization can be computed in parallel across independent LD blocks, and as it only needs to be computed once per reference LD panel, We believe RSSp is especially well-suited to analysis of bio-bank data where multiple phenotypes are measured in the same set of individuals.

CHAPTER 4

DISCOVERING RISK GENES OF PRE-TERM BIRTH VIA FINE-MAPPING GENOME-WIDE ASSOCIATION STUDY SUMMARY STATISTICS [1]

4.1 Introduction

Spontaneous preterm birth (PTB), defined as spontaneous labor and birth before 37 weeks of gestation, is a leading cause of infant mortality and morbidity. While PTB is widely believed to have a genetic component, the broad etiologic heterogeneity and contribution of environmental factors have frustrated efforts to identify causal genes and characterize their mechanisms[24]. Recently, several loci have been linked to the risk of PTB and gestation length in a large genome-wide association study (GWAS)[93]. However, the causal variants and their target genes remain to be detected.

Indeed, identifying causal variants driving association signals is a common problem in post-GWAS analysis. Because of extensive linkage disequilibrium (LD) in human genome, a single causal variant can generate associations in many nearby SNPs in LD. One way to address this challenge is to identify functional variants in trait-associated loci, which are more likely to be causal variants. This effort has been largely focused on annotating regulatory functions of non-coding genome. Surveys of GWAS associations have found that the majority of GWAS signals come from non-coding regions. It is believed that the majority of this functional variation is regulatory in nature. Rather than modifying the function of a gene product directly by modifying its structure, regulatory variants alter the activity of a gene-product by modulating the abundance of the gene-product through regulatory mechanisms — by increasing or decreasing the baseline rate of transcription initiation of the gene's DNA to RNA, or by acting downstream of transcription initiation. Given genomic and epigenomic annotations, candi-

1. Much of this chapter contains material from the paper: Transcriptome and regulatory maps of decidua-derived stromal cells inform gene discovery in preterm birth (doi.org/10.1101/2020.04.06.017079), which is currently in press in Science Advances.

date variants in trait-associated loci can be prioritized based on their locations in regulatory elements. A major challenge in this approach comes from the fact that the regulation of transcription is highly dynamic across tissue, across cell-type, and across environmental stimuli. To help bridge this gap, large consortia, such as ENCODE, have generated a wealth of annotations of putative gene regulatory elements across a diverse set of human cell types and tissues. Using these resources, methods have been developed to identify putative causal variants from GWAS loci, and been successfully used to study traits ranging from type 2 diabetes[84] to blood cell traits[57].

However, there is a paucity of genomic and epigenomic data available for tissues related to pregnancy in general and to PTB in particular. To fill this gap in knowledge, we characterized the transcriptional and chromatin landscapes of cultured mesenchymal stromal cells (MSCs) collected from human placental membranes and decidualized MSCs, also known as decidual stromal cells or DSCs. These cells — collected from women following both term and preterm pregnancy — play critical roles in promoting successful pregnancy, interfacing with fetal cells throughout pregnancy, and the timing of birth. We then built a computational framework that integrated these decidua-derived stromal cell annotations with the results of a large GWAS of gestational duration to facilitate discovery of PTB genes.

This integrated analysis revealed a significant enrichment of heritability estimates for gestational duration in decidua-derived stromal cell genomic regions marked by open chromatin or histone marks. Leveraging those functional annotations in a Bayesian statistical framework, we discovered additional loci associated with gestational duration and improved fine mapping in regions associated with gestational duration. Finally, using promoter capture Hi-C (pcHi-C), we linked functionally-annotated gestational age-associated variants to their putative target genes. More generally, these functional annotations and our analytic pipeline should prove a valuable resource for studying other pregnancy-related conditions, such as preeclampsia and recurrent miscarriage, as well as conditions associated with endometrial

dysfunction, such as endometriosis and infertility.

Statistical fine-mapping analyses, including our own, are focused on identifying individual causal variants. Despite the best efforts, it is generally difficult to narrow down to a single or few causal variants in trait-associated regions. For example, in a high-powered GWAS of type 2 diabetes (T2D), only in a quarter of associated loci, the credible sets (the union of all SNPs that, with high probability, contains the causal signal), have less than 10 SNPs[57]. One way to address this challenge is to combine signals from SNPs likely targeting the same genes. The intuition is that, after fine-mapping a locus, we may still have a large uncertainty of exact causal variant, but if most of candidate variants target the same gene, then that gene is likely to be the risk gene at the locus. This strategy will both increase the statistical signal and provide more interpretable results. We thus developed a computational procedure that summarizes statistical evidence from GWAS fine-mapping at the level of genes. Applying this procedure to GWAS of gestation length, we identified some highly plausible genes, such as HAND2 and WNT4, the transcription factors that are believed to be important for differentiation of pregnancy related cell types. Interestingly, the genes from this analysis are significantly enriched with those differentially expressed after treating cells with pregnancy related stimuli, including cAMP and Trophoblast condition medium (TCM), providing evidence that disruption of transcriptional response to these stimuli is one mechanism of gestation length variation and pre-term birth.

4.2 Materials and Methods

4.2.1 Functional Genomic Data in MSCs

Placentas were collected from six African American women (≥ 18 years old) following spontaneous labor. Three of the women delivered at term (≥ 37 weeks), and three delivered preterm. All were vaginal deliveries of singleton pregnancies. Within 1 hour of delivery, 5 × 5 cm pieces

of the membranes were sampled from a distant location of the rupture site. Pieces were placed in DMEM-HAMS F12 media containing 10% FBS and 1% pen/strep. Samples were kept at 4°C and processed within 24 hours of tissue collection.

Primary MSC were derived from three women who delivered at term and three who delivered preterm using cells isolated from the decidua parietalis. To model the process of decidualization, cells were treated with medroxyprogesterone acetate (MPA) and cAMP for 48 hours and a paired set of untreated samples was cultured in parallel for 48 hours. To model the trophoblast invasion process, cells were treated with Trophoblast Conditioned Medium (TCM).

Three replicates of of each cell line were studied to assess experimental variability in the three conditions. Each of the 27 samples (3 individual lines x 3 replicates x 3 conditions) were assayed to generate transcriptomes (RNA-seq). Open chromatin (ATAC-seq) was assayed for the decidualized cells and the TCM treated samples, and histone modification (ChIP-seq) maps for H3K27ac, H3K4me1 and H3K4me3 marks were assayed in the control and decidualized cells.

Chromatin interaction was measured using promoter capture Hi-C in cultured primary decidua-derived mesenchymal stromal/stem cells (MSCs) and in vitro differentiated decidual stromal cells (DSCs) as well as in TCM treated cells.

4.2.2 Detecting differential gene expression

We used the the pseudo-alignment tool `salmon`[65] to obtain transcript-level abundance estimates (using gencode 19 as a source of transcripts). The abundance estimates were loaded into R using the `tximeta` package,[55] which was also used to summarize the transcript-level abundance estimates into gene-level abundance estimates. Genes with counts lower than 10 were excluded from consideration for differential expression, as were genes for which the gene level abundance estimate was above zero in less than 5 samples. We then used the R

package `DESeq2`[54] to identify differentially expressed genes. To formally test the hypothesis that individuals with term births respond differently to either decidualization media or TCM as compared to individuals with preterm births, we test for an interaction between the preterm effect and each of the two treatment effects. By coding the term vs preterm status using a "sum" coding, a nonzero estimate of either of these interaction terms indicates that the response to treatment (i.e the change in gene expression) differs between term and preterm samples. To capture the individual level effects, instead of comparing to a particular individual, we again used a sum coding, meaning for each term and preterm, 1 covariate captures the difference between the individual 1 and individual three,and another captures the difference between individual 2 and individual three. Like passage number,which was also included in the `DESeq2` model, individual-level effects were captured in the model, but the significance of the effect-size estimates were not tested. For the main effect tests (i.e term vs preterm, control vs decidualized, and TCM vs decidualized) we used the Wald test functionality for null hypothesis significance testing. `DESeq2` includes composite null hypothesis testing functionality when using the Wald test; instead of testing against the null hypothesis that $\beta = 0$, one can test against the hypothesis that $|\beta| \leq \theta$ where θ is some threshold value. Rather than adding a fold-change cutoff on top of a test against an effect size of 0, with the composite test the FDR results remain interpretable: p-values and adjusted p-values correspond to the specific null hypothesis of interest. This composite null-hypothesis testing was used with a log fold-change threshold of 0.2.

4.2.3 Fine-mapping GWAS summary statistics using functional annotations

Fine mapping proceeded in three stages. In the first stage we partitioned the genome into 1,703 regions approximately independent regions using breakpoints derived from the `ldetect` method[5]. Next, we constructed a SNP-level prior probability of causality, informed by the functional genomic data. To estimate the functionally informed SNP-level prior, We employed

a Bayesian hierarchical model TORUS[87]. TORUS uses SNP-level annotations and GWAS summary statistics to estimate the extent to which SNPs with functional genomic annotations are likely to be causal for a trait of interest. TORUS takes as input GWAS summary statistics and genomic annotations, and for each annotation outputs enrichment estimates that correspond to estimates from a logistic regression: the additive change in log odds for a variant being causal, conditioned on all other annotations being held constant. We ran TORUS with the gestational age GWAS summary statistics and the reproducible H3K27ac and H3K4me1 peaks from the `dec`-treated samples, the pcHi-C contact regions, and the union of all ATAC-seq peaks to obtain enrichment estimates. A SNP-level prior was constructed from those enrichment estimates. Lastly, fine mapping was performed using a summary statistics-based version of the "Sum of Single Effects" model (SuSiE[85]). In the summary statistics-based version of SuSiE, the inputs are the GWAS summary statistics in a region, the SNP-level prior for every GWAS variant, and an estimate of the LD between variants. As an estimate of LD, we used the unrelated European individuals from the 1000 Genomes project as a reference panel. SuSiE (as implemented in the R package "susieR") was run on the 33 regions most believed to have one or more causal variants as estimated by TORUS. For each region, SuSiE was run with a uniform prior (default setting of SuSiE) and with an informed prior learned by TORUS. The parameter L of SuSiE (maximum number of causal variants) is set at 3 when running SuSiE. To be conservative, the pip for all SNPs in each region were multiplied by $1 - \text{FDR}_{\text{TORUS}}$ to approximate the TORUS posterior probability that the region contained at least one causal variant.

4.2.4 Gene-level summary of fine-mapping results

We developed a method for summarizing our fine-mapping results at the gene level. SNP-level PIPs were summarized as gene-level PIPs based on several possible mechanisms SNPs can target genes, while accounting for uncertainty of SNP-to-gene mapping. Intuitively, if we are confident that a SNP targets a gene, then we should assign the PIP of that SNP to that

gene. In most cases, we will have uncertainty of the target genes, so we should allocate the PIP of the SNP in a weighted fashion to putative target genes, with the weights specified according to possible biological mechanisms.

