Statistical methods for genomic sequencing data

Alan Min

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
William Stafford Noble, Chair
Elizabeth Thompson
Daniela Witten

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Statistical methods for genomic sequencing data

Alan Min

Chair of the Supervisory Committee:
Professor William Stafford Noble
Genome Sciences

Genomic sequencing data has revolutionized our understanding of the genetic basis of biological processes. The cost of sequencing the first human genome was estimated to be greater than 50 million dollars. However, with the advent of next generation sequencing, that cost has decreased to a few hundred dollars. It is thus now possible to use sequencing technology to understand nuanced aspects of the cell, both on the population and at the single-cell level. In this dissertation, we present three projects that develop statistical methods for analyzing genomic data.

In the first project, we discuss how heritability estimators based on single nucleotide polymorphisms are affected under alternative structures of linkage disequilibrium. We demonstrate that linkage disequilbrium has the potential to bias modern estimators of heritability. In the second project, we investigate a sequencing-based assay that measures local chromatin structure. In this context, we propose a prior that allows a latent Dirichlet allocation model chromatin accessibility data to leverage auxiliary data. In the third project, we consider the connection between sequence data and epigenomic or expression data in the context of multitask learning models. We demonstrate that this multitask learning setup can lead to inaccurate models, when genomic features that are irrelevant for one task are erroneously assigned significance in a related task.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to acknowledge my advisor Bill Noble who has provided tremendous guidance and support. I remember when I first met Bill in the genome science class he was teaching, and contacting him about potentially doing research with him. Despite being from a different department, Bill was willing to meet with me and work with me, introducing me to aspects of computational biology that I did not previously understand. I am also grateful to Elizabeth Thompson, who approached me at the Statistical Genetics Journal Club about a potential project. Elizabeth was incredibly supportive through my first paper and taught me so much about heritability and statistical genetics. Throughout the rest of my PhD, Elizabeth has always been someone I felt I could go to for advice. I would also like to thank Saonli Basu, who along with Elizabeth, worked with me on the the heritability project and spent countless hours discussing simulations with me and providing important feedback and contributing to writing the manuscript. Thank you to Daniela Witten, who served on my reading committee and was always willing to discuss problems I was facing in my projects. Thank you to Ali Shojaie for serving on my committee and offering great feedback on my projects. Thank you to Sara Mostafavi, who served as my graduate school representative and offering her insights in our projects.

In addition to faculty mentorship, I was helped a tremendous amount by other members of the Noble lab and members of the genome science community. I am thankful for Tim Durham, who mentored me through the LDA project. Tim had tremendous patience with me as I worked through tough simulations and faced difficulty with evaluation of our method. Tim offered important feedback and suggestions and also gave invaluable detailed feedback on many versions of the manuscript. Tim helped finish up the LDA project, despite the fact

that he had graduated and had started a new job. I will always remember the chats that Tim and I had after meetings ended about hiking or biking around Seattle. Thank you to Jacob Schreiber, who mentored me in the multitask model, as well as another imputation project. Jacob helped me with becoming a better coder, teaching me about how to use various machine learning packages and offering suggestions about ideas to push forward in our projects. Jacob also provided humor in our meetings, often bringing an air of fun and excitement to our projects even when work was hard. Thank you to Louis Gevirtzman from the Waterston lab who provided coding support when I was working on the LDA project. Thank you to all the members of the Noble lab and the Genome Sciences department who I have had the great pleasure of calling friends.

Thank you to my friends from my cohort from the statistics department. Thank you to Sarah Teichman, for all the hikes, the paddleboarding, Ravenna brewing, frisbee, and so much more. Thank you to Anna Neufeld for the late nights eating nachos and working, runs throughout Seattle, hikes in the mountains, bike rides around Washington, and everything else. Thank you to Anupreet Porwal, for the many restaurants we got to try, the nights out, the great tennis battles, and the great laughter. Thank you to Michael Pearce, for SMSDBC, climbing sessions at SBP, backpacking and hiking, and for the good times. Thank you to all of the members of the statistics community who I have developed invaluable friendships with. Thank you to all the friends I have made around Seattle who have offered support and friendship.

Finally, thank you to my mother who has always been willing to lend a listening ear or support me in whatever endeavors I choose to pursue.

DEDICATION

Mt. Rainier

Chapter 1
INTRODUCTION

One of the first widespread methods to determine genomic sequence was Sanger sequencing, which used electrophoresis and the incorporation of dideoxynucleotides to determine the base pair composition of sequences [Sanger and Coulson, 1975]. Although this method was able to accurately determine genomic sequence, it was slow and unable to produce large quantities of reads by today's standards—even automated machines employing these techniques could only read up to 1,000 sequences per day by 1987. Gradual improvement has been made since then, including improvement to the polymerase used to attach the dideoxynucleotides, improved preprocessing steps, more rapid electrophoresis, and other automation steps [Shendure et al., 2017]. The Human Genome Project [International Human Genome Sequencing Consortium., 2001] spurred further development of genomic sequencing technology, leading up to modern next generation sequencing techniques (NGS). These NGS technologies used sequencing by fluorescence, where as the sequence is amplified, different bases would fluoresce in different colors [Balasubramanian et al., 2003]. This approach can be conducted in a massively parallel way by generating a large array of immobilized templates [Shendure et al., 2005]. These advances, along with many more, have brought the estimated cost of sequencing the human genome down from fifty million down to thousands or even only hundreds of dollars [National Human Genome Research Institute, 2021]. The efficiency of NGS has enabled collection of genomic data at a scale never before seen, enabling scientific discovery.

With the human genome sequenced, and genomic sequencing technology on the rise, it became possible to identify base-pair positions in the genome that tended to vary between individuals. These positions, single-nucleotide polymorphisms (SNPs), were thought to contribute to individual physical traits, and have been critical to studying many human genetic

diseases through genome-wide association studies [Uffelmann et al., 2021]. One way to identify SNPs employs a similar technique to whole genome sequencing by using fluorescence to identify the SNP. This technology creates a library of primers that anneal next to the SNP, and upon extension, results in the identification of the SNP [Kim and Misra, 2007]. Many variants of this technology exist, including some that are commercially available, such as the MassARRAY [Gabriel and Ziaugra, 2004] and Affymetrix [Nishida et al., 2008]. With these technologies available, efforts have been made to identify SNPs in large populations of humans, such as the 1000 Genomes Project [1000 Genomes Project Consortium, 2015], which was an effort to genotype over 1000 individuals from 26 different populations in order to understand human variability, or the TOPMed program [Taliun et al., 2021], which identified SNPs in over 50,000 samples with the goal of understanding the genetic basis of heart, lung, blood, and sleep diseases.

1.1 Heritability

One of the many accomplishments enabled by the increased output of SNP genotyping is in understanding heritability, a measure of how much of the variance in a physical trait can be explained by genomic information.

To define what heritability is, we first begin with defining a model of phenotypes and genotypes, following [Crow and Kimura, 1970]. We assume that at each locus within a population, individuals have one of two alleles, say A_1 and A_2. For our purposes, we only consider loci which are biallelic and we only consider additive effects. Furthermore, for now, we assume that there is only one locus of interest. Assume that the overall mean phenotypic value of the trait is $\bar{a}$, and that the average phenotypic value for individuals with specific alleles are

$$Y_g = \begin{cases} \bar{a} + 2\alpha_1 & \text{if individual has } A_1A_1 \text{ alleles} \\ \bar{a} + \alpha_1 + \alpha_2 & \text{if individual has } A_1A_2 \text{ alleles} \\ \bar{a} + 2\alpha_2 & \text{if individual has } A_2A_2 \text{ alleles} \end{cases} \tag{1.1}$$

where α_1 and α_2 can be thought of as deviations from the average phenotypic value due to having either the A_1 or A_2 allele respectively, and Y_g is the mean phenotype for individuals with the specified alleles. We call α_1 and α_2 additive effects. We can also define the frequency of the alleles A_1 and A_2 in the population of interest as p_1 and p_2. We then define the genic variance as the variance due to A_1 and A_2, which can be calculated to be

$$V_g = 2p_1p_2(\alpha_1 - \alpha_2)^2 \tag{1.2}$$

by using the fact that α_1 and α_2 are deviations from the mean. We can then use this variance to define "narrow sense" heritability, by defining

$$h^2 = V_g/Var(Y) \tag{1.3}$$

Here, Y is the total phenotypic variance, including any environmental effects. We call this the narrow sense heritability because V_g only includes variance due to additive effects of the alleles. Although it is not considered in this work, if we were to instead include all variance effects of the alleles, we would call the quantity "broad sense" heritability.

We have here considered the case of one marker, but in the case of multiple loci A_{1i} and A_{2i}, where i is the index of the locus, then we can instead write

$$V_g = \sum_i 2p_{1i}p_{2i}(\alpha_{1i} - \alpha_{2i})^2 \tag{1.4}$$

We will now discuss methods used to estimate heritability, starting from historic methods that used kinship values following [Crow and Kimura, 1970]. We define the kinship values between two individuals as

$$k_0 = \text{P(No genes are identical by descent at the locus)} \tag{1.5}$$

$$2k_1 = \text{P(One gene is identical by descent, but not both)} \tag{1.6}$$

$$k_2 = \text{P(Both genes are identical by descent at the locus)} \tag{1.7}$$

We note that these values can be calculated by knowing the pedigree of the family. If we are only interested in additive effects, and we assume that environmental effects are independent

of the genes, and there is no inbreeding between individuals, it can then be derived that if we have two individuals with phenotypic values Y_1 and Y_2, then

$$Cor(Y_1, Y_2) = (k_1 + k_2)h^2 \tag{1.8}$$

where Cor is the correlation function, and h^2 is the narrow sense heritability. This equation leads to

$$Cor(Y_1, Y_2)/(k_1 + k_2) \tag{1.9}$$

as an estimate for narrow sense heritability. We note that although this more closely resembles a method of moments estimator, it is also possible to use an identity by descent strategy with a likelihood approach to estimate heritability [Evans et al., 2018a]. We also note that this estimator is similar in form to Haseman-Elston regression [Haseman and Elston, 1972], which we later formulate as an estimate of heritability as well.