To infer causal genes, we denote Z_g as the indicator variable (unobserved) of whether gene g is a causal gene. Also denote D as all the GWAS data. Our goal is to estimate $P(Z_g = 1|D)$. If we assume there is a single causal variant in a region of interest, this probability can be related to PIPs of all SNPs by summing over the possible causal variant:

$$P(Z_g = 1|D) = \sum_i P(Z_g = 1|\gamma_i = 1) \cdot P(\gamma_i = 1|D) \tag{4.1}$$

where γ_i is the indicator of whether SNP i is the causal variant. In the equation, $P(\gamma_i = 1|D)$ is simply the PIP of SNP i, p_i, which is computed from fine-mapping analysis. The term $P(Z_g = 1|\gamma_i = 1)$ is the probability that gene g is the causal gene behind SNP i, assuming i is the causal variant. It can also be interpreted as the proportion of the PIP of SNP i assigned into gene g. To obtain this proportion, we denote w_{ig} as the probability that SNP i perturbs gene g. Intuitively, if a SNP is confidently assigned to a gene, e.g. it is in an exon, then $w_{ig} = 1$. We assume that each locus has a single causal gene, i.e. a causal SNP acts on the phenotype through one causal gene, then we should have the constraint that $\sum_g P(Z_g = 1|\gamma_i = 1) = 1$ for all SNPs. This leads to the following "normalization":

$$P(Z_g = 1|\gamma_i = 1) = \frac{w_{ig}}{\sum_g w_{ig}} \tag{4.2}$$

Our weighting scheme specifies w_{ig} based on several possible mechanisms a variant changes gene function. Specifically, if a SNP is in the 5′ UTR (or 2kb upstream of the 5′ UTR), 3′ UTR, or exon of a gene, then the weight is 1. A SNP that is within a promoter-capture HiC contact, (as called by CHiCAGO [14]) of the promoter of a gene is also assigned weight of 1. Note that it is possible that a SNP is in Hi-C contact with multiple genes, and at the same lies in the gene

boy (UTR or exon) of another gene, then each gene will receive weight of 1, and the PIP of that SNP will be equally partitioned among all genes, as described in Equation 4.2. For all variants that lay outside genes or HiC contacts, we assign the SNPs to target genes based on SNP-gene distances in a weighted fashion, with nearby genes receiving higher weights. This weighting scheme is supported by the observation that most target genes of enhancers are located close to the genes, typically within 100Kb[32]. For a given SNP, we consider all genes with 1Mb that are inside the LD blocks defined in fine-mapping. The weight of a gene decays exponentially with distance to the gene according to the following function

$$w_{i,g} = e^{\frac{-d_{i,g}}{100000}} \tag{4.3}$$

with $d_{i,g}$ being the distance between variant i and the TSS of gene g. We use a parameter of 100kb here so that overall genes within 100kb of enhancers receiving most of the weights, while still allowing long-range interactions. For example, a gene 50kb away from a SNP would have weight of 0.6, and a gene 200kb away has a weight of 0.14. The PIP of a SNP would then be partitioned among all nearby genes according to Equation 4.2.

4.3 Results

4.3.1 Fine-mapping Loci associated with Gestational Duration GWAS using Functional Annotations in Decidualized Stromal Cells

We developed a computational procedure to integrate the decidua stromal cell functional maps with genetic map of reproductive traits. We posited that integrating functional maps in these pregnancy-relevant cells and leveraging statistical methods to fine-map associations would result in 1) identifying candidate causal variants in each associated locus, 2) linking those variants to their target genes, and 3) discovering additional loci and genes associated

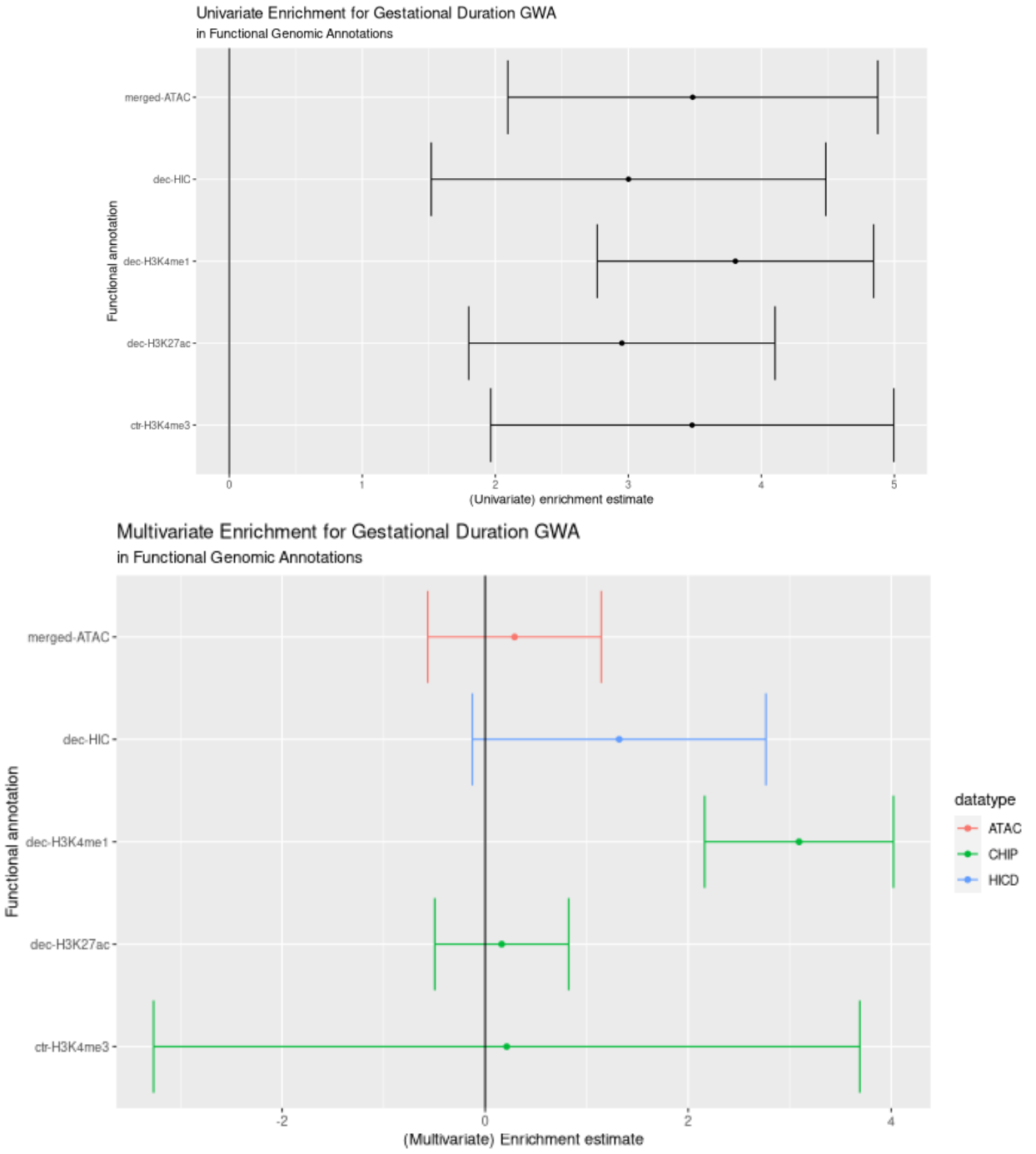

Figure 4.1: (a) Enrichment for gestational duration GWAS signal in regulatory maps of decidua-derived stromal cells. TORUS was run on each annotation separate. (b) Joint enrichment analysis for gestational duration GWAS signal in all annotations using TORUS.

with gestational duration.

We first leveraged the enrichments of DSC annotations to create Bayesian prior probabilities for a variant being causal. Using prior probabilities informed by functional annotations of SNPs could increase the accuracy of fine-mapping, as shown in recent studies (8, 41). We chose H3K27ac, H3K4me1, and pcHi-C interactions from the decidualized cells, and H3K4me3 from untreated cells, and ATAC-seq peaks from any of the cells as functional genomic annotations to create informative priors using TORUS[87]. To assign a prior to each SNP, TORUS uses genome-wide summary statistics of GWAS and the functional annotations to assess how informative each annotation is in predicting causal variants. TORUS analysis shows that the functional annotations are often enriched with GWAS signals (Figure 4.1). SNPs associated with functional annotations will then be assigned higher prior probabilities. Additionally, TORUS computes statistical evidence at the level of genomic blocks, defined as the probability that a block (determined by LD) contains at least one causal SNP. Without including any histone marks or chromatin accessibility annotations, TORUS implicated six autosomal blocks in the genome at FDR < 0.05, including five of the six genome-wide significant autosomal loci identified in the GWAS ($p < 5x10\text{-}8$). By including the functional genomic annotations from endometrial stromal cells, the number of high confidence blocks increased to ten, including all six that were significant in the gestational duration GWAS and four that were not significant in the GWAS.

We next performed computational fine mapping on the top these ten blocks, with the informative priors learned by TORUS, using SuSiE[85]. Conceptually, SuSiE is a Bayesian version of the step-wise regression analysis commonly used in GWAS (i.e. conditioning on one variant, and testing if there is any remaining signal in a region). SuSiE accounts for the uncertainty of causal variants in each step, and reports the results in the form of posterior including probabilities (PIPs). The PIP of a variant ranges from 0 to 1, with 1 indicating full confidence that the SNP is a causal variant. If a region contains a single causal variant, the PIPs of all SNPs in

the region should approximately sum to 1.

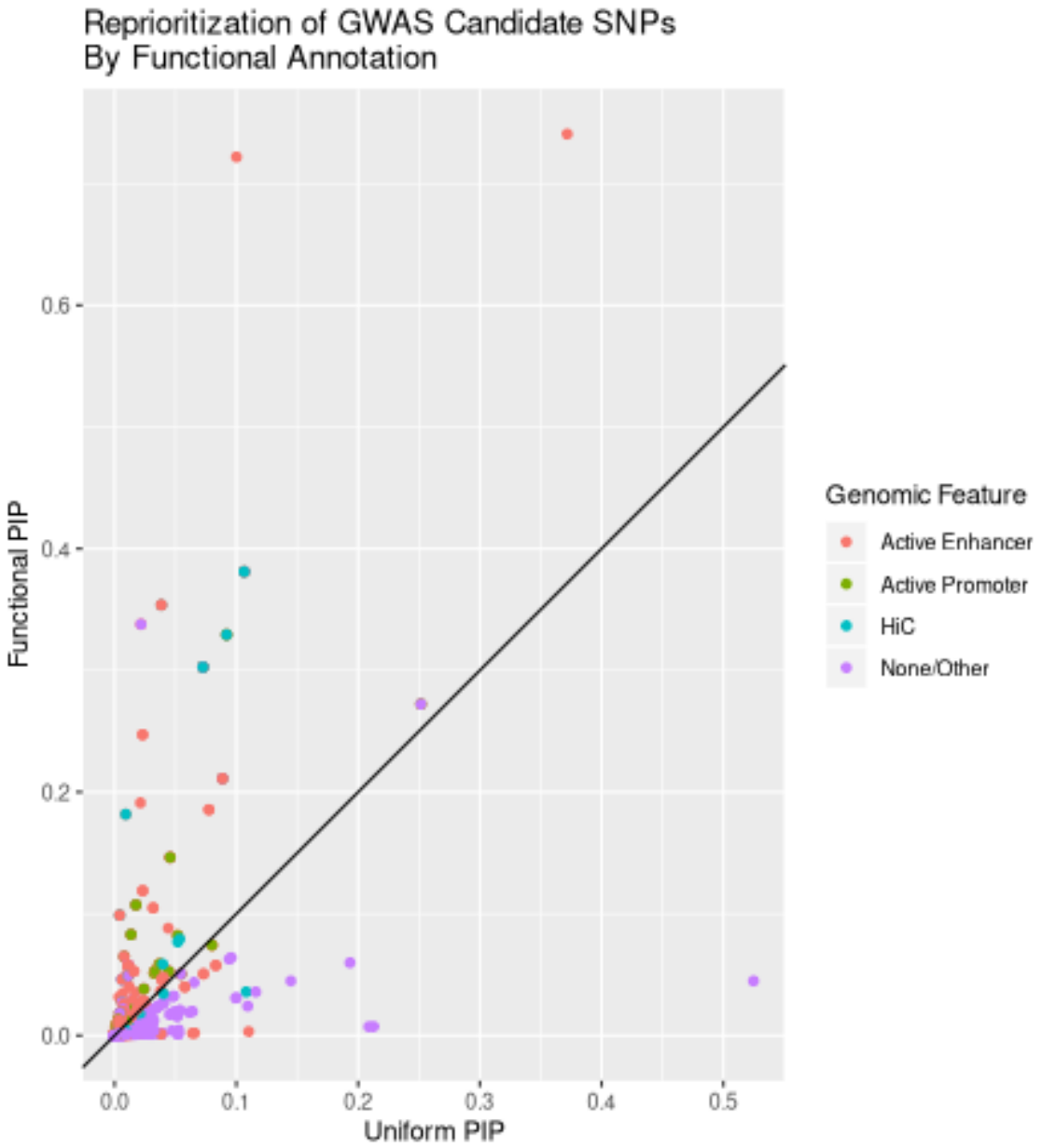

Figure 4.2: PIPs of SNPs using uniform vs. functional priors in SuSiE (each dot is a SNP). The functional prior of a SNP is based on SNP annotations and is estimated using TORUS. Figure originally published in Sakabe et al. 2020 [71].

Region	SNP	Location	GWAS p-value	Functional Prior	PIP	Functional annotations	Likely target
chr3:137.4-139.9M	rs147843771	chr3:138843356	3.8E-08	8.3e-05	0.74	K4me1	FOXL2
chr3:137.4-139.9M	rs17315501	chr3:139029676	1.7E-07	9.9e-05	0.21	K4me1, K4me3 ATAC, K27ac	
chr5:156.6-158.8M	rs2946164	chr5:157884706	3.0E-26	8.3e-05	0.72	K4me1	
chr4:174.3-176.6M	rs13141656	chr4:174728703	3.9E-07	5.1e-04	0.38	K4me1,K27ac K4me3, ATAC pcHi-C	HAND2 (pcHi-C)
chr4:174.3-176.6M	rs7663453	chr4:174729014	4.5E-07	5.1e-04	0.33	K4me1, K27ac K4me3, pcHi-C	HAND2 (pcHi-C)
chr2:73.2-75.6M	rs13387174	chr2:74206685	4.7E-07	3.8e-04	0.35	Hi-C, K4me1,K27ac	WBP1 (pcHi-C)
chr2:73.2-75.6M	rs13390332	chr2:74207357	2.e-07	9.9e-05	0.18	K4me1, K27ac	WBP1 (pcHi-C)
chr3:122.0-123.5M	rs4677884	chr3:123062970	4.1E-09	8.3e-05	0.34	K4me1, ATAC	
chr1:22.8-23.1M	rs56318008	chr1:22470407	2.3E-12	4.3e-04	0.3	K4me1, K4me3, ATAC, pcHi-C	WNT4 (promoter)
chr1:22.8-23.1M	rs55938609	chr1:22470451	2.3E-12	4.3e-04	0.3	K4me1, K4me3 ATAC, pcHi-C	WNT4 (promoter)
chr1:22.8-23.1M	rs3820282	chr1:22468215	6.4E-13	1.1e-04	0.27	K4me1, K4me3, ATAC	WNT4 (promoter)
chr3:154.7-156.0M	rs4679761	chr3:155868039	5.0E-09	9.9e-05	0.24	K4me1, K27ac	KCNAB1 (pcHi-C)
chr3:154.7-156.0M	rs9882088	chr3:155867092	5.5E-09	8.3e-05	0.19	K4me1	
chr3:126.2-128.2M	rs3122173	chr3:127889287	5.4E-12	3.2e-04	0.18	K4me1, pcHi-C	GATA2 (pcHi-C)
chr3:126.2-128.2M	rs2999048	chr3:127878416	2.0E-12	8.3e-05	0.12	K4me1, K27ac, pcHi-C	GATA2 (pcHi-C)
chr3:126.2-128.2M	rs1554535	chr3:127895986	1.2E-11	3.8e-04	0.10	K4me1, pcHi-C, K27ac	GATA2 (pcHi-C)

Table 4.1: Most probable SNPs identified from computational fine-mapping of regions associated with gestational duration

Table 4.1, continued: Most probable SNPs identified from computational fine-mapping of regions associated with gestational duration. Functional annotations are based on data from endometrial stromal cells. We list an annotation if the SNP is located in a sequence with that annotation in either untreated or decidualized condition. We list the pcHi-C annotation if the SNP is within 1 kb of a region involved in a pcHi-C interaction. We call a gene the target of a SNP if (1) the SNP is located in the promoter (< 1 kb of TSS) of that gene; or (2) the promoter of that gene has a pcHi-C interaction with a region within 1 kb of the SNP. In the case of rs147843771 at the FOXL2 locus, the target was defined by literature evidence[28].

Including the priors defined by TORUS using DSC functional annotations significantly improved fine-mapping (Table 4.1, Figure 4.2). For example, only one SNP reached PIP > 0.3 across all 10 blocks using the default setting under SuSiE (uniform prior, treating all SNPs in a block equally). This reflects the general uncertainty of pinpointing causal variants due to LD: e.g., a strong GWAS SNP in close LD with 9 other SNPs would have PIP about 0.1. By using the annotation-informed priors, 8 SNPs in six different blocks reached PIP > 0.3 4.2. Table 4.1 summarizes the most probable causal variants in eight blocks (fine-mapping in the remaining two blocks produced large credible sets with no high-PIP SNPs) as well as their likely target genes based on promoter assignment or chromatin interactions from pcHi-C. We note that our results of the WNT4 locus identified rs3820282 as the likely causal variant. This is consistent with our previous results demonstrating experimentally that the T allele of this SNP disrupts the binding of estrogen receptor 1 (ESR1)(5). This SNP was among the 3 most likely SNPs in our fine-mapping study, with a PIP of 0.27 (Table 4.1).

We highlight the results from two regions. In the first, two adjacent SNPs (311 bp apart), rs13141656 and rs7663453, on chromosome 4q34 did not reach genome-wide significance in the GWAS (p = 3.9 ×10-7 and 4.5 ×10-7, respectively). After using functional annotations in decidua-derived stromal cells, the block containing these SNPs was highly significant (TORUS q-value = 0.02), suggesting the presence of at least one causal variant in this block. The two SNPs together explained most of the PIP signal in the block (PIP 0.38 and 0.33, respectively, Table 1). The two SNPs are located in a region of open chromatin in endometrial stromal cells,

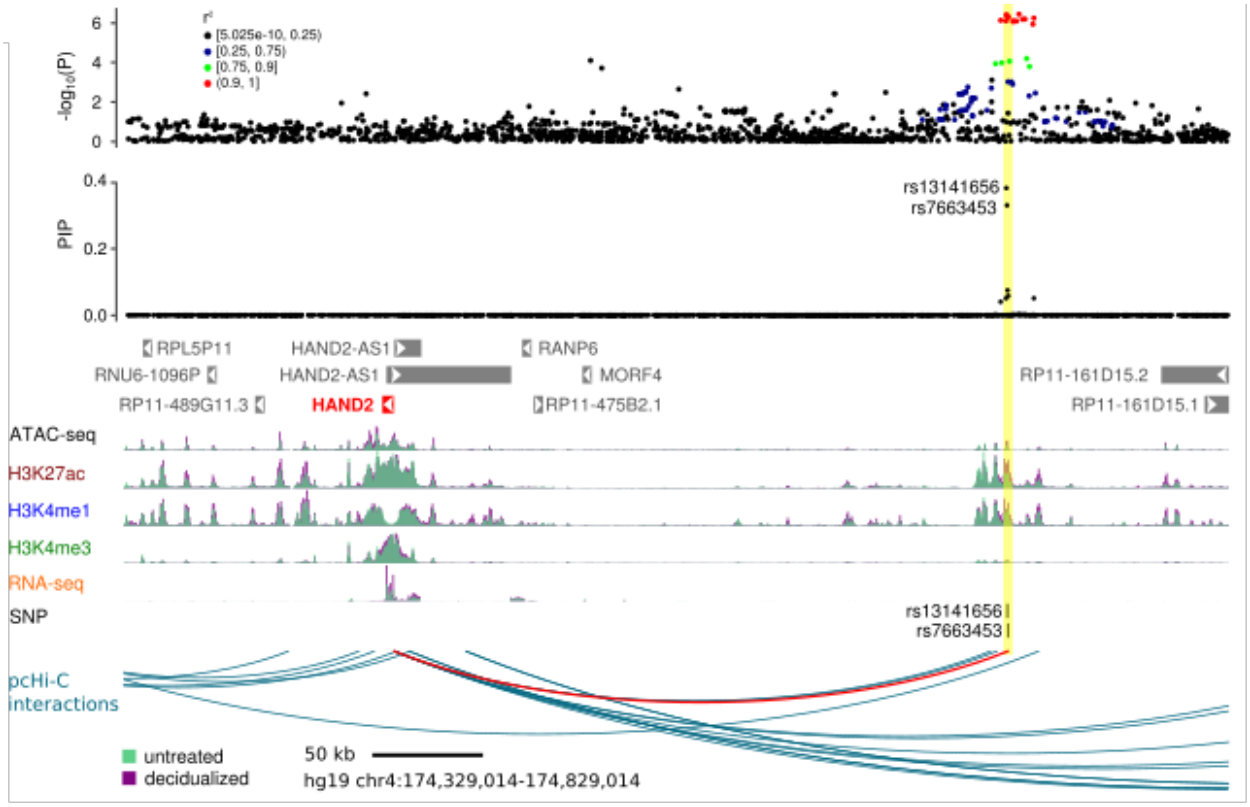

Figure 4.3: Likely causal variants near HAND2 and their functional annotations. The upper panel shows the significance of SNP association in the GWAS and the middle panel shows fine-mapping results (PIPs) in the region. The vertical yellow bar highlights the two SNPs with high PIPs. These SNPs are located in a region annotated with ATAC-seq, H3K27ac, H3K4me1 and H3K4me3 peaks (bottom). This putative enhancer also had increased ATAC-seq, H3K27ac and H3K4me1 levels in decidualized samples and interacts with the HAND2 promoter (red arc). Figure originally published in Sakabe et al. 2020 [71].