In later work, heritability was also studied using twin data, comparing the correlation between physical characteristics of twins that were raised separately compared to those that were raised together [Bouchard Jr et al., 1990]. The goal was to eliminate the possibility that a shared environment confounded estimates of heritability. For example, in Eq. 1.9, if the two individuals are siblings, they share both environmental factors and genetic factors. Hence, the phenotypic correlation would be higher than expected from solely genetic reasons, and we may overestimate heritability. When comparing twins that are raised apart, it is possible to remove the common environment factor.

In twin studies, it can be difficult to find enough twins for the study, especially when comparing twins raised together versus apart. Hence, it has been proposed to study heritability through apparently unrelated individuals [Visscher et al., 2006]. Other modern methods also use this approach [Yang et al., 2011, Speed et al., 2012, Haseman and Elston, 1972, Schwartzman et al., 2019]. The goal was to both avoid the shared common environment of studying relatives and to avoid the difficulty of finding subjects for a study that relied on having siblings. In these methods using unrelated individuals, kinship is unknown, and hence often the "genetic relatedness matrix" (GRM) is used. To define the GRM, we consider only

biallelic genes, and we name one of the alleles the reference allele. We then count the number of occurences of the reference allele. We consider then a matrix of genotypes, $\mathbf{G}$, where for where individual i and locus j, if A_1 is the reference allele and A_2 is the alternate allele, then

$$G_{ij} = \begin{cases} 0 & \text{If individual } i \text{ has } A_2A_2 \text{ alleles at locus } j \\ 1 & \text{If individual } i \text{ has } A_1A_2 \text{ alleles (or } A_2A_1\text{) at locus } j \\ 2 & \text{If individual } i \text{ has } A_1A_1 \text{ alleles at locus } j \end{cases} \tag{1.10}$$

We define $\mathbf{\Gamma}$ as the normalized $\boldsymbol{G}$. Then the GRM defined as

$$M^{-1}\Gamma\Gamma^T \tag{1.11}$$

Where M is the number of loci of interest, and T denotes the transpose operation. The GRM accounts for realized relatedness of the individuals. In other words, the amount of correlation between the genotypes of an individual. We can then use the GRM to estimate heritability.

These modern approaches typically take either a method of moments (MoM) or likelihood approach. Whereas the MoM estimators analytically solve for quantities of interest using multiple moments, likelihood approaches create a probabilistic model and optimize for the unknown quantities of interest to achieve an estimate. As an example of a MoM estimator, we can formulate an estimator based off of Haseman and Elston [1972] by assessing the relationship between phenotypic correlation and shared genomic content. As an example of a likelihood estimator, Yang et al. [2011] create a likelihood based on a normal distribution with variance components determined by genomic relatedness (more details about both estimators in Section 2.4.2). Another innovation has been that rather than using traditional maximum likelihood to estimate variance components, restricted maximum likelihood (REML) has been used [Corbeil and Searle, 1976]. As another distinction between estimators is that some use a fixed-effects model, in which assume that there are fixed values corresponding to effect sizes of each SNP, whereas other estimators use a random-effects model, assuming that the effect sizes are all drawn from a distribution.

In the first project of this thesis, titled **Comparing Heritability Estimators under Alternative Structures of Linkage Disequilibrium**, we will present a project exploring the ways in which heritability estimates based on SNPs, or narrow sense heritability, can be affected by linkage disequilbrium (LD), or correlation between positions of the genome. This is similar to how Browning and Browning [2011] and Lin et al. [2022] investigated ways in which heritability is affected by population structure. In our work, we propose various ways in which LD can manifest in the genome, and study the effects of LD on multiple estimators. We considered various estimators, including both the MoM estimators and the likelihood based methods. We were able to conduct a theoretical analysis of the properties of the MoM estimators due to their computationally tractable form. This allowed us to explicitly write out an equation of the bias in the heritability estimate due to LD for some of the estimates. On the other hand, we were unable to analytically understand the properties of the likelihood based estimators as there was no closed form representation of the estimators. Hence, for those estimators, we conducted simulation studies by simulating SNPs and LD between them. Ultimately, we found that although some methods aim to adjust their estimates based on the correlation in the data, they are still vulnerable to correlation in the input data.

1.2 Functional genomics and latent Dirichlet allocation

Beyond pure sequence information, sequencing technology has enabled functional genomic measurements through variants of sequencing both in bulk sequence and at single-cell resolution. For example, it is often of interest defining regions of euchromatin and heterochomatin. Whereas euchromatin is comprised of open and actively transcribed genomic regions, heterochromatin is condensed and very rarely transcribed [Saksouk et al., 2015]. Furthermore, it may be of interest to identify how unique cell types differ in their chromatin accessibility, requiring the ability to sequence each cell individually. Although there are a variety of approaches to answering these two questions, scATAC-seq [Buenrostro et al., 2015] is one example that has enabled answering both these questions simultaneously. This technique uses Tn5 transposase to insert adapters into genomic regions, but can only do so in ac-

cessible regions of the genome. Because sequencing can only occur when the adapters are inserted, sequencing reads identify regions of the genome that were accessible. Furthermore, scATAC-seq employs barcoding technology to enable unique identification of cells. Each cell is assigned a unique barcode, which is sequenced along with the genomic reads, which disambiguates the cells. In post-processing, calls are often made for the genomic reads to "peak" regions, or regions that have enriched scATAC-seq signal. scATAC-seq has been used to identify cell types in mouse [Cusanovich et al., 2018], understand differentiation [Jiang et al., 2023], and identify cancer subtypes [Wang et al., 2021], as a few examples.

One challenge with scATAC-seq data is its high dimensionality. Often, sequencing reads are mapped to either peak regions or genes, resulting in a data matrix with number of rows corresponding to the number of cells, and number of columns corresponding to the number of genes or peaks. Generally, the goal of analyzing scATAC-seq data is to better understand cell types, which can be challenging with the dimensionality issues. Many methods have been proposed to analyze scATAC-seq data. A few popular examples are ChromVAR [Schep et al., 2017], SnapATAC [Fang et al., 2021], and cisTopic [González-Blas et al., 2019].

ChromVAR used an approach where peaks were mapped to motifs of interest and deviation from the expectation was calculated. The expected number of reads mapping to a peak is the average fraction of reads mapping to that peak across all cells, scaled by the number of reads in our cell of interest. Precisely, if x_{ij} ts the number of reads for a cell i and a peak j, then the expected number reads for cell i and peak j is

$$E = \frac{\sum_i x_{ij}}{\sum_i \sum_j x_{ij}} \sum_j x_{ij} \tag{1.12}$$

and if M is a matrix where

$$m_{k,j}$$

is 1 if motif k is present in peak j. Then the transformation is made to

$$Y = \frac{MX^T - ME^T}{ME^T} \tag{1.13}$$

This approach reduces the dimensionality of the raw peaks and can lead to improved resolution of fine-grained cell types. SnapATAC also does normalization steps, although involving more steps than can be described here, and ultimately takes eigenvalues of the data to reduce the dimensionality.

We focus on the methodology employed by cisTopic, which was based off of latent Dirichlet allocation (LDA). LDA was originally developed to be used as a tool to analyze corpuses of text and provide topic annotations for individual articles [Blei et al., 2003], but has since been applied successfully to single cell data, improving interpretability of cell types and aiding in the resolution of fine-grained cell types [González-Blas et al., 2019]. Generally, LDA is a Bayesian model that assumes the following generative process, following [Darling, 2011].

- For each topic t, $\phi_t \sim Dirichlet(\beta)$
- For each cell c, $\theta_c \sim Dirichlet(\alpha)$
- For each read r_i in c, Topic $t_i \sim Discrete(\theta_c)$, and $r_i \sim Discrete(\phi_{t_i})$

where in the context of scATAC-seq data, LDA takes as input a matrix of cells by scATAC-seq peaks, and outputs both a cell-topic matrix θ and a topic-gene matrix ϕ, and β and α are vector hyperparameters of length equal to the number of topics. The cell-topic matrix includes information about what the topic composition of each cell is, and the topic-gene matrix includes information about what the gene composition of each topic is. Using the LDA-based approach improves interpretability and also offers the flexible Bayesian framework.

In our second project, titled **Matrix prior for data transfer between single cell data types in latent Dirichlet allocation**, we investigated ways to improve analysis of scATAC-seq data using topic models such as LDA. We realized that although LDA has shown great promise in large datasets, it may still struggle on smaller datasets. Furthermore, despite great advances in sequencing technology, single-cell sequencing can still be prohibitively expensive to conduct at a large scale. Hence, we aimed to create a method to leverage information from

large atlas level single-cell datasets in the analysis of smaller, lower budget datasets. Prior methods using LDA have used a prior for the topic-peak matrix that we call the "uniform" prior, which is a Dirichlet prior with equal values for all parameters. That is, the prior assumes that all of the topics have an equal likelihood for all of the peaks. We hypothesized that we could improve on this by introducing a nonuniform "matrix" prior, which assigns different probabilities to different peaks based on the topic at hand. We hence proposed a matrix prior that took the results of LDA applied to a large atlas data and encoded it in the analysis of a smaller dataset. We showed that in certain cases, this would improve the quality of analyses in the smaller dataset.

1.3 Sequence to function models and multitask learning

In these first two projects, we investigated the use of genomic sequence information on its own to predict SNP heritability, and also investigated LDA analysis of cell populations using functional genomic data, in particular with scATAC-seq data. Many modern models consider the intersection of sequence information and functional genomic information [Zhou and Troyanskaya, 2015, Avsec et al., 2021a,b, Kelley et al., 2018, 2016, Zhou et al., 2018, Chen et al., 2022]. That is to say, these models take raw genomic sequence as input and aim to predict as output chromatin profiles, for instance ATAC-seq signal, ChIP-seq signal, or many other potential signals. The commonality of these functional tracks is that they have a measurement, often based on genomic sequencing reads, at each point along the genome. For example, SEI [Chen et al., 2022] takes as input any 4 kilobase window along the human genome, and predicts 21,907 chromatin profiles. These chromatin profiles may be different assays, or they may be data from different cell types. The goal of these tools is twofold: first, the authors aim to identify motifs associated with certain features of the genome, and second, the authors hope to achieve *in silico* mutagenesis by learning the semantics of the genome.