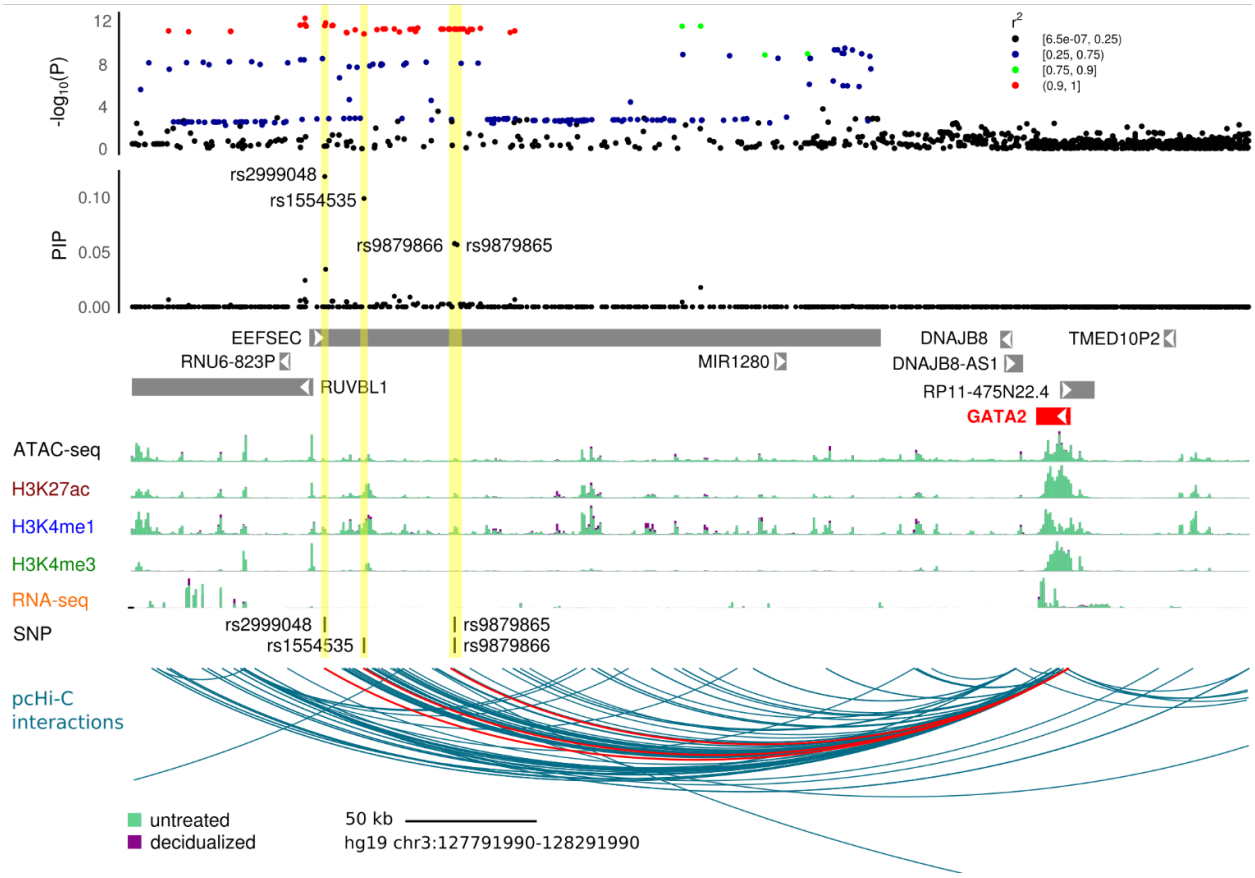

Figure 4.4: Fine-mapping a GWAS locus of gestational duration: likely causal variants near GATA2 and their functional annotations. The upper panel shows the significance of the SNPs in the GWAS and the middle panel shows fine-mapping results (PIPs) in the region. The vertical yellow bar highlights the four SNPs with high PIPs. These SNPs are located in a region annotated with ATAC-seq,H3K27ac, H3K4me1 and H3K4me3 peaks (lower panel). The sequences containing the four SNPs all interact with the GATA2 promoter (red arcs). rs2999048 is spanned by an H3K4me1 peak in 3/129 tissues of the Epigenome Roadmap data set whereas rs1554535 is not spanned by enhancer marks in any tissue. rs9879865 and rs9879866 are spanned by H3K27ac or H3K4me1 peaks in 24 and 26 tissues, respectively. Figure originally published in Sakabe et al. 2020[71]

with enhancer activity marked by both H3K27ac and H3K4me1 (Figure 4.3). Only 9 of the 129 tissues from the Epigenome Roadmap(11) also had H3K27ac, H3K4me1 or H3K4me3 peaks spanning the rs13141656 locus and only 2 spanning the rs7663453 locus. In addition, this putative enhancer is bound by multiple transcription factors, including GATA2, FOXO1, NR2F2 and PGR, based on ChIP-seq data. The only physical interaction of this enhancer in the pcHi-C data in decidualized stromal cells is with the promoter of the HAND2 gene, located 277 kb away (Figure 4.3). Summing over the PIPs of all SNPs whose nearby sequences interact with HAND2 via chromatin looping gives an even higher probability, 0.89, suggesting that HAND2 is very likely to be the causal gene in this region (Table 4.1). HAND2 is an important transcription factor that mediates the effect of progesterone on uterine epithelium[52]. Thus, in this example we identified a novel locus, the likely causal variant(s), the enhancers they act on, and an outstanding candidate gene for gestational duration and PTB.

The second example focuses on the locus showing a strong GWAS association with gestational duration on chromosome 3q21. The lead SNP, rs144609957 (GWAS $p = 4 \times 10^{-13}$), is located upstream of the EEFSEC (Eukaryotic Elongation factor, Selenocysteine-TRNA Specific) gene. There is considerable uncertainty of the causal variants in this region, with 50 SNPs in the credible set and the lead SNP explaining only a small fraction of signal (PIP = 0.02). Among all 12 SNPs with PIP > 0.01, 11 have functional annotations, most commonly H3K4me1 and pcHi-C interactions. Interestingly, for nine SNPs (first 3 shown in Table 4.1), the sequences in which they are located physically interact with the promoter of GATA2 in the pcHi-C data, but not with any other promoters in the region (Figure 4.4). The PIPs of all SNPs in the genomic regions that likely target GATA2 through chromatin looping sum to 0.68 (Table 4.1). Thus, despite uncertainty of causal variants in this region, our results implicate GATA2 as a candidate causal gene in endometrial stromal cells. GATA2 is a master regulator of embryonic development and differentiation of tissue-forming stem cells[31]. As support for the possible role of GATA2 in pregnancy, GATA2 deficient mice show defects in embryo implantation

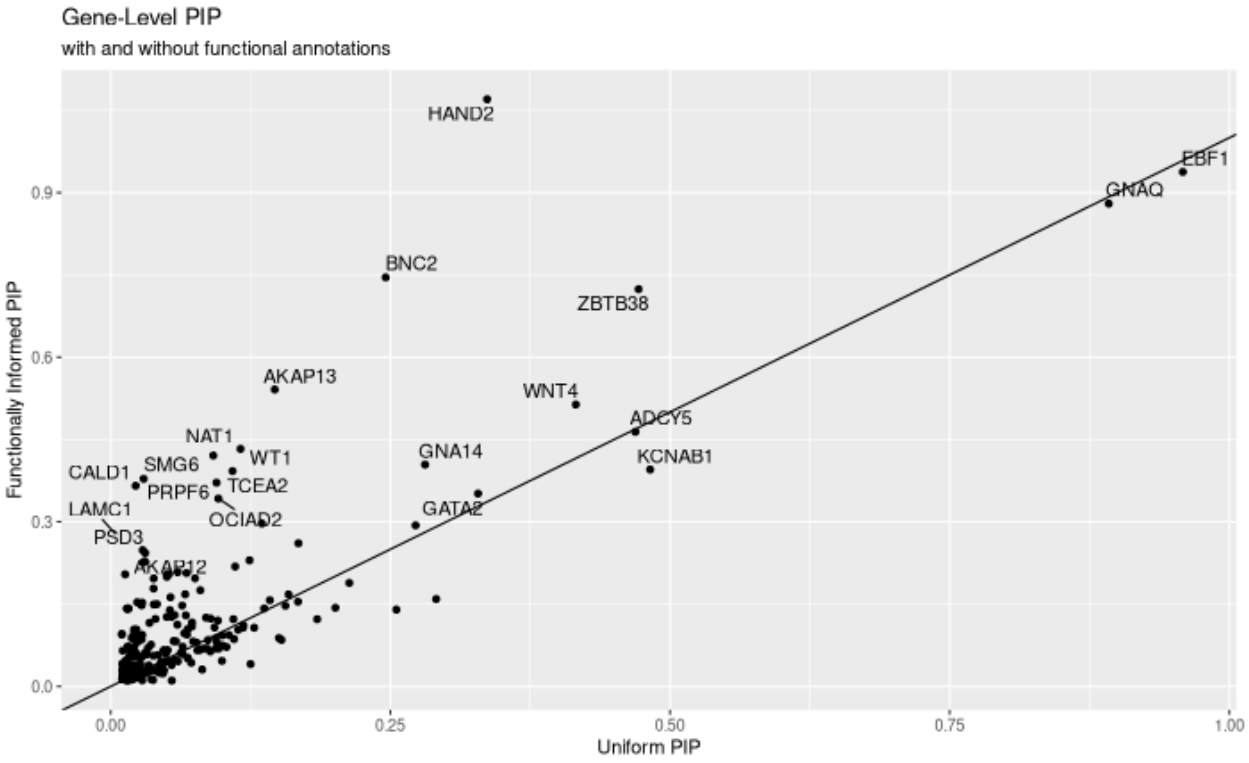

Figure 4.5: Gene-level PIPs under the functionally informed model as compared to the uniform model. The same variant-gene assignment was used for both models. Genes above the black diagonal line have a higher gene-level PIP under the functionally-informed model compared to the uniform model.

and endometrial decidualization[70], making this another excellent candidate causal gene for gestational duration and PTB.

4.3.2 Gene-level summary of variant fine-mapping suggests candidate genes of gestation length

While our fine-mapping analysis highlights some putative causal variants and potential targets, in most cases, the PIPs of the variants are small, limiting our knowledge of the potential mechanisms in most loci. Given that we are ultimately interested in the genes involved in pregnancy phenotypes, we developed a computational procedure to summarize variant PIPs from fine-mapping, at the level of genes (see Methods). These "gene PIPs" provide a conve-

Differential Expression Test	Gene-level PIP/DE status	No. Genes
Control vs Decidualized	significantly DE,Not high PIP	4592
Control vs Decidualized	significantly DE,High PIP	18
Control vs Decidualized	not significantly DE,Not high PIP	16565
Control vs Decidualized	not significantly DE,High PIP	31
TCM vs Decidualized	significantly DE,Not high PIP	1439
TCM vs Decidualized	significantly DE,High PIP	8
TCM vs Decidualized	not significantly DE,Not high PIP	19718
TCM vs Decidualized	not significantly DE,High PIP	41
Term vs Preterm	significantly DE,Not high PIP	168
Term vs Preterm	not significantly DE,Not high PIP	20989
Term vs Preterm	not significantly DE,High PIP	49

Table 4.2: Intersection of differential expression results with gene-level fine mapping results. A gene is considered "high PIP" if the gene-level PIP exceeded 0.15, and is considered significantly differentially expressed if the FDR adjusted p-value is less than 0.1. A visual representation of these results can be seen in Figure 4.6

nient summary of the evidence of genes, and can be thought of as approximating the probability that a gene has an effect on the phenotype. Our procedure integrates genomic annotations about where a SNP is located relative to the gene body, as well as promoter capture HiC (PC-HiC) data that links variants to gene promoters. Meanwhile, the procedure accounts for uncertainty of SNP-to-gene mapping, with a SNP potentially assigned to multiple genes.

We applied this procedure to an expanded set of GWAS loci of gestation length. Specifically, we used TORUS to compute the "region level" FDR (see Methods), which summarizes how likely a block contains at least one causal variant. To include as many causal signals as possible, we use a loose FDR cutoff of 0.5. Given that many regions may be false positives, the SNP level PIPs from SuSIE may not be valid, or un-calibrated, since SuSiE assumes that there is at least one causal signal per block. To address this issue, we multiply the SNP level PIPs from SuSiE fine-mapping in any region, by (1 - FDR) of that region, which approximates the probability that the region contains at least one causal signal. Using these "calibrated" PIPs of all SNPs in a total of 33 regions, we computed gene-level PIPs of 756 genes in these regions. For comparison, we also use the same procedure to compute gene PIPs but with SNP-level

PIPs from SuSiE fine-mapping using uniform prior (i.e. no functional annotations used). We found that this procedures leads to 7 genes with high confidence, PIP > 0.6, and a number of genes with suggestive evidence with PIP between 0.2 and 0.6 (Figure 4.5). The results from using SuSiE PIPs without any functional information include significantly fewer genes, with only 2 genes having a PIP above 0.6.

Some of the candidate genes we found have plausible connections with gestation. One top gene is HAND2 (PIP > 1), which as discussed earlier is an important transcription factor mediates the effect of progesterone on uterine epithelium. GNAQ (PIP = 0.89) encodes Guanine nucleotide-binding protein G(q) sub-unit alpha. In mice with GNAQ knockout, uterine growth were significantly reduced. WNT4 (PIP = 0.5) is a transcription factor in the important WNT signaling pathway, and it promotes female sex development and represses male sex development. EBF1 (PIP = 0.02) is a transcription factor essential for B-cell development. Given the close interaction between B cell development and pregnancy[101], the function of EBF1 in gestation length seems plausible and worthy of further investigation.

Taking advantage of PIP summaries of more than 700 genes, we next investigated whether gestation length genes may be involved in transcriptional response of MSCs to fetal-derived stimuli. Our hypothesis is that some of these candidate genes may be important for how cells response to these stimuli, and then when disrupted by genetic variations, may lead to higher risk of PTB. We perform differential expression analysis in three comparisons, cells treated by decidualization signal (cAMP) vs. control, cells treated by Trophoblast condition medium (TCM) vs. dedidualized cells, and cells collected from preterm labor placental and from term labor. Using DEseq2, we identified 4610, 1447 and 168 differentially expressed genes (DEGs), in the decidualized vs control, TCM vs decidualized, and term vs preterm comparisons, respectively, at FDR < 0.1.

We plotted the PIPs and p-values (after multiple testing adjustment) from DEG analysis of all genes under three conditions in Figure 4.6. At gene PIP cutoff of 0.15 and p-value cutoff

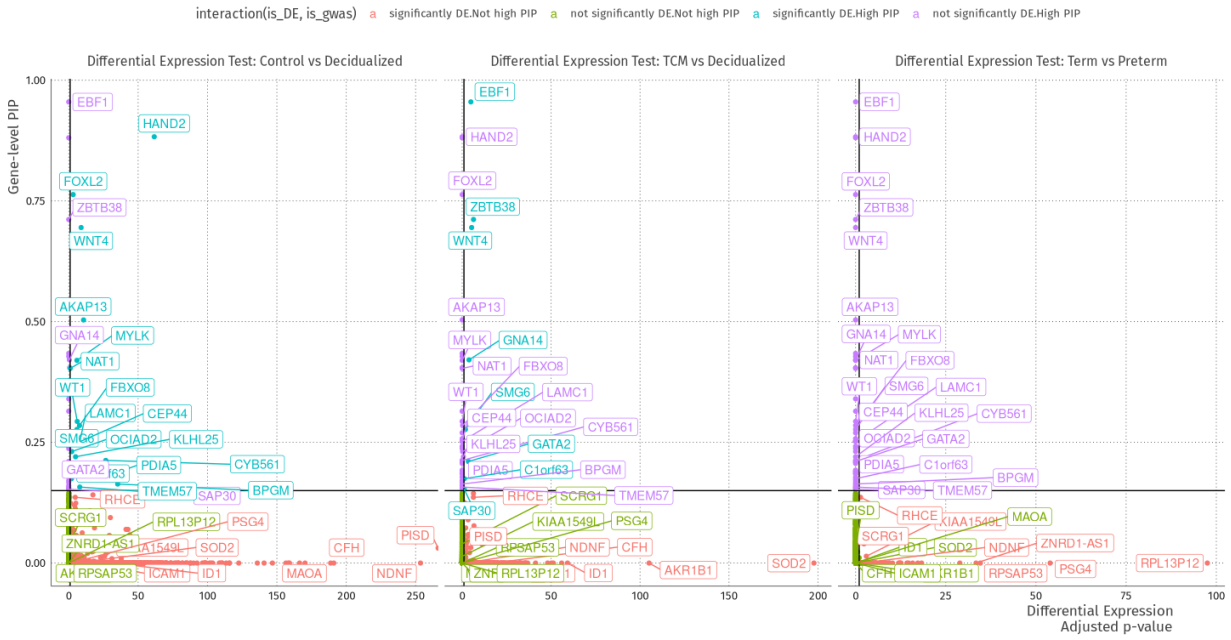

Figure 4.6: Differential expression adjusted -log10 p-value for each of three differential expression tests: control vs decidualized, decidualized vs TCM-treated, and term vs preterm, along with the corresponding gene-level PIP. The vertical and horizontal lines indicated thresholds that were used to test for the enrichment of high PIP genes for differentially expressed genes. High PIP genes were significantly enriched for both differential control vs decidualized genes ($p = 0.0264$), and for differential decidualized vs TCM-treated genes ($p = 0.0344$), by Fisher's exact test.

of 0.05, we found significant enrichment of gestation length candidate genes from GWAS in DEGs, in two comparisons. Among all 756 genes in gestation length GWAS regions, 14 genes with PIP > 0.15 show differential expression in the decidualization condition, while 18 genes with lower PIPs show evidence of DEG. This represents 2.2 fold enrichment, with $p = 0.04$ (Fisher's exact test). In the TCM vs. decidulization comparison, 6 genes with PIP > 0.15 are DEGs, representing 2.9 fold enrichment over genes with PIPs below the cutoff ($p = 0.03$). Given that few genes show DEGs between pre-term and term labor, we did not find evidence of enrichment of gestation genes in DEGs.

Take together, we have provided a large list of candidate genes of gestation length and PTB, combining statistical fine-mapping and functional genomic data sets in relevant cell types. The joint analysis with differential expression analysis provides support to the hypothesis that disruptions of transcriptional responses to pregnancy relevant conditions, including decidualization and response to TCM, are potential mechanisms of variation of gestation length and PTB.

4.4 Discussion

The lack of complete independence between our functional genomic annotations makes it difficult to delineate their individual effects but we have nonetheless highlighted the importance of enhancers and of gene regulation in endometrial stromal cells in modulating the effects of GWAS variants on gestational duration. This is consistent with both the known tissue-specific roles of enhancers and the observation that over 90% of GWAS loci reside outside of the coding portion of the genome and are enriched in regions of open chromatin and enhancers[61].

Integrating transcriptional and chromatin annotations of gene regulation from MSCs and DSCs improved our ability to discover novel GWAS loci and identify likely causal SNPs and genes associated with gestational duration. We illustrate how our integrated platform identified a novel causal locus and candidate gene (HAND2) associated with gestational duration,

as well as refined the annotation of loci that had been previously identified. Our data suggest that in endometrial stromal cells GATA2 is likely the target gene of enhancers harboring SNPs associated with gestational duration. This does not exclude the possibility that the nearest gene to the associated SNPs, EEFSEC, may be a target gene in other cell types.

Both of these examples highlight transcription factors that are essential for endometrial development or decidualization. The fact that neither GATA2 nor HAND2 were identified as potential candidate genes in previous GWASs of gestational duration or PTB supports our approach and the importance of using functional annotations from cell types relevant to pregnancy to fine map and identify candidate genes for the pregnancy-related traits. Overall, the integrated analyses performed in this study resulted in the identification of both novel GWAS loci and novel candidate genes for gestational duration, as well as maps of the regulatory architecture of these cells and their response to decidualization.