Because the problem of predicting chromatin profiles from sequence involves complex input sequences, and because the relationship between the input and the output is not clear,

researchers tackling this problem often rely on deep-learning approaches using a convolutional neural network architecture [LeCun et al., 2015]. A standard fully connected neural network architecture involves multiplying inputs by weight parameters, which are targets of optimization, and summing together the weighted inputs into separate "nodes" within a "layer". Deep-learning approaches often combine many layers of these connections, resulting in a model that is capable of learning complex nonlinear relationships between inputs and outputs if trained properly. Convolutional neural networks (CNN) add together inputs in a windowed fashion, thus reducing the number of parameters to learn and often improving performance. Training even with the reduced parameters of a CNN can be computationally intensive, however. Hence, researchers often employ two techniques. First, researchers oversample regions that they are interested in. For example, if the researcher is aiming to predict scATAC-seq data, they may create a training set where half of the training sequences come from peak regions, and the other half come from non-peak regions, even though most of the genome is not in a peak region. Second, researchers use multitask learning, a method in which a latent representation of a sequence is learned using all of the available genomic tracks, but a few final layers in a CNN model are added to produce the final output.

In our third project, entitled **Data leakage in sequence-based multitask genomics models**, we aim to tie together the previous two projects by investigating the task of predicting genomics data using sequence. We recognized that the common practice of oversampling regions of the genome and creating a multitask model with a shared latent representation to create sequence to function models may result in spurious results. We propose two possible reasons for this: peak selection or the shared latent representation of the models. We consider comparing fibroblast versus induced pluripotent stem cells (iPSC) as a running example to illustrate these issues, based on biological plausibility. Fibroblast cells are found in connective tissues and are fully differentiated, compared to iPSC cells, which are in early stages of differentiation. This leads to some critical differences in our biological expectations of the cell types. In our running example, we explore the accessibility of the Oct-Sox heterodimer transcription binding factor. Oct4 and Sox2 are two transcription factors that are active in

stem cells which are responsible for promoting transcription of differentiation factors [Ambrosetti et al., 1997, Tomioka et al., 2002]. They form a heterodimer and recognize a DNA motif, and we expect that in iPSC cells, these motifs will be in accessible regions of DNA. On the other hand, because Oct4 and Sox2 are not expressed in fibroblast cells, we expect no change in expression for these cells. Contrary to these biological expectations, we see that in many published models, fibroblast cells respond *in silico* to the Oct-Sox motif. We explore this phenomenon in the Sei model [Chen et al., 2022], a seqeunce-based multitask genomic model, and provide plausible reasons for it through simulation.

1.4 Organization of this dissertation

We will present each of these projects in their own chapter, following this one. They each have subsections and separate introductions and details about experiments and results. Supplementary materials for each chapter are found at the end of the chapter.

Chapter 2

COMPARING HERITABILITY ESTIMATORS UNDER ALTERNATIVE STRUCTURES OF LINKAGE DISEQUILIBRIUM

This chapter is adapted with minimal modification from:

Alan Min, Elizabeth Thompson, and Saonli Basu. Comparing heritability estimators under alternative structures of linkage disequilibrium. *G3*, 12(8):jkac134, 2022

2.1 Author Contribution

ET and SB conceived the idea of studying heritability estimators with LD structure. AM and ET conducted theoretical analysis of estimators. AM conducted and designed simulation studies with the estimators. AM and SB contributed to the introduction section. All authors contributed to the discussion section. AM contributed to the methods and results sections. All authors read and approved the manuscript.

2.2 Abstract

The SNP heritability of a trait is the proportion of its variance explained by the additive effects of the genome-wide single nucleotide polymorphisms (SNPs). The existing approaches to estimate SNP heritability can be broadly classified into two categories. One set of approaches model the SNP effects as fixed effects and the other treats the SNP effects as random effects. These methods make certain assumptions about the dependency among individuals (familial relationship) as well as the dependency among markers (linkage disequilibrium, LD) to provide consistent estimates of SNP heritability as the number of individuals increases. While various approaches have been proposed to account for such dependencies, it remains

unclear which estimates reported in the literature are more robust against various model misspecifications. Here we investigate the impact of different structures of LD and familial relatedness on heritability estimation. We show that the performance of different methods for heritability estimation depends heavily on the structure of the underlying pattern of LD and the degree of relatedness among sampled individuals. Moreover, we establish the equivalence between the two method-of-moments estimators, one using a fixed-SNP-effects approach, and another using a random-SNP-effects approach.

2.3 Introduction

Fundamental to the study of inheritance is the partitioning of the total phenotypic variation into genetic and environmental components Visscher et al. [2008]. Using family studies, the phenotypic variance-covariance matrix can be parameterized to include the variance of an additive genetic effect, and an environmental effect [Lynch et al., 1998]. Specific family designs, such as twin studies can accommodate both shared and nonshared environmental effects. The ratio of the genetic variance component to the total phenotypic variance is the proportion of genetically controlled variation and is termed as the 'narrow-sense heritability'. As shown in the recent review of more than 17,000 twin studies [Polderman et al., 2015], heritability provides useful information on the ability of a model to identify causal genetic markers in a genome-wide association study (GWAS), is used to estimate familial recurrence risk of disease, and informs the genetic architecture of the trait (e.g., through partitioning by genomic region or tissue-specific expression).

GWASs seek to understand the relationship between phenotypic traits and millions of single-nucleotide polymorphisms (SNPs), a type of genetic variant. Linear models are widely used in the field of statistical genetics to assess both individual and cumulative contribution of genetic variants on a trait. The individual contribution is assessed by treating each variant as a fixed effect (fixed-SNP-effect model) while adjusting for relevant covariates in a linear regression [Dicker, 2014, Schwartzman et al., 2019, Bulik-Sullivan et al., 2015] or by treating each variant as a random effect (random-SNP-effect model) by using a linear mixed effect

model [Yang et al., 2010, 2011, Speed et al., 2012]. The fixed-SNP-effect based approaches model individuals as independent, but incorporate the dependencies among the markers explicitly into the model. On the other hand, the random-SNP-effect models use the genetic relatedness among individuals to improve the efficiency of estimation of genetic variance. Nowadays, with the increasing ability to sequence many genetic variants in large cohort studies (UK Biobank Bycroft et al. [2018], Precision Medicine cohort Collins and Varmus [2015], and the Million Veterans Program Gaziano et al. [2016] are a few such examples), there is significant interest to estimate the cumulative contribution of the genome-wide causal variants. Often we assess such cumulative contribution by estimating the proportion of variance explained by the additive effects of the causal variants in the genome; that is, the "SNP heritability".

The random-SNP-effect models assume an infinitesimal model for the SNP effects and use of genome-wide SNP data on distantly related individuals [Yang et al., 2010, 2011, Lee et al., 2011, Yang et al., 2012, Lee et al., 2012, Speed et al., 2012, Bulik-Sullivan et al., 2015] to estimate the pairwise genetic relatedness between sampled individuals. These approaches assume that each causal SNP makes a random contribution to the phenotype, and these contributions are correlated between individuals who have similar genotypes. By partitioning the phenotypic covariance matrix among all individuals into a genetic similarity matrix and a random variation matrix, the approach estimates the proportional contribution of the genetics to the total phenotypic variation. The estimation of the heritability parameter heavily depends on the estimation of a high-dimensional genetic relationship matrix (GRM). The matrix is usually estimated from the observed data on M markers for all n individuals in the cohort. Two methods of estimation are used to estimate heritability under this model. One is a likelihood-based approach, which includes GCTA [Yang et al., 2010] and LDAK [Speed et al., 2012]. It uses restricted maximum likelihood (REML) estimation technique Corbeil and Searle [1976] to estimate heritability. The other approach uses method-of-moments technique to estimate heritability, such as HE regression, a method based off Haseman and Elston [1972]. A major advantage of this mixed effect model approach is that it can account for re-

lated individuals, but the general recommendation is to exclude individuals with relatedness greater than 0.025 in the estimation of heritability [Yang et al., 2011] due to shared environment. These approaches do not explicitly account for the linkage disequilibrium (LD) among the markers, and REML-based estimators have been shown to be sensitive to the patterns of LD [Speed et al., 2012].

There have been attempts to rectify such bias due to LD by partitioning the genome into regions with different LD structures and by assigning different genetic variance parameter to each partition [Evans et al., 2018b]. Such correction has been shown to improve the bias in heritability estimation for REML-based estimators. However, such corrections are often *ad hoc* and the performance depends on the underlying LD structure. Recently, Pazokitoroudi et al. [2020] used a similar partitioning strategy on the HE regression estimator. However, it is not clear if LD will impact the MOM estimator in the same way as it does the REML-based estimators. Moreover, the performance of the MOM estimators in presence of LD has not been studied extensively.

The fixed-SNP-effect approaches assume SNP effects are arbitrary and fixed [Dicker, 2014, Schwartzman et al., 2019], thus giving more flexibility to each SNP effect. The proposed estimators are consistent and asymptotically normal in high-dimensional linear models with Gaussian predictors and errors, where the number of causal predictors m is proportional to the number of observations n; in fact, consistency holds even in settings where $m/n \to \rho$, where $0 < \rho < \infty$. This set of approaches cannot easily accommodate relatives in the model, and thus the consistency of the estimator is derived under the assumption that the sampled individuals have independent genotypes. These approaches directly incorporate the LD among SNPs into the model and have been shown to provide consistent estimates of heritability under the correctly specified LD model for $n > M$. However, these methods make different approximations to derive the heritability estimator for $n < M$, since the LD matrix is not invertible then. The properties and differences between these approximation-based estimators [Dicker, 2014, Hou et al., 2019] are not well studied for $n < M$.

In this paper, we take a closer look at these random-SNP-effects and fixed-SNP-effects

model for heritability estimation using both likelihood and MOM approaches. This paper provides an analytical comparison of two popular MOM estimators from each of these categories. We aim to understand the fundamental differences or similarities between the principles of these two lines of approaches. We present a set of simulation studies with varying structures of LD and compare the performance of a wide array of estimators. We further provide some theoretical results that justify the observed simulation performance. We demonstrate through theoretical derivations as well as simulation studies that the potential impact of LD on a random-SNP-effect model based estimator [Haseman and Elston, 1972] depends on the extent and structure of correlation of the causal and non-causal variants. We also show that the fixed-SNP-effect model estimator proposed by Dicker [2014] is essentially equivalent to the Haseman-Elston method-of-moments estimator [Haseman and Elston, 1972] for $n < M$.