However, there are some limitations. Our results are based on only a small number of individuals, which may not be enough to fully capture the regulatory landscape of endometrial stromal cells. In addition, the individuals were African American and the GWAS results were obtained from Caucasians individuals and therefore, it is possible that the GWAS results do not match functional annotations in a different population, which could lead to erroneous conclusions. Another limitation is the fact that we focused on only one cell type, albeit one that plays a central role in pregnancy. Future studies that include fetal cells from the placenta and uterine or cervical myometrial cells could reveal additional processes that contribute to gestational duration and PTB, such as those related to fetal signaling and the regulation of labor, respectively. Second, to maximize power we focused on a GWAS of gestational duration and not PTB per se. While previous GWAS have shown that all PTB loci were among the gestational age loci[93], we realize that some of the loci that we identified could be related to normal variation in gestational duration and not specifically to PTB. Nonetheless, our findings contribute to our understanding of potential mechanisms underlying the timing of human

gestation, about which we still know little. Lastly, although our ChIP-seq results revealed an association between GATA2 binding and decidualization, confirming the role of this transcription factor in decidual cell biology([31], [70]), and studies in murines support its role in endometrial processes[70], we do not yet have direct evidence showing that perturbations in the expression of GATA2, or any of the other target genes identified, influence the timing of parturition in humans. Future studies will be needed to directly implicate the expression of these genes in gestational duration or PTB. Our study highlights the importance of generating functional annotations in pregnancy-relevant cell types to inform GWASs of pregnancy-associated conditions. Our results suggest that the expression of two transcription factors, GATA2 and HAND2, in endometrial stromal cells may regulate transcriptional programs that influence the timing of parturition in humans, which could lead to the identification of biomarkers of or therapeutic targets for PTB.

CHAPTER 5

CONCLUSION

The classical definition of a gene is a "unit of heredity" [46]. In the time before the central dogma of molecular biology, the gene was defined as that which gave rise to (i.e caused) that which was heritable — no distinction was made between that which was inherited and that which "caused" the heritable trait. In light of the central dogma, we now distinguish between that which is inherited, the gene, encoded in DNA, and that which gives rise to that which is heritable, which is commonly referred to as the "gene product". Rather than genes being the agents of both heritability and causality, we consider the gene the heritable precursor to the causal agent, which we dub the "gene product". In this dissertation I have examined the relationship between gene, gene product and phenotype from three perspectives.

With FGEM, I developed a statistical model that exploited the dual nature of the gene, gene-product relationship to find new causal genes and enriched pathways. The FGEM model assumes that if two gene products have similar characteristics they have similar prior probabilities of being causally related to a trait of interest. It similarly assumes that the genes most associated with a trait (by means of a gene-based test) causally influence the trait by a shared set of mechanisms that are reflected in the gene's characteristics. The study of cancer, in particular the problem of identifying mutational driver genes, provided ideal setting for the application of FGEM. Cancer is a disease of unregulated cell growth, believed to be caused (at least in part) by the accumulation of somatic mutations. The patterns of somatic point mutation in cancer cells, as summarized at the gene level by `driverMAPS`[96], provided clear signal of which genes are most likely to harbor an excess of somatic mutations. The biological process GO annotations provide a rich yet precise language for characterizing gene products. FGEM leveraged both of these datasets to identify new driver subtype-specific genes and pathways.

One caveat to consider when evaluation the results of FGEM, especially in the context of identifying cancer genes, is the extent to which FGEM may impart a "status-quo bias" in its

results. When embarking on an enrichment analysis of a gene-level association study for a trait that has never previously been studied, one might assume that genes with a high level of association with the trait are as likely to have gene-level annotations as genes with a low or intermediate level of statistical association. Cancer is not such a trait. As cancer biology is a major sub-discipline of biology, one must keep in mind that genes believed to be involved in cancer are more likely to have annotation, and it is further possible that annotations reflecting putative cancer mechanisms are in some sense "overrepresented" among annotations. As an example, a gene with a `driverMAPS` FDR < 0.1 has on average 19 Gene Ontology annotations, while a gene with a higher FDR has on average only 8.6 annotations. As a consequence, if a true causal gene operates by a mechanism not reflected in its annotations, the gene will be de-prioritized by FGEM.

I believe FGEM could also be useful in other settings where gene-based tests are commonly used, such as in the analysis of the contribution of rare and *de novo* variants to disease risk. It requires only evidence of gene functions in a phenotype, as measured by Bayes factors (BFs), as input. However, care needs to be taken to ensure that valid input data is given to FGEM. If BFs are not properly calibrated, then FGEM may not produce correct result. For example, in a recent study of *de novo* mutations in autism, the BFs of genes, when smaller than 1 (providing evidence against being a risk gene), are automatically assigned to 1[72]. As a result, overall BF distribution would be significantly inflated. It may be helpful thus, for a user of FGEM, to check BF distribution first to check if there is any anomaly before running FGEM. When the gene-based tests do not produce BFs, but p-values or other types of summary statistics (e.g. estimated effects and standard errors), one cannot directly use FGEM. However, it is possible to calculate BFs of genes using Empirical Bayes procedures developed for multiple testing, for example, Adaptive Shrinkage (ASH) method[79].

With RSSp I explored the phenotype genotype relationship at its most diffuse. I demonstrated how the infinitesimal model — a statistical model of inheritance that dates back over

a century — remains highly relevant in relating phenotype to genotype. Using a Bayesian approach, I combined the RSS likelihood with a normal prior and developed a method for heritability from summary statistics that outperformed the state of the art, LD score regression, across a variety of genetic architectures. In addition, we found that external LD reference panels, when of the same approximate size and from the same population as the GWAS cohort, can make suitable stand-ins when used for heritability estimation.

There are trade-offs to the fact that RSSp works with GWAS summary statistics and not individual-level data. This means that the quality of the heritability inference is (almost) entirely reliant on the quality of the summary statistics. If the original study suffered from population stratification, that bias will propagate into the heritability estimates of RSSp. In LDSC, population structure induced inflation of GWAS statistics is captured by the intercept term in the regression model. It is unclear, however, that this fully addressed the problem of population structure. Indeed, the distortion of GWAS summary statistics by population structure can be subtle. It was found, for example, that the polygenic risk scores (PRSs) based on summary statistics of height are confounded by population structure and result in misleading findings in the study of polygenic adaptation[4]. A future direction in RSSp is to generalize the model to accommodate some level of population structure.

Of additional concern for me, even in the absence of studies with population stratification, is the extent to which Europeans have become a de facto "model organism" in statistical genetics, and how little my work has done to push in the opposite direction. I have built a tool that 1) is most useful in combination with a large "reference" LD panel that is as similar to the original dataset as possible and 2) provides no functionality for testing whether the reference LD panel is appropriate for the GWAS. My method is not alone in this regard, and it is not irrational that scientists, without the tools to detect whether association is the result of stratification or not, and without large reference panels of non-Europeans readily available, try to avoid the issue by designing their study to only include European ancestry individuals.

My hope is that incentives related to funding and publication will start to push towards the creation of reference panels in less studied populations, so that this becomes less of an issue.

One main simplifying assumption of RSSp is the normal prior distribution of effect sizes of all variants. Recent work has shown more complex relationship of effect sizes of variants and their MAF and LD structure. For deleterious traits, higher MAF is generally associated with lower effect sizes, because of negative selection against large effect variants. Importantly, this dependency may vary across the traits, and can be learned for specific traits of interest[77]. The relationship of LD with effect size is more complex, but generally, low LD structure is associated with younger alleles, which tend to have bigger effects because of less time for natural selection to work[34]. But regions with low recombination rates, on the other hand, tend to harbor variants with larger effects because selection is less effective in such regions[34]. Methods have been developed to incorporate such dependency of effect sizes on MAF and LD into the prior model of effect sizes, e.g. LDAK and Stratified LDSC[77][34]. All these methods, however, are all method-of-moment estimators. Likelihood based model is possible. In fact, it is straightforward to extend RSSp so that effect sizes follow normal distribution with different variance for each variant, based on its MAF and LD. The challenge is that the computational technique of eigen-decomposition cannot be applied, so computational burden of full likelihood inference is daunting.

Despite that simple prior distribution being a limitation of RSSp, we note that RSSp estimated heritability is not the point estimate of the h^2 parameter, rather, it is the sum of the mean posterior estimate of PVE of each SNP. This "summary PVE" approach is closely related to the recently proposed Generalized Random Effect (GRE) estimator [41]. The GRE estimator is essentially a frequentist estimator of summary PVE. As argued in the GRE paper, this summary PVE approach is robust to the violation of the prior distribution. The GRE estimator, however, requires the number of samples to be substantially larger than the number of variants, which is a limitation in practice. RSSp does not have such limitation, and as we have

shown, performs well with relatively small (10K) sample sizes.

With the preterm birth project I integrated functional genomic annotation with GWAS data to identify new target genes. Applied to preterm birth, a trait thought to significant "missing heritability", we implicated the genes GATA2 and HAND2, genes that were not previously identified as contributing to gestational duration heritability. I found that through functionally-informed fine mapping, gene-level integration of these fine mapping results, and intersection with disease-relevant, tissue-specific differential expression results, pieces of the genotype-phenotype map reveal themselves. AKAP13 was among the genes that benefitted the most under the functionally informed fine-mapping, with a gene-level PIP of >0.5 with the functional model and a gene-level PIP of only 0.153 under the uniform model. The top SNP for AKAP13 has a GWAS p-value of only 1.185×10^{-6}, which contributes 0.03013043 of the PIP under the functional model, but with the contribution of other high-prior SNPs it obtains a relatively high gene-level PIP. We find that AKAP is significantly differentially expressed between decidualized and control, with an FDR adjusted p-value of 2.07×10^{-11}. AKAP13 is further interesting as it has been demonstrated to augment progesterone signaling in uterine fibroid cells[62]. This is interesting is that fibroids are characterized by altered extracellular matrices and increased stiffness. Mechanical stretch of the uterus is thought to be one of the key mechanical signals that regulates uterine development during pregnancy[73].

I am left with the impression that the human genetics field is still grappling with the "disappointment" of GWAS: studies are summarized either in terms of the number of loci that reach extremely stringent genome-wide significance threshold, or in terms of the polygenic SNP-heritability. Due to LD, it is often not possible to identify with certainty the sparse "true" set of causal variants that give rise to a GWAS signal. Even with sophisticated fine mapping techniques, the credible set of variants can be quite large, as was the case with the GATA2 locus. Despite this uncertainty at the variant level, there may in fact be very little uncertainty as to what the causal gene at the locus is, if there is strong evidence that all of the possible causal

variants are in an enhancer that is only known to loop to one gene. It is my hope that in the near future, genomic annotation, and methods for reporting uncertainty in genomic annotation, will improve to the point where geneticists will be able to simultaneously integrate over uncertainty in causal variant assignment, genomic annotation, and variant-to-gene assignment to improve gene mapping.

In particular, I think there are two major opportunities to extend my current method of gene-level fine-mapping summary. First, my variant-to-gene assignments are relatively simple and miss other important information. For instance, alternative splicing is one major mechanism of regulating gene functions, and variants affecting splicing can be important for complex traits[53]. Splicing variants are often located in intronic regions, and such variant-gene relationship is missed in my current method. A recently proposed Activity-by-contact (ABC) score is another example where SNP to gene function can be potentially assigned[32]. In the ABC score, the possible impact of a variant or DNA sequence on a gene is defined by the product of the regulatory activity of the sequence and the physical contact frequency of the sequence with a gene promoter from HiC data. My current assignment scheme only uses a binary assignment using promoter-capture Hi-C data, and may miss many pairs that are physically close in 3D. Secondly, when calculating the score of a gene, our method does not explicitly model "allelic heterogeneity" (AH), which describes the phenomenon that a locus may contain multiple risk variants. It has been reported that AH could be common, reaching as high as 50% in some complex traits[40]. It is very likely that in the AH case, multiple risk variants in a locus would target the same gene. With properly modeling of AH, I think it is possible to further improve the accuracy of fine-mapping causal genes.

REFERENCES

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, New York, NY, USA, 2002.

[2] Gaia Andreoletti, Lipika R Pal, John Moult, and Steven E Brenner. Reports from the fifth edition of cagi: The critical assessment of genome interpretation. *Human mutation*, 40(9):1197–1201, 2019.

[3] Douglas Bates and Dirk Eddelbuettel. Fast and elegant numerical linear algebra using the RcppEigen package. *Journal of Statistical Software*, 52(5):1–24, 2013.

[4] Jeremy J Berg, Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, Yair Field, Evan August Boyle, Xinjun Zhang, Fernando Racimo, Jonathan K Pritchard, et al. Reduced signal for polygenic adaptation of height in uk biobank. *Elife*, 8:e39725, 2019.

[5] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283, 2016.

[6] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006.

[7] R.P. Brent. *Algorithms for Minimization Without Derivatives*. Prentice-Hall series in automatic computation. Prentice-Hall, 1972.

[8] Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized tcga data from broad gdac firehose stddata 2015 06 01 run, 2015.

[9] bulik. ldsc, Nov 2020. [Online; accessed 25. Nov. 2020].

[10] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, 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(3):291–295, Feb 2015.

[11] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[12] William S Bush, Matthew T Oetjens, and Dana C Crawford. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nature Reviews Genetics*, 17(3):129, 2016.

[13] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, Sep 1995.

[14] Jonathan Cairns, Paula Freire-Pritchett, Steven W Wingett, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, Stefan Schoenfelder, Biola-Maria Javierre, Cameron Osborne, Peter Fraser, and Mikhail Spivakov. Chicago: Robust detection of dna looping interactions in capture hi-c data. *Genome Biology*, 17:127, 2016. R package version 4.0.2.

[15] Peter Carbonetto and Matthew Stephens. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for il-2 signaling genes in type 1 diabetes, and cytokine signaling genes in crohn's disease. *PLoS Genetics*, 9(10):e1003770, Oct 2013.

[16] Marc Carlson. *GO.db: A set of annotation maps describing the entire Gene Ontology*, 2020. R package version 3.11.0.

[17] J.M. Chambers and T. Hastie. *Statistical Models in S*. Wadsworth & Brooks/Cole computer science series. Wadsworth & Brooks/Cole Advanced Books & Software, 1992.

[18] Charlotte Chandeck and Wolter J. Mooi. Oncogene-induced cellular senescence. *Advances in Anatomic Pathology*, 17(1):42–48, 2010.

[19] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[20] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52, 2010.

[21] CreRecombinase. simu, Nov 2020. [Online; accessed 25. Nov. 2020].

[22] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[23] Francis HC Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[24] Krista S Crider, Nedra Whitehead, and Rebecca M Buus. Genetic variation associated with preterm birth: a huge review. *Genetics in Medicine*, 7(9):593–604, 2005.

[25] Sarah W Davies, Samuel V Scarpino, Thanapat Pongwarin, James Scott, and Mikhail V Matz. Estimating trait heritability in highly fecund species. *G3: Genes, Genomes, Genetics*, 5(12):2639–2645, 2015.

[26] Christiaan A de Leeuw, Joris M Mooij, Tom Heskes, and Danielle Posthuma. Magma: generalized gene-set analysis of gwas data. *PLoS Comput Biol*, 11(4):e1004219, 2015.

[27] Thomas Derrien, Jordi Estellé, Santiago Marco Sola, David G. Knowles, Emanuele Raineri, Roderic Guigó, and Paolo Ribeca. Fast computation and applications of genome mappability. *PLoS ONE*, 7(1):e30377, Jan 2012.

[28] Michal Elbaz, Ron Hadas, Louise M Bilezikjian, and Eran Gershon. Uterine foxl2 regulates the adherence of the trophectoderm cells to the endometrial epithelium. *Reproductive Biology and Endocrinology*, 16(1):12, 2018.

[29] Luke M Evans, Rasool Tahmasbi, Scott I Vrieze, Gonçalo R Abecasis, Sayantan Das, Steven Gazal, Douglas W Bjelland, Teresa R De Candia, Michael E Goddard, Benjamin M Neale, et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature genetics*, 50(5):737–745, 2018.

[30] JD French and SL Edwards. The role of noncoding variants in heritable disease. *Trends in Genetics*, 2020.

[31] Tohru Fujiwara. Gata transcription factors: basic principles and related human disorders. *The Tohoku Journal of Experimental Medicine*, 242(2):83–91, 2017.

[32] Charles P Fulco, Joseph Nasser, Thouis R Jones, Glen Munson, Drew T Bergman, Vidya Subramanian, Sharon R Grossman, Rockwell Anyoha, Benjamin R Doughty, Tejal A Patwardhan, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of crispr perturbations. *Nature genetics*, 51(12):1664–1669, 2019.

[33] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, and et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, Aug 2015.

[34] Steven Gazal, Hilary K Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M Neale, Alexander Gusev, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics*, 49(10):1421, 2017.

[35] Pinghua Gong and Jieping Ye. A modified orthant-wise limited memory quasi-newton method with convergence analysis. In *International Conference on Machine Learning*, pages 276–284, 2015.

[36] Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. Intogen-mutations identifies cancer drivers across tumor types. *Nature Methods*, 10(11):1081–1082, 2013.

[37] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, Mar 2011.

[38] Trevor Hastie, Robert Tibshirani, and Ryan J. Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *CoRR*, 2017.

[39] Xin He, Stephan J. Sanders, Li Liu, Silvia De Rubeis, Elaine T. Lim, James S. Sutcliffe, Gerard D. Schellenberg, Richard A. Gibbs, Mark J. Daly, Joseph D. Buxbaum, and et al.

Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genetics*, 9(8):e1003671, Aug 2013.

[40] Farhad Hormozdiari, Anthony Zhu, Gleb Kichaev, Chelsea J-T Ju, Ayellet V Segrè, Jong Wha J Joo, Hyejung Won, Sriram Sankararaman, Bogdan Pasaniuc, Sagiv Shifman, et al. Widespread allelic heterogeneity in complex traits. *The American Journal of Human Genetics*, 100(5):789–802, 2017.

[41] Kangcheng Hou, Kathryn S. Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nature Genetics*, 51(8):1244–1251, Jul 2019.

[42] Lin Hou and Hongyu Zhao. A review of post-gwas prioritization approaches. *Frontiers in Genetics*, 4, 2013.

[43] Rachael P Huntley, Tony Sawford, Maria J Martin, and Claire O'Donovan. Understanding how and why the gene ontology and its annotations evolve: the go within uniprot. *GigaScience*, 3(1):4, 2014.

[44] Iuliana Ionita-Laza, Seunggeun Lee, Vlad Makarov, Joseph D. Buxbaum, and Xihong Lin. Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, 92(6):841–853, Jun 2013.

[45] Longda Jiang, Zhili Zheng, Ting Qi, Kathryn E. Kemper, Naomi R. Wray, Peter M. Visscher, and Jian Yang. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.*, 51(12):1749–1755, Dec 2019.

[46] 1857-1927 Johannsen, W. (Wilhelm). *Arvelighedslærens elementer : forelæsninger holdte ved Københavns universitet.* Gyldendal ; Nordisk forlag, København : Kristiania, 1905.

[47] Matthew C Keller and William L Coventry. Quantifying and addressing parameter indeterminacy in the classical twin design. *Twin Research and Human Genetics*, 8(3):201–213, 2005.

[48] William C Knowler, RC Williams, DJ Pettitt, and A Gm Steinberg. Gm3; 5, 13, 14 and type 2 diabetes mellitus: an association in american indians with genetic admixture. *American journal of human genetics*, 43(4):520, 1988.

[49] David Lamparter, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLOS Computational Biology*, 12(1):e1004714, Jan 2016.

[50] Joy E Lawn and Mary Kinney. Preterm birth: now the leading cause of child death worldwide, 2014.

[51] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[52] Quanxi Li, Athilakshmi Kannan, Francesco J DeMayo, John P Lydon, Paul S Cooke, Hiroyuki Yamagishi, Deepak Srivastava, Milan K Bagchi, and Indrani C Bagchi. The antiproliferative action of progesterone in uterine epithelium is mediated by hand2. *Science*, 331(6019):912–916, 2011.

[53] Yang I Li, Bryce Van De Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. Rna splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, 2016.

[54] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12), Dec 2014.

[55] Michael I. Love, Charlotte Soneson, Peter F. Hickey, Lisa K. Johnson, N. Tessa Pierce, Lori Shepherd, Martin Morgan, and Rob Patro. Tximeta: Reference sequence checksums for provenance identification in rna-seq. *PLOS Computational Biology*, 16(2):e1007664, Feb 2020.

[56] Matthew Lyon, Shea J Andrews, Ben Elsworth, Tom R Gaunt, Gibran Hemani, and Edoardo Marcora. The variant call format provides efficient and robust storage of gwas summary statistics. May 2020.

[57] Anubha Mahajan, Daniel Taliun, Matthias Thurner, Neil R Robertson, Jason M Torres, N William Rayner, Anthony J Payne, Valgerdur Steinthorsdottir, Robert A Scott, Niels Grarup, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature genetics*, 50(11):1505–1513, 2018.

[58] Julian B Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, Joanna MM Howson, Adam Auton, Simon Myers, Andrew Morris, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics*, 44(12):1294, 2012.

[59] Iñigo Martincorena, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041, 2017.

[60] Joan Massagué. Tgf-β signal transduction, 1998.

[61] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[62] Sinnie Sin Man Ng, Soledad Jorge, Minnie Malik, Joy Britten, Szu-Chi Su, Charles R Armstrong, Joshua T Brennan, Sydney Chang, Kimberlyn Maravet Baig, Paul H Driggers, et al. A-kinase anchoring protein 13 (akap13) augments progesterone signaling in uterine fibroid cells. *The Journal of Clinical Endocrinology & Metabolism*, 104(3):970–980, 2019.

[63] Hoang T Nguyen, Julien Bryois, April Kim, Amanda Dobbyn, Laura M Huckins, Ana B Munoz-Manchado, Douglas M Ruderfer, Giulio Genovese, Menachem Fromer, Xinyi Xu, et al. Integrated bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome medicine*, 9(1):114, 2017.