Our findings in this paper do not demonstrate any particular advantage of the fixed-SNP-effect models over the random-SNP-effect models in presence of LD, at least for the case when heritability is estimated using a genome-wide marker model and when $n < M$. One could partition the genome into small segments to account for the differences in genome-wide LD structure and handle the influence of LD better using a fixed-SNP-effect estimator for each partition separately [Hou et al., 2019]. However, there is a potential overfitting issue in having a separate heritability parameter for every partitioned segment.

This rest of the paper is organized as follows: first, different methods to estimate heritability are explained and analytical formulae to compare their performance under different LD structures are presented. We then describe strategies to simulate LD and relatedness structure and to evaluate both fixed-SNP-effects and random-SNP-effects models. Finally, results are presented and discussed.

Variable	Definition
n	Number of individuals in an analysis
M	Total number of markers in an analysis
m	The number of causal markers in an analysis. A marker is considered causal if it has a non-zero direct effect on phenotype.
$i,\ k$	Typically used to index individuals, $i, k = 1, ..., n$.
$j,\ \ell$	Typically used to index markers, $j, \ell = 1, ..., m$ or M.
p_j	The population frequency of locus j
σ_g^2	The variance in phenotype attributable to genotypic effects
σ_e^2	The variance in phenotype attributable to environment.
r_{jl}	The genotypic correlation between loci j and l in the population
ϕ_{ik}	Genotypic correlation between individuals i and k
$\boldsymbol{G}$	A matrix of genotypes with n rows and M columns
$\boldsymbol{\Gamma}_A$	A matrix of normalized genotypes with n rows and M columns (Equation (2.1))
$\boldsymbol{\Gamma}_C$	A matrix formed by the m columns of $\boldsymbol{\Gamma}_A$ that correspond to the causal markers
$\boldsymbol{\Psi}$	The GRM calculated using all markers; an $n \times n$ matrix. $M^{-1}\boldsymbol{\Gamma}_A\ \boldsymbol{\Gamma}'_A$.
$\boldsymbol{\Sigma}$	The $M \times M$ marker LD matrix calculated from all markers. $n^{-1}\boldsymbol{\Gamma}'_A\ \boldsymbol{\Gamma}_A$
$\boldsymbol{\Sigma}^*$	The true $M \times M$ marker LD matrix calculated from all markers. $E(n^{-1}\boldsymbol{\Gamma}'_A\ \boldsymbol{\Gamma}_A)$
$\boldsymbol{\beta}$	An m-vector of effects of causal loci on phenotype, or sometimes an M-vector augmented by 0's (Equation (2.4))
$\boldsymbol{y}$	An n-vector of phenotypic values of individuals.

Table 2.1: Glossary of notation used

2.4 Materials and Methods

2.4.1 Genotypes, Phenotypes, and Heritability Estimation

We consider a set of n individuals from a homogeneous population, typed at M SNP markers, assumed to be in Hardy-Weinberg equilibrium. Note that notation is also listed Table 2.1. Assume an n × M matrix of genotypes $\mathbf{G} = (G_{ij})$, where $G_{ij} = 0, 1, 2$ is the number of copies of the reference allele for individual i at locus j with population frequency p_j. Thus $G_{ij}, i = 1, 2, \ldots, n$, has mean $2p_j$ and variance $2p_j(1 - p_j)$, $j = 1, 2, \ldots, M$. The vector of standardized genotypes for individual i at marker j is given by

$$\Gamma_{ij} = \frac{G_{ij} - 2p_j}{\sqrt{2p_j(1 - p_j)}} \tag{2.1}$$

so that Γ_{ij} has mean 0 and variance 1.

The matrix of standardized genotypes for all markers, $\mathbf{\Gamma}_A = (\Gamma_{ij})$, carries information on the relatedness of individuals, and the LD among markers. While $\mathrm{E}(\Gamma_{ij}\Gamma_{i\ell}) = r_{j\ell}$ is the genotypic correlation between loci within an individual, $\mathrm{E}(\Gamma_{ij}\Gamma_{kj}) = \phi_{ik}$ measures the genotypic correlation between individuals. We define the GRM $\mathbf{\Psi}$ as in Yang et al. [2010]

$$\mathbf{\Psi} = M^{-1}\mathbf{\Gamma}_A \, \mathbf{\Gamma}_A{}' \tag{2.2}$$

and we define the LD matrix as

$$\mathbf{\Sigma} = n^{-1}\mathbf{\Gamma}'_A \, \mathbf{\Gamma}_A. \tag{2.3}$$

where we use the single quote ($'$) to denote the transpose of a matrix. In large samples, the empirical allele frequencies $(2n)^{-1}\sum_{i=1}^{n} G_{ij}$ can be used as an estimate of the population frequency p_j of Equation (2.1) in forming the matrices $\mathbf{\Psi}$ and $\mathbf{\Sigma}$.

Suppose that the first m of the M markers are *causal*, having a direct impact on phenotype. We denote by $\mathbf{\Gamma}_C$ the matrix consisting of the m columns of $\mathbf{\Gamma}_A$ corresponding to the causal markers, and adopt the classical trait model of Fisher [1918]. The phenotype of individual i is given by

$$y_i = \sum_{j=1}^{m} \Gamma_{ij}\beta_j + \epsilon_i \tag{2.4}$$

where β_j, Γ_{ij} and ϵ_i are mutually independent and have mean 0. Then $\mathrm{E}(y_i|\mathbf{\Gamma}_C) \equiv 0$ so that $\mathrm{var}(\mathrm{y_i}) = \mathrm{E}(\mathrm{var}(\mathrm{y_i} \mid \mathbf{\Gamma_C}))$. In SNP heritability estimation, σ_g^2 is the phenotypic variance attributable to SNPs, and σ_e^2 is the phenotypic variance attributable to environmental factors. In a random-SNP-effect model, we assume $\beta_j \sim N(0, \sigma_g^2/m)$, and $\epsilon_i \sim N(0, \sigma_e^2)$. Then

$$\mathrm{var}(\mathrm{y_i} \mid \mathbf{\Gamma_C}) \;=\; \mathrm{var}(\sum_{\mathrm{j}=1}^{\mathrm{m}} \Gamma_{\mathrm{ij}}\beta_{\mathrm{j}} + \epsilon_{\mathrm{i}}) \;=\; \sum_{\mathrm{j}=1}^{\mathrm{m}} \Gamma_{\mathrm{ij}}^2 \mathrm{var}(\beta_{\mathrm{j}}) + \mathrm{var}(\epsilon_{\mathrm{i}}) \;=\; \frac{\sigma_{\mathrm{g}}^2}{\mathrm{m}} \sum_{\mathrm{j}=1}^{\mathrm{m}} \Gamma_{\mathrm{ij}}^2 + \sigma_{\mathrm{e}}^2$$

Then either conditionally on $\mathbf{\Gamma}_C$ as m becomes large, or taking expectations over Γ_{ij}, $\mathrm{var}(\mathrm{y_i}) = \sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2$. On the other hand, a fixed-SNP-effect model assumes that $\boldsymbol{\beta}$ is a fixed quantity, with $\beta_j = 0$ for non-causal markers. In that case, we define $\sigma_g^2 = \boldsymbol{\beta}'\mathbf{\Sigma}^*\boldsymbol{\beta}$, and note $\mathrm{E}(\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{i}\ell}) = \mathrm{r_{j\ell}}$ and because of normalization, $\mathrm{E}(\Gamma_{\mathrm{ij}}^2) = 1$ so that

$$\mathrm{var}(\mathrm{y_i}) \;=\; \mathrm{var}(\sum_{\mathrm{j}=1}^{\mathrm{m}} \Gamma_{\mathrm{ij}}\beta_{\mathrm{j}} + \epsilon_{\mathrm{i}}) \;=\; \sum_{\mathrm{j}=1}^{\mathrm{m}} \sum_{\ell=1}^{\mathrm{m}} \beta_{\mathrm{j}}\beta_{\ell}\mathrm{E}(\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{i}\ell}) + \mathrm{var}(\epsilon_{\mathrm{i}}) \;=\; \sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2$$

Thus in either case, the phenotypic variance $\mathrm{var}(\mathrm{y_i}) = \sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2$ and SNP-heritability is $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$. If phenotypes are standardized to have variance 1, then $\sigma_g^2 = h^2$ and $\sigma_e^2 = 1 - h^2$. More generally, estimation of heritability is primarily concerned with estimation of σ_g^2, the estimate of h^2 being then obtained by dividing by the empirical variance of the phenotypes $y_i,\ i = 1, ..., n$.

2.4.2 Overview of Estimators

In our overview of the methodologies for heritability estimation, we concentrate on method-of moments estimation and likelihood-based estimation for the random-SNP-effect models. We further compare these estimators with the fixed-SNP-effect method of moments model based estimators [Dicker, 2014, Schwartzman et al., 2019]. The Supplementary Material (end of this chapter) provides more details on these estimators.

For the likelihood methods, we consider the GCTA [Yang et al., 2011] and LDAK [Speed et al., 2012] approaches. In brief, GCTA is a random-SNP-effect model derived under asumptions similar to those of Section 2.4.1. The approach uses REML [Patterson and Thompson,

1971] to estimate σ_g^2 and σ_e^2. It estimates heritability assuming that phenotypes are drawn from a multivariate Normal distribution, where the log likelihood function is

$$-\frac{n}{2}\log(2\pi) - \frac{1}{2}(\log\det(\sigma_g^2\boldsymbol{\Psi} + \sigma_e^2\boldsymbol{I}) + \mathbf{y}'(\sigma_g^2\boldsymbol{\Psi} + \sigma_e^2\boldsymbol{I})^{-1}\mathbf{y}) \tag{2.5}$$

LDAK [Speed et al., 2012] uses a similar model, except reweighting the SNP markers to adjust for LD. More details on the GCTA and LDAK approaches can be found in Supplementary Section S1. Note that σ_g^2 is only identifiable when $\boldsymbol{\Psi}$ is not the identity matrix.

For the method-of-moments estimators, we first considered a Haseman-Elston (HE) estimator [Haseman and Elston, 1972], an estimator from the random-SNP-effect approach category. The estimator of σ_g^2, derived in Supplementary Section S2.1, has the form

$$\widetilde{\sigma_g^2} \;=\; \frac{S_{Y\Psi}}{S_{\Psi\Psi}} \;=\; \frac{\sum_k \sum_{i<k} y_i y_k \Psi_{ik}}{\sum_k \sum_{i<k} \Psi_{ik}^2} \tag{2.6}$$

An estimate of heritability is then given by dividing by the empirical variance of phenotypes Y_i. Further properties of this estimator in the case of no LD are given in Supplementary Section S2.1.

We also considered two method-of moments estimators from the fixed-SNP-effect approach category, which we denote Dicker-1 and Dicker-2 [Dicker, 2014]. Dicker-1 is applicable in the case of no LD. It is derived and discussed in the Supplementary Section 2.2, and takes the form

$$\tilde{\sigma}_g^2 \;=\; (n(n+1))^{-1}(\|\boldsymbol{\Gamma_A}'y\|^2 \;-\; M\mathbf{y}'\mathbf{y}) \;\;=\;\; (n(n+1))^{-1}(M\mathbf{y}'\boldsymbol{\Psi}\mathbf{y} \;-\; M\mathbf{y}'\mathbf{y}) \tag{2.7}$$

We consider this estimator primarily for comparison with the HE estimator: see Supplementary Section S2.1 and Section 2.4.3.

Throughout this paper, we refer to the estimator in Equation (2.7) as Dicker-1, but we also present a variant of Dicker-1, which we denote Dicker-1-Σ. In the presence of LD, if $\boldsymbol{\Sigma}$ is invertible, Dicker-1-Σ is

$$\begin{aligned}\tilde{\sigma}_g^2 \;&=\; (n(n+1))^{-1}((\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma_A}'\mathbf{y})'(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma_A}'\mathbf{y}) \;-\; M\mathbf{y}'\mathbf{y}) \\ &=\; (n(n+1))^{-1}(\mathbf{y}'\boldsymbol{\Gamma}_A\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma_A}'\mathbf{y} \;-\; M\mathbf{y}'\mathbf{y})\end{aligned} \tag{2.8}$$

However, in many cases, $\mathbf{\Sigma}$ is not invertible because $M > n$, and hence we do not consider Dicker-1-Σ in our simulations. In these cases, to address the LD, Dicker [2014] derives an estimator which we denote Dicker-2. This estimator uses moments of the trace of the LD matrix $\mathbf{\Sigma}$ to correct for LD, resulting in an estimator of σ_g^2

$$\tilde{\sigma}_g^2 \;=\; \frac{\mu_1(\mathbf{\Gamma}'_A\mathbf{y})'(\mathbf{\Gamma}'_A\mathbf{y}) \;-\; M\mu_1^2\mathbf{y}'\mathbf{y}}{n(n+1)\mu_2} \tag{2.9}$$

where

$$\mu_1 \;=\; \frac{tr(\mathbf{\Sigma})}{M} \text{ and } \mu_2 \;=\; \frac{tr(\mathbf{\Sigma}^2)}{M} - \frac{(tr(\mathbf{\Sigma}))^2}{Mn} \tag{2.10}$$

Further details of the Dicker-2 estimator are given Supplementary Section S3.3.

2.4.3 Impact of Linkage Disequilibrium

Impact of LD on the Haseman-Elston Estimator

In this section we consider the impact of marker mispecification and marker LD on the numerator and denominator of the HE estimator, and hence on the estimate of σ_g^2. We assume unrelated individuals but correlated markers, so that $\mathrm{E}(\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{k}\ell}) \;=\; 0$ if $i \neq k$, but $\mathrm{E}(\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{i}\ell}) \;=\; \mathrm{r}_{\mathrm{j}\ell}$, with $-1 \leq r_{j\ell} \leq 1$, and $r_{jj} = 1$.

We split the markers into m causal markers C and $(M - m)$ non-causal markers F. Note that all markers are used in the GRM: $\Psi = M^{-1}\mathbf{\Gamma}_A\,\mathbf{\Gamma}'_A$, but that only causal markers $\mathbf{\Gamma}_C$ contribute to the phenotype $\mathbf{y}$. For convenience, assume that the first m markers are causal: $C = \{1, ..., m\}$ and $F = \{(m+1), ..., M\}$. Then, following the same derivation as in Supplementary Section S2.1, for $i \neq k$ we obtain,

$$\mathrm{E}(\mathrm{y_i}\Psi_{\mathrm{ik}}\mathrm{y_k}) = \mathrm{M}^{-1}\mathrm{E}\left(\sum_{\mathrm{j}=1}^{\mathrm{m}}\sum_{\ell=1}^{\mathrm{m}}\beta_{\mathrm{j}}\beta_\ell(\sum_{\mathrm{w}=1}^{\mathrm{M}}\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{iw}}\Gamma_{\mathrm{kw}}\Gamma_{\mathrm{k}\ell})\right) \tag{2.11}$$

If the β_j have mean 0, and are uncorrelated, we have only terms in $j = \ell$, and this reduces

to

$$\mathrm{E}(\mathrm{y_i}\Psi_{\mathrm{ik}}\mathrm{y_k}) = M^{-1}\mathrm{E}\left(\sum_{\mathrm{j=1}}^{\mathrm{m}}\beta_{\mathrm{j}}^2(\sum_{\mathrm{w=1}}^{\mathrm{M}}\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{iw}}\Gamma_{\mathrm{kw}}\Gamma_{\mathrm{kj}})\right) \;=\; \mathrm{M}^{-1}\mathrm{E}\left(\sum_{\mathrm{j=1}}^{\mathrm{m}}\beta_{\mathrm{j}}^2(\sum_{\mathrm{w=1}}^{\mathrm{M}}\mathrm{r}_{\mathrm{jw}}^2)\right)$$
$$= \;(mM)^{-1}\sigma_g^2\sum_{j=1}^{m}\sum_{w=1}^{M}r_{jw}^2$$

here using that individuals i and k are independent, and that β_j^2 has expectation σ_g^2/m. Then

$$\mathrm{E}(\mathrm{S_{Y\Psi}}) = \;\mathrm{E}(\sum\sum\nolimits_{\mathrm{i<k}}\mathrm{y_i}\Psi_{\mathrm{ik}}\mathrm{y_k}) \;=\; (\mathrm{n}(\mathrm{n}-1)/2)\mathrm{E}(\mathrm{y_i}\Psi_{\mathrm{ik}}\mathrm{y_k})$$
$$= \;\frac{n(n-1)\sigma_g^2}{2mM}\sum_{j=1}^{m}\sum_{w=1}^{M}r_{jw}^2 \;=\; \frac{n(n-1)\sigma_g^2}{2mM}(R_{CC} \;+\; R_{CF}) \tag{2.12}$$

where for convenience we denote the sums of squared correlations

$$R_{CC} = \;\sum_{j=1}^{m}\sum_{\ell=1}^{m}r_{j\ell}^2 \;\text{ among causal markers}$$
$$R_{CF} = \;\sum_{j=1}^{m}\sum_{\ell=m+1}^{M}r_{j\ell}^2 \;\text{ between causal and non-causal markers}$$
$$\text{and } R_{FF} = \;\sum_{j=m+1}^{M}\sum_{\ell=m+1}^{M}r_{j\ell}^2 \;\text{ among non-causal markers}$$

Considering similarly the denominator of the HE estimator,

$$\mathrm{E}(\Psi_{\mathrm{ik}}^2) \;=\; \mathrm{M}^{-2}\sum_{\mathrm{j=1}}^{\mathrm{M}}\sum_{\mathrm{\ell=1}}^{\mathrm{M}}\mathrm{E}(\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}}\Gamma_{\mathrm{i\ell}}\Gamma_{\mathrm{k\ell}})) \;=\; \mathrm{M}^{-2}\sum_{\mathrm{j=1}}^{\mathrm{M}}\sum_{\mathrm{\ell=1}}^{\mathrm{M}}\mathrm{r}_{\mathrm{j\ell}}^2$$

so that

$$\mathrm{E}(\mathrm{S}_{\Psi\Psi}) \;=\; \sum\sum\nolimits_{\mathrm{i<k}}\mathrm{E}(\Psi_{\mathrm{ik}}^2) \;=\; \frac{\mathrm{n}(\mathrm{n}-1)}{2\mathrm{M}^2}(\mathrm{R_{CC}} \;+\; 2\;\mathrm{R_{CF}} \;+\; \mathrm{R_{FF}})$$

leading finally to the ratio of expectations of $S_{Y\Psi}$ and $S_{\Psi\Psi}$

$$\frac{M}{m}\sigma_g^2\;\frac{R_{CC}+R_{CF}}{R_{CC} \;+\; 2\;R_{CF} \;+\; R_{FF}} \tag{2.13}$$

Equation (2.13) approximates the expectation of the HE estimator and gives insight into its bias. First, if there is no LD, $R_{CC} = m$, $R_{CF} = 0$, and $R_{FF} = (M - m)$, giving the

results of Section S2. Second, if the GRM contains only causal markers $M = m$, then LD among these causal markers does not cause bias, as approximated by Equation (2.13). Third, if additional markers F are not in LD with each other, nor with the causal markers C, $R_{CF} = R_{FF} = 0$, and again no bias results. Note that generally inclusion of additional markers in the GRM is less serious than omission of causal markers. If $\mathbf{\Gamma}_A$ is missing causal markers j then Equation (2.11) will not include the contributions of those β_j and $S_{Y\Psi}$ will be decreased, but $S_{\Psi\Psi}$ will not (on average) be affected, leading to underestimation of σ_g^2.

In some special cases biases cancel out. Consider first a special case of causal markers in regions of "average LD"; suppose all $r_{j\ell} = s$ for $j \neq \ell$. Then $R_{CC} = m + m(m-1)s^2$, $R_{CF} = m(M-m)s^2$ and $R_{FF} = (M-m) + (M-m)(M-m-1)s^2$ and some arithmetic shows there is no bias. Two other examples occur in the simulations of Section 2.5. In both the autocorrelation and block simulations, causal and non-causal markers are alternating. Then $M = 2m$ and $R_{FF} = R_{CC}$, and Equation (2.13) again shows there is no bias. This is demonstrated in the simulation results in Figures 2.1 and 2.2. We note that although we only show the case of $M = 2m$, we show in Supplementary Section S3.1 that the approximate theoretical bias is also quite small for other ratios of causal to noncausal markers, and there is no bias for equally sized blocks.