[64] Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, et al. Functional snps in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature genetics*, 32(4):650–654, 2002.

[65] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417–419, 2017.

[66] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573, 2014.

[67] Yuliya Pylayeva-Gupta, Elda Grabocka, and Dafna Bar-Sagi. Ras oncogenes: weaving a tumorigenic web. *Nature Reviews Cancer*, 11(11):761–774, Oct 2011.

[68] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[69] Noah A. Rosenberg, Lucy Huang, Ethan M. Jewett, Zachary A. Szpiech, Ivana Jankovic, and Michael Boehnke. Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, 11(5):356–366, May 2010.

[70] Cory A Rubel, San-Pin Wu, Lin Lin, Tianyuan Wang, Rainer B Lanz, Xilong Li, Ramakrishna Kommagani, Heather L Franco, Sally A Camper, Qiang Tong, et al. A gata2-dependent transcription network regulates uterine progesterone responsiveness and endometrial function. *Cell reports*, 17(5):1414–1425, 2016.

[71] Noboru Sakabe, Ivy Aneas, Nicholas Knoblauch, Debora R. Sobreira, Nicole Clark, Cristina Paz, Cynthia Horth, Ryan Ziffra, Harjot Kaur, Xiao Liu, Rebecca Anderson, Jean Morrison, Virginia C. Cheung, Chad Grotegut, Timothy E. Reddy, Bo Jacobsson, Mikko Hallman, Kari Teramo, Amy Murtha, John Kessler, William Grobman, Ge Zhang, Louis J. Muglia, Sarosh Rana, Vincent J. Lynch, Gregory E. Crawford, Carole Ober, Xin He, and Marcelo A. Nóbrega. Transcriptome and regulatory maps of decidua-derived stromal cells inform gene discovery in preterm birth. *bioRxiv*, page 2020.04.06.017079, Apr 2020.

[72] F. Kyle Satterstrom, Jack A. Kosmicki, Jiebiao Wang, Michael S. Breen, Silvia De Rubeis, Joon-Yong An, Minshi Peng, Ryan Collins, Jakob Grove, Lambertus Klei, Christine Stevens, Jennifer Reichert, Maureen S. Mulhern, Mykyta Artomov, Sherif Gerges, Brooke Sheppard, Xinyi Xu, Aparna Bhaduri, Utku Norman, Harrison Brand, Grace Schwartz, Rachel Nguyen, Elizabeth E. Guerrero, Caroline Dias, Catalina Betancur, Edwin H. Cook, Louise Gallagher, Michael Gill, James S. Sutcliffe, Audrey Thurm, Michael E. Zwick, Anders D. Børglum, Matthew W. State, A. Ercument Cicek, Michael E. Talkowski, David J. Cutler, Bernie Devlin, Stephan J. Sanders, Kathryn Roeder, Mark J. Daly, Joseph D. Buxbaum, Branko Aleksic, Richard Anney, Mafalda Barbosa, Somer Bishop, Alfredo Brusco, Jonas Bybjerg-Grauholm, Angel Carracedo, Marcus C.Y. Chan, Andreas G. Chiocchetti, Brian H.Y. Chung, Hilary Coon, Michael L. Cuccaro, Aurora Curró, Bernardo Dalla Bernardina, Ryan Doan, Enrico Domenici, Shan Dong, Chiara Fallerini, Montserrat Fernández-Prieto, Giovanni Battista Ferrero, Christine M. Freitag, Menachem Fromer, J. Jay Gargus, Daniel Geschwind, Elisa Giorgio, Javier González-Peñas, Stephen Guter, Danielle Halpern, Emily Hansen-Kiss, Xin He, Gail E. Herman, Irva Hertz-Picciotto, David M. Hougaard, Christina M. Hultman, Iuliana Ionita-Laza, Suma Jacob, Jesslyn Jamison, Astanand Jugessur, Miia Kaartinen, Gun Peggy Knudsen, Alexander Kolevzon, Itaru Kushima, So Lun Lee, Terho Lehtimäki, Elaine T. Lim, Carla Lintas, W. Ian Lipkin, Diego Lopergolo, Fátima Lopes, Yunin Ludena, Patricia Maciel, Per Magnus, Behrang Mahjani, Nell Maltman, Dara S. Manoach, Gal Meiri, Idan Menashe, Judith Miller, Nancy Minshew, Eduarda M.S. Montenegro, Danielle Moreira, Eric M. Morrow, Ole Mors, Preben Bo Mortensen, Matthew Mosconi, Pierandrea Muglia, Benjamin M. Neale, Merete Nordentoft, Norio Ozaki, Aarno Palotie, Mara Parellada, Maria Rita Passos-Bueno, Margaret Pericak-Vance, Antonio M. Persico, Isaac Pessah, Kaija Puura, Abraham Reichenberg, Alessandra Renieri, Evelise Riberi, Elise B. Robinson, Kaitlin E. Samocha, Sven Sandin, Susan L. Santangelo, Gerry Schellenberg, Stephen W. Scherer, Sabine Schlitt, Rebecca Schmidt, Lauren Schmitt, Isabela M.W. Silva, Tarjinder Singh, Paige M. Siper, Moyra Smith, Gabriela Soares, Camilla Stoltenberg, Pål Suren, Ezra Susser, John Sweeney, Peter Szatmari, Lara Tang, Flora Tassone, Karoline Teufel, Elisabetta Trabetti, Maria del Pilar Trelles, Christopher A. Walsh, Lauren A. Weiss, Thomas Werge, Donna M. Werling, Emilie M. Wigdor, Emma Wilkinson, A. Jeremy Willsey, Timothy W. Yu, Mullin H.C. Yu, Ryan Yuen, Elaine Zachi, Esben Agerbo, Thomas Damm Als, Vivek Appadurai, Marie Bækvad-Hansen, Rich Belliveau, Alfonso Buil, Caitlin E. Carey, Felecia Cerrato, Kimberly Chambert, Claire Churchhouse, Søren Dalsgaard, Ditte Demontis, Ashley Dumont, Jacqueline Goldstein, Christine S. Hansen, Mads Engel Hauberg, Mads V. Hollegaard, Daniel P. Howrigan, Hailiang Huang, Julian Maller, Alicia R. Martin, Joanna Martin, Manuel Mattheisen, Jennifer Moran, Jonatan Pallesen, Duncan S. Palmer, Carsten Bøcker Pedersen, Marianne Giørtz Pedersen, Timothy Poterba, Jesper Buchhave Poulsen, Stephan Ripke, Andrew J. Schork, Wesley K. Thompson, Patrick Turley, and Raymond K. Walters. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*, 180(3):568–584.e23, 2020.

[73] Oksana Shynlova, Ruth Kwong, and Stephen J Lye. Mechanical stretch regulates hyper-

trophic phenotype of the myometrium during pregnancy. *Reproduction*, 139(1):247, 2010.

[74] Siming Zhao. drivermaps run results for 20 tumor types from tcga, 2019.

[75] Zbyslaw Sondka, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, Oct 2018.

[76] Doug Speed and David J Balding. Sumher better estimates the snp heritability of complex traits from summary statistics. *Nature genetics*, 51(2):277–284, 2019.

[77] Doug Speed, Na Cai, Michael R Johnson, Sergey Nejentsev, and David J Balding. Reevaluation of snp heritability in complex human traits. *Nature genetics*, 49(7):986–992, 2017.

[78] Jeffrey P. Spence and Yun S. Song. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 5(10):eaaw9206, Oct 2019.

[79] Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

[80] stephenslab. ldshrink, Nov 2020. [Online; accessed 26. Nov. 2020].

[81] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.*, 12(3):e1001779, Mar 2015.

[82] Joseph Swift and Gloria M Coruzzi. A matter of time—how transient transcription factor interactions create dynamic gene regulatory networks. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1860(1):75–83, 2017.

[83] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, 11 2018.

[84] Matthias Thurner, Martijn Van De Bunt, Jason M Torres, Anubha Mahajan, Vibe Nylander, Amanda J Bennett, Kyle J Gaulton, Amy Barrett, Carla Burrows, Christopher G Bell, et al. Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. *Elife*, 7:e31977, 2018.

[85] Gao Wang, Abhishek K Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv*, page 501114, 2019.

[86] Kai Wang, Mingyao Li, and Hakon Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854, 2010.

[87] Xiaoquan Wen, Yeji Lee, Francesca Luca, and Roger Pique-Regi. Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129, 2016.

[88] Xiaoquan Wen and Matthew Stephens. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The Annals of Applied Statistics*, 4(3):1158–1182, Sep 2010.

[89] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, and et al. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, Jun 2010.

[90] Nengjun Yi and Degui Zhi. Bayesian analysis of rare variants in genetic association studies. *Genetic epidemiology*, 35(1):57–69, 2011.

[91] Ronald Yurko, Max G'Sell, Kathryn Roeder, and Bernie Devlin. A selective inference approach for false discovery rate control using multiomics covariates yields insights into disease risk. *Proceedings of the National Academy of Sciences*, 2020.

[92] Rong W Zablocki, Andrew J Schork, Richard A Levine, Ole A Andreassen, Anders M Dale, and Wesley K Thompson. Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics*, 30(15):2098–2104, 2014.

[93] Ge Zhang, Bjarke Feenstra, Jonas Bacelis, Xueping Liu, Lisa M Muglia, Julius Juodakis, Daniel E Miller, Nadia Litterman, Pan-Pan Jiang, Laura Russell, et al. Genetic associations with gestational duration and spontaneous preterm birth. *New England Journal of Medicine*, 377(12):1156–1167, 2017.

[94] Martin J Zhang, Fei Xia, and James Zou. Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature communications*, 10(1):1–11, 2019.

[95] Ming Zhao, Lopa Mishra, and Chu-Xia Deng. The role of tgf-β/smad4 signaling in cancer. *International Journal of Biological Sciences*, 14(2):111–123, 2018.

[96] Siming Zhao, Jun Liu, Pranav Nanga, Yuwen Liu, A. Ercument Cicek, Nicholas Knoblauch, Chuan He, Matthew Stephens, and Xin He. Detailed modeling of positive selection improves detection of cancer driver genes. *Nature Communications*, 10(1):3399, 2019.

[97] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*, 9(2):e1003264, Feb 2013.

[98] Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The Annals of Applied Statistics*, 11(3):1561–1592, Sep 2017.

[99] Xiang Zhu and Matthew Stephens. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature Communications*, 9(1), Oct 2018.

[100] Xiang Zhu and Matthew Stephens. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature communications*, 9(1):1–14, 2018.

[101] Katharina B Ziegler, Damián Oscar Muzzio, F Matzner, Imke Bommer, MS Ventimiglia, K Malinowsky, J Ehrhardt, Marek Zygmunt, and F Jensen. Human pregnancy is accompanied by modifications in b cell development and immunoglobulin profile. *Journal of reproductive immunology*, 129:40–47, 2018.