In other cases, there can be bias. For example, if causal markers are in regions of high LD, then (per marker) R_{CC} dominates over R_{FF}, and σ_g^2 will be overestimated, while if causal markers are in regions of low LD R_{FF} in the denominator will dominate, and σ_g^2 will be underestimated. The case of duplication of markers also considered in the simulations (Section 2.5.2) is different, and Equation (2.13) again provides an estimate of the bias. In this example, there is no LD in the m causal markers, so $R_{CC} = m$. The genotypes at subset of d of these markers are replicated r additional times, but these replicates are non-causal. So $M = m + rd$. $R_{CF} = rd$ and $R_{FF} = r^2 d$. Then Equation (2.13) reduces to

$$\frac{M}{m}\sigma_g^2 \frac{m + rd}{m + 2rd + r^2 d} \;=\; \sigma_g^2 \, \frac{(m+rd)^2}{m(m + rd(2+r))} \tag{2.14}$$

Note that if no markers are replicated ($d = 0$) or all markers are replicated ($d = m$) then there

is no bias. Note also that the result only depends on the proportion of markers replicated. If $d = gm$ then Equation (2.14) reduces to $(1+rg)^2\sigma_a^2/(1+2rg+r^2g)$. Although the expectation of the ratio is approximated by the ratio of expectations in Equation (2.13), simulation shows this approximation gives an accurate estimate of the bias: see Supplementary Section 3.1 (Figure S3).

Impact of LD on the Dicker-1 Estimator

Through our simulations, we found that the Dicker-1 estimator had a generally greater bias than the HE estimator (Section 2.5.2). This is because the Dicker-1 estimator is derived from a linear expression of quadratic forms which is inflated due to LD. On the other hand, the HE estimator is a ratio of quadratic forms, and LD inflates both the numerator and the denominator, which potentially reduces the overall effect of LD. We recall from equation 2.7 that the Dicker-1 estimator takes the form $\tilde{\sigma}_g^2 = (n(n+1))^{-1}(M\mathbf{y}'\mathbf{\Psi}\mathbf{y} - M\mathbf{y}'\mathbf{y})$. Through calculations shown in detail in Supplementary Section S3.2, we show that

$$E(\tilde{\sigma}_g^2) \approx \sigma_g^2\ (n+1)^{-1}(K + \frac{(n+M-2)}{m}(R_{CC}+R_{CF}) - M)$$

We note that $R_{CC}+R_{CF} \geq m$, and hence, from this approximation, we have that $\tilde{\sigma}_g^2$ is greater than σ_g, and the magnitude to which it is greater increases as R_{CC} and R_{CF} increase.

Equivalence of Haseman-Elston and Dicker-2

As will be shown in Section 2.5.2, estimates from the Dicker-2 and HE regression were very similar, although Dicker-2 explicitly models LD. Analytically, under certain normalization schemes, the two estimators are effectively equivalent. This suggests that efforts to correct for LD in the Dicker-2 framework do not ensure improved performance of this estimator compared to the HE estimator.

We begin by reformulating HE regression. We recall from Equation (2.6) that the HE estimator is given by

$$\widetilde{\sigma_g^2} = \frac{S_{Y\Psi}}{S_{\Psi\Psi}} = \frac{\sum_k \sum_{i<k} y_i y_k \Psi_{ik}}{\sum_k \sum_{i<k} \Psi_{ik}^2}$$

We can rewrite this in matrix form, giving us

$$S_{Y\Psi} = \frac{\mathbf{y}'\mathbf{\Psi}\mathbf{y} - \mathbf{y}'diag(\mathbf{\Psi})\mathbf{y}}{2}$$

Under Hardy-Weinberg Equilibrium, the GRM should have values that are approximately 1 on the diagonal. We assume y is normalized to have variance 1, which results in

$$S_{Y\Psi} \approx \frac{\mathbf{y}'\mathbf{\Psi}\mathbf{y} - n}{2}$$

Now we consider $S_{\Psi\Psi}$. Noting that $tr(\mathbf{\Psi}'\mathbf{\Psi}) = \sum_i \sum_j \Psi_{ij}^2$,

$$\begin{aligned} S_{\Psi\Psi} &= \frac{tr(\mathbf{\Psi}'\mathbf{\Psi}) - diag(\mathbf{\Psi})'diag(\mathbf{\Psi})}{2} \\ &\approx \frac{tr(\mathbf{\Psi}'\mathbf{\Psi}) - n}{2} \end{aligned}$$

Together, the HE estimator is approximately

$$\frac{\mathbf{y}'\mathbf{\Psi}\mathbf{y} - n}{tr(\mathbf{\Psi}'\mathbf{\Psi}) - n} \tag{2.15}$$

Now consider the equations for Dicker-2. First note that in Equations (2.9) and (2.10), $\mu_1 = 1$ since the genotypes are normalized to have variance 1. Next, we use the property of traces that $tr(ABCD) = tr(DABC)$ to calculate

$$\begin{aligned} \mu_2 &= \frac{1}{M} tr(\frac{1}{n^2}\mathbf{\Gamma}'_A\mathbf{\Gamma}_A\mathbf{\Gamma}'_A\mathbf{\Gamma}_A) - \frac{1}{Mn}\left(tr(\frac{1}{n}\mathbf{\Gamma}'_A\mathbf{\Gamma}_A)\right)^2 \\ &\approx \frac{1}{M} tr(\frac{1}{n^2}\mathbf{\Gamma}'_A\mathbf{\Gamma}_A\mathbf{\Gamma}'_A\mathbf{\Gamma}_A) - \frac{M}{n} \\ &= \frac{M}{n^2} tr(\frac{1}{m^2}\mathbf{\Gamma}_A\mathbf{\Gamma}'_A\mathbf{\Gamma}_A\mathbf{\Gamma}'_A) - \frac{M}{n} \\ &= \frac{M}{n^2} tr(\mathbf{\Psi}'\mathbf{\Psi}) - \frac{M}{n} \end{aligned}$$

Since $n \approx n+1$, for large n, we have that the Dicker 2 estimator (Equation 2.9) is approximately

$$\frac{\mathbf{y}'\mathbf{\Gamma}_A\mathbf{\Gamma}'_A\mathbf{y} - M\mathbf{y}'\mathbf{y}}{(\frac{M}{n^2}tr(\mathbf{\Psi}'\mathbf{\Psi}) - \frac{M}{n})(n)(n+1)} \approx \frac{\mathbf{y}'\mathbf{\Psi}\mathbf{y} - n}{tr(\mathbf{\Psi}'\mathbf{\Psi}) - n}$$

Which is the same as Equation (2.15).

2.4.4 Impact of Relatedness of Individuals on Moments Estimators

Under the assumption of independence of individuals, the SD of the HE estimator of σ_g^2 or of h^2 increases with the number of markers M (Supplementary Section S2). This arises because in the limit, the matrix $\mathbf{\Psi}$ converges in probability to the identity matrix, i.e. all off-diagonal terms converge in probability to 0. This leads to poor behavior of the HE estimator because the numerator and denominator of the HE estimator converge in probability to 0. However, this is an artefact of the assumption of complete independence (unrelatedness) of individuals. In any real sample, regardless of the extent of correction for population structure, there will always be variation in the degree of relatedness of individuals, even if any single pairwise relatedness measure is small. Note that the original formulation of HE estimators [Haseman and Elston, 1972] made use of the genetic similarity between known relatives. In this section, we therefore consider the case where individuals may be related, so standardised genotypes Γ_{ij} and Γ_{kj} are no longer independent. For simplicity we ignore LD: that is Γ_{ij} and $\Gamma_{k\ell}$ are independent, for $j \neq l$, whether or not $i = k$.

Under relatedness and inbreeding it remains the case that $\mathrm{E}(\Gamma_{\mathrm{ij}}) = 0$, but $\mathrm{var}(\Gamma_{\mathrm{ij}}) = (1 + \mathrm{F_i})$ [Crow and Kimura, 1970] and $\mathrm{E}(\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}}) = \phi_{\mathrm{ik}}$, where F_i is the inbreeding coefficient of individual i, and ϕ_{ik} is the relatedness of i and k, or twice the coefficient of kinship between i and k. To consider the HE estimator (2.6), for $i \neq k$,

$$\begin{aligned}
\mathrm{E}(\Psi_{\mathrm{ik}}^2) = \; & M^{-2} \sum_{j=1}^{M} \sum_{\ell=1}^{M} E(\Gamma_{ij}\Gamma_{kj}\Gamma_{i\ell}\Gamma_{k\ell}) \\
= \; & M^{-2} \sum \sum_{j \neq \ell} E(\Gamma_{ij}\Gamma_{kj}\Gamma_{i\ell}\Gamma_{k\ell}) \; + \; M^{-2} \sum_{j=1}^{M} E(\Gamma_{ij}^2\Gamma_{kj}^2) \\
= \; & M^{-2}(M(M-1))(\mathrm{E}(\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}}))^2 \; + \; \mathrm{M}^{-1}\mathrm{E}(\Gamma_{\mathrm{ij}}^2\Gamma_{\mathrm{kj}}^2)) \\
= \; & (\mathrm{E}(\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}}))^2 \; + \; \mathrm{M}^{-1}(\mathrm{E}(\Gamma_{\mathrm{ij}}^2\Gamma_{\mathrm{kj}}^2) - (\mathrm{E}(\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}}))^2)
\end{aligned} \tag{2.16}$$

Hence as $M \to \infty$ $S_{\Psi\Psi}$ tends to

$$\mathrm{E}(\sum_{\mathrm{k}}\sum_{\mathrm{i<k}}\Psi^2_{\mathrm{ik}}) \;\longrightarrow\; \sum_{\mathrm{k}}\sum_{\mathrm{i<k}}\phi^2_{\mathrm{ik}}$$

We can also calculate

$$\begin{aligned}
\mathrm{E}(\mathrm{y_i y_k}\Psi_{\mathrm{ik}}) &= \frac{1}{M}\mathrm{E}\left[\left(\sum_{\ell=1}^{\mathrm{m}}\Gamma_{\mathrm{i}\ell}\beta_\ell\right)\left(\sum_{\mathrm{w}=1}^{\mathrm{m}}\Gamma_{\mathrm{kw}}\beta_{\mathrm{w}}\right)\left(\sum_{\mathrm{j}=1}^{\mathrm{M}}\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}}\right)\right] \\
&= \frac{1}{M}\mathrm{E}\left[\left(\sum_{\ell=1}^{\mathrm{m}}\Gamma_{\mathrm{i}\ell}\Gamma_{\mathrm{k}\ell}\beta^2_\ell\right)\left(\sum_{\mathrm{j}=1}^{\mathrm{M}}\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}}\right)\right] \\
&= \frac{\sigma^2_g}{mM}\mathrm{E}\left[\left(\sum_{\ell=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{M}}\Gamma_{\mathrm{i}\ell}\Gamma_{\mathrm{k}\ell}\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}}\right)\right] \\
&= \frac{\sigma^2_g}{mM}\mathrm{E}\left[\sum_{\mathrm{j}=1}^{\ell-1}\sum_{\ell=2}^{\mathrm{m}}\Gamma_{\mathrm{i}\ell}\Gamma_{\mathrm{k}\ell}\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}} + \sum_{\mathrm{j}=\ell+1}^{\mathrm{M}}\sum_{\ell=1}^{\mathrm{m}}\Gamma_{\mathrm{i}\ell}\Gamma_{\mathrm{k}\ell}\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}} + \sum_{\ell=1}^{\mathrm{m}}\Gamma^2_{\mathrm{il}}\Gamma^2_{\mathrm{kl}}\right] \\
&= \frac{\sigma^2_g}{mM}(mM-m)(\mathrm{E}(\Gamma_{\mathrm{i}\ell}\Gamma_{\mathrm{k}\ell}))^2 + \frac{\sigma^2_{\mathrm{g}}}{\mathrm{M}}\mathrm{E}(\Gamma^2_{\mathrm{il}}\Gamma^2_{\mathrm{kl}}) \\
&= \sigma^2_g(\mathrm{E}(\Gamma_{\mathrm{i}\ell}\Gamma_{\mathrm{k}\ell}))^2 + \frac{\sigma^2_{\mathrm{g}}}{\mathrm{M}}(\mathrm{E}(\Gamma^2_{\mathrm{il}}\Gamma^2_{\mathrm{kl}}) - (\mathrm{E}(\Gamma_{\mathrm{i}\ell}\Gamma_{\mathrm{k}\ell}))^2)
\end{aligned}$$

and $S_{Y\Psi}$ tends to

$$\mathrm{E}(\sum_{\mathrm{k}}\sum_{\mathrm{i<k}}\mathrm{y_i y_k}\Psi_{\mathrm{ik}}) \;\longrightarrow\; \sigma^2_{\mathrm{g}}\sum_{\mathrm{k}}\sum_{\mathrm{i<k}}\phi^2_{\mathrm{ik}}$$

Thus, contrary to the results of Supplementary Section S2.1 for unrelated individuals, the SD of the HE estimator no longer increases as $M \to \infty$, but rather will depend on the magnitude of $(\sum_k \sum_{i<k} \phi^2_{ik})$. Although this sum may be small, if even any of the ϕ_{ik} are non-zero it is strictly positive, and eventually relatedness will bound the SD of the estimator of σ^2_g.

Relatedness poses greater problems for the Dicker-1 estimator (Equation 2.7) which involves the diagonal terms of the GRM matrix Ψ. Considering the expected quadratic form

$$\mathrm{E}(\mathrm{M}\ \mathbf{y'\Psi y}) \;=\; \sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{k}=1}^{\mathrm{n}}\mathrm{E}\left((\sum_{\mathrm{j}=1}^{\mathrm{m}}\Gamma_{\mathrm{ij}}\beta_{\mathrm{j}}+\epsilon_{\mathrm{i}})(\sum_{\mathrm{w}=1}^{\mathrm{M}}\Gamma_{\mathrm{iw}}\Gamma_{\mathrm{kw}})(\sum_{\ell=1}^{\mathrm{m}}\Gamma_{\mathrm{k}\ell}\beta_\ell+\epsilon_{\mathrm{k}})\right)$$

Now, by expanding and simplifying, even the coefficient of σ_e^2 is no longer mn but

$$\sum_{i=1}^{n}\sum_{w=1}^{m}\mathrm{E}(\Gamma_{\mathrm{iw}}^2) \;=\; \mathrm{M}(\mathrm{n}+\sum_{\mathrm{i}=1}^{\mathrm{k}}\mathrm{F_i})$$

while that of σ_g^2 is, as in Supplementary Section 2.2

$$(M-1)\;\mathrm{E}(\sum_{\mathrm{i}}\Gamma_{\mathrm{ij}}^2\Gamma_{\mathrm{iw}}^2)+\mathrm{E}(\sum_{\mathrm{i}}\sum_{\mathrm{k}}\Gamma_{\mathrm{ij}}^2\Gamma_{\mathrm{kj}}^2) \;=$$
$$(M-1)\sum_{i=1}^{n}(\mathrm{E}(\Gamma_{\mathrm{ij}})^2)^2+\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{E}(\Gamma_{\mathrm{ij}}^4) \;+\; \sum_{\mathrm{i}}\sum_{\mathrm{k}\neq\mathrm{i}}\mathrm{E}(\Gamma_{\mathrm{ij}}^2\Gamma_{\mathrm{kj}}^2)$$

This expectation now involves not only $(1+F_i)^2$, and ϕ_{ik}^2 but also higher order moments.

Although the derivation of distributional properties of the Dicker method-of-moments estimators depends critically on the assumption of $2n$ independent genomes, there is nothing in the derivation of Section 2.4.3 that assumes $\mathbf{\Psi}$ is diagonal. Indeed, the trace equation

$$n^2\;tr(\mathbf{\Sigma}^2) \;=\; tr(\mathbf{\Gamma}'_A\;\mathbf{\Gamma}_A\;\mathbf{\Gamma}'_A\;\mathbf{\Gamma}_A) \;=\; tr(\mathbf{\Gamma}_A\;\mathbf{\Gamma}'_A\;\mathbf{\Gamma}_A\;\mathbf{\Gamma}'_A) \;=\; M^2\;tr(\mathbf{\Psi}^2)$$

used in showing the approximate equivalence of the HE and Dicker-2 estimators, suggests that the Dicker-2 accommodation of LD in the absence of relatedness is alternatively accommodating relatedness in the absence of LD. Thus, as will be seen in the results of Section 2.5.4, the close equivalence of the Dicker-2 and HE estimators should hold under relatedness, and, as seen from equation (2.16) above, the standard deviation will no longer increase indefinitely as $M\to\infty$.

2.4.5 Simulation strategy

We performed simulation studies to assess the impact of LD structure and relatedness of individuals on heritability estimation. Each simulated data set consisted of genotypes $\mathbf{G}$ at M markers (m causal markers) for n unrelated individuals. The marker allele frequencies were those of a randomly chosen subset of markers from the 1000 Genomes Project from Chromosome 1 in the African population [Clarke et al., 2017]. This set of frequencies was

filtered to have allele frequency less than .95 and greater than .05 and was fixed over data set simulations.

Genotypes are standardized using their empirical allele frequencies. Phenotypes were simulated for n individuals, given their genotypes at the m causal markers, in accordance with the linear model of Equation (2.4):

$$y_i = \sum_{j=1}^{m} \Gamma_{ij}\beta_j + \epsilon_i \quad \text{for } i = 1, ..., n \tag{2.17}$$

For the chosen value of h^2, $(0 < h^2 < 1)$, the m-vector of genetic effects $\boldsymbol{\beta}$ was simulated with independent components $\beta_j \sim N(0, h^2/m)$ for $j = 1, ..., m$. The independent residual effects $\epsilon_i \sim N(0, 1 - h^2)$ for $i = 1, ..., n$. Thus, for the purposes of the simulation $\sigma_g^2 = h^2$, $\sigma_e^2 = 1 - h^2$, and $\text{var}(y_i) = 1$, with h^2 set to 0.8 for all simulations (see Section 2.4.1).

We implemented the Dicker and Haseman Elston estimators in R Version 4.0.2 as described in Supplementary Section S2. We used GCTA [Yang et al., 2011] and LDAK [Speed et al., 2012] as representative likelihood estimators, both of which are described in more detail in Supplementary Section S1. For every simulated data set, we applied each of these estimators. We also report a gold standard estimator to assess the performance of these different methods. The gold standard estimate is calculated assuming we know the true values of $\boldsymbol{\beta}$: the empirical variance of $\boldsymbol{\Gamma_C\beta}$ is divided by the empirical variance of the phenotypes. This gold standard estimator can be expressed as

$$\frac{(\boldsymbol{\Gamma}_C\boldsymbol{\beta} - \overline{\boldsymbol{\Gamma}_C\boldsymbol{\beta}})'(\boldsymbol{\Gamma}_C\boldsymbol{\beta} - \overline{\boldsymbol{\Gamma}_C\boldsymbol{\beta}})}{(\mathbf{y} - \overline{\mathbf{y}})'(\mathbf{y} - \overline{\mathbf{y}})} \tag{2.18}$$

In simulation study 1, we assessed the impact of different LD structures on heritability estimation. We generated genotypes assuming three kinds of LD structure: autocorrelated, block, and repeat. More details of the LD structures are given in Supplementary Section S4. Each data set was simulated with a new $\boldsymbol{\beta}$ and $\mathbf{G}$. For each LD structure, we studied the impact of both sample size n and the number of causal markers m on heritability estimation, and simulated data sets at five levels of LD. For each LD structure and level, we generated 500 simulated data sets.

For the autocorrelation and block structures, we considered the following combinations of n and m: (1) $n = 1000, m = 100$, (2) $n = 200, m = 500$, (3) $n = 200, m = 1500$, and (4) $n = 2000, m = 500$. Comparing (1) and (2) provides insight on differences in estimates of h^2 depending on if $n > m$ or $m > n$, whereas (2) and (3) compares estimates with different number of causal markers, and (2) and (4) compares estimates with different numbers of individuals. We first generated genotypes at $M = 2m$ markers. We used marker correlations $\rho = 0, 0.2, 0.4, 0.6,$ and $0.8.$, as detailed in Supplementary Section S4 (Note that $\rho = 0$ is the no-LD case.) The markers were then assigned to be alternating causal and non-causal ($m = M/2$).

For the repeat structure, we considered the cases: (1) $n = 1000, m = 200$, (2) $n = 200, m = 1000$, (3) $n = 200, m = 3000$, and (4) $n = 2000, m = 1000$. In this case, we first simulated genotypes for the m independent causal markers. The genotypes at the first 10% of markers were then repeated r times, where $r = 0, 2, 4, 6,$ or 8. (Note that $r = 0$ is the no-LD case.) The repeat copies of the markers are non-causal, so the number of non-causal markers is $0.1rm$, and $M = m + 0.1rm$. In Supplementary Figure S3 (panels C & F) the first m markers are causal, and the last $(M - m)$ are the non-causal repeat copies.

In simulation study 2, we investigated the behavior of likelihood models by plotting log-likelihood values (Equation 2.5) as a function of σ_g^2 and σ_e^2. The GRM $\mathbf{\Psi}$ in Equation (2.5) was calculated using Equation (2.2). Of interest was the relationship between the shape of the log-likelihood function and the number of individuals and causal markers, and the shape of the likelihood as the number of repeats increased. From the results of simulation study 1, we hypothesized that the shape would be different when $m > n$, where GCTA underestimated heritability, compared to when $m < n$, where GCTA overestimated (comparing (i) and (iv) of Figure 2.3). The combinations of numbers of markers and individuals was the same as with the repeats in Simulation Study 1, and the allele frequencies were taken from the AFR sample of the 1000 Genomes Project, as before. We include plots with no repeated markers (Figure 2.4A) and with 10% of the markers repeated 8 times (Figure 2.4B).

The log likelihoods minus the maximum log likelihood were plotted. Likelihoods were

truncated at the 60% quantile of B(i) for rows (i) and (iv), and at the 60% quantile of A(ii) for rows (ii) and (iii). These cutoffs were chosen because they were the plots that had the lowest 60% quantile. Plots were generated for (i) $n = 1000, m = 200$ (ii) $n = 200, m = 1000$, (iii) $n = 200, m = 3000$ (iv) $n = 2000, m = 1000$, in following with simulation study 1. We averaged log likelihoods of 100 simulated data sets with grid spacing 0.05. Due to differences in ranges, there is a shared color bar between (i) and (iv), and a different shared color bar between (ii) and (iii). A red dot is used to mark the location of the maximum log likelihood.

In simulation study 3, we assessed the impact of related individuals on heritability estimation. We simulated 1st, 2nd, and 3rd cousins using the rres package in R [Wang et al., 2017] as well as unrelated individuals to illustrate our findings in Section 2.4.4. The segment length option in rres was set to 3000 centimorgans. Using the same set of allele frequencies as previously, we simulated marker genotypes for 400 individuals, in 10 40-ships. A k-ship is defined to be a set of k cousins related to a certain degree. Each cousinship is unrelated with all other cousinships. The number of markers ranged from 400 to 4000 in steps of 400. Phenotypes were generated using Equation (2.17). For every combination of cousinships and number of markers, we simulated 500 sets of 10 40-ships. A visualization of the GRM of the dataset as shown in Supplementary Figure S4, using 1000 markers.

2.5 Results

2.5.1 Simulation Study 1: Bias and Variance when $\rho = 0$ or $r = 0$ (no LD)

The special case of no LD in Simulation Study 1 is shown in Figures 2.1, 2.2, and 2.3 in panels of the upper row at the left-hand point of each point. These figures verify that the estimators were generally unbiased in estimating the heritability. One exception is in LDAK, where when $n = 200$, LDAK seemed to underestimate heritability.

Although we generally observed no bias in the estimators under independent markers, we saw that the estimators had a wide range of variances. In the cases $n = 1000, M = 200$ and $n = 2000, M = 1000$ (columns (i) and (iv) in Figures 2.1-2.3), the variance of the GCTA,

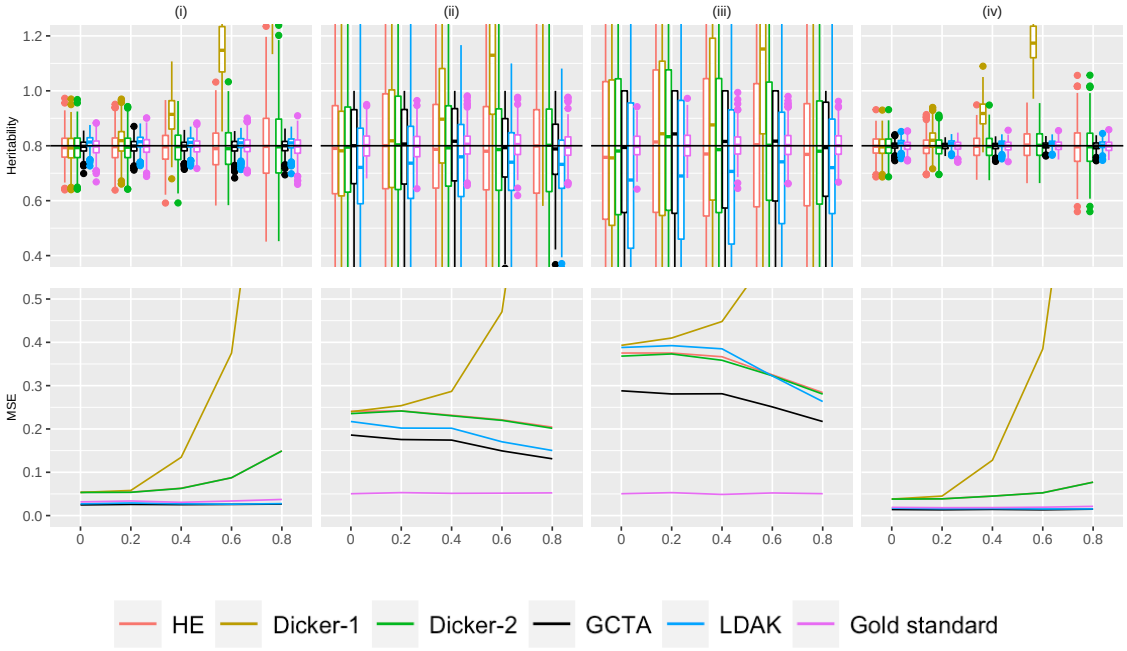

Figure 2.1: Simulation Study 1A (autocorrelated markers). On the top row, the X-axis plots the parameter ρ, the autocorrelation correlation coefficient between simulated markers as described in Supplementary Section S4. Estimates of h^2 using different estimators are plotted along the Y- axis. The value n refers to the number of individuals simulated. The value M is the total number of markers simulated, where half of the markers are causal, set in an alternating fashion, as described in Section 2.4.5. We consider (i) $n = 1000, m = 100$ (ii) $n = 200, m = 500$, (iii) $n = 200, m = 1500$ (iv) $n = 2000, m = 500$. 500 data sets were simulated for each condition. A horizontal line is shown at $h^2 = .8$, the simulated truth. On the bottom row, the X-axis is the parameter ρ, and the MSE of each of the estimators is plotted on the Y-axis.

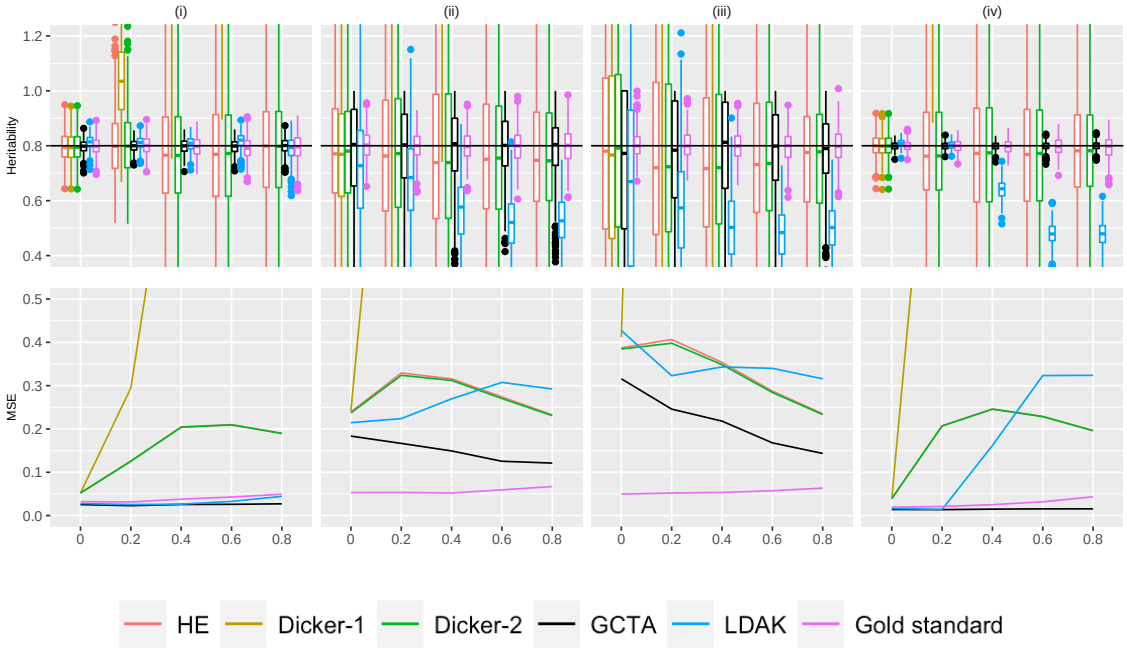

Figure 2.2: Simulation Study 1B (block markers). On the top row, the X-axis plots the parameter ρ, the block correlation coefficient between simulated markers as described in Supplementary Section S4. Estimates of h^2 using different estimators are plotted along the Y- axis. The value n refers to the number of individuals simulated. The value M is the total number of markers simulated, where half of the markers are causal, set in an alternating fashion, as described in Section 2.4.5. We consider (i) $n = 1000, m = 100$ (ii) $n = 200, m = 500$, (iii) $n = 200, m = 1500$ (iv) $n = 2000, m = 500$. 500 data sets were simulated for each condition. A horizontal line is shown at $h^2 = .8$, the simulated truth. On the bottom row, the X-axis is the parameter ρ, and the MSE of each of the estimators is plotted on the Y-axis.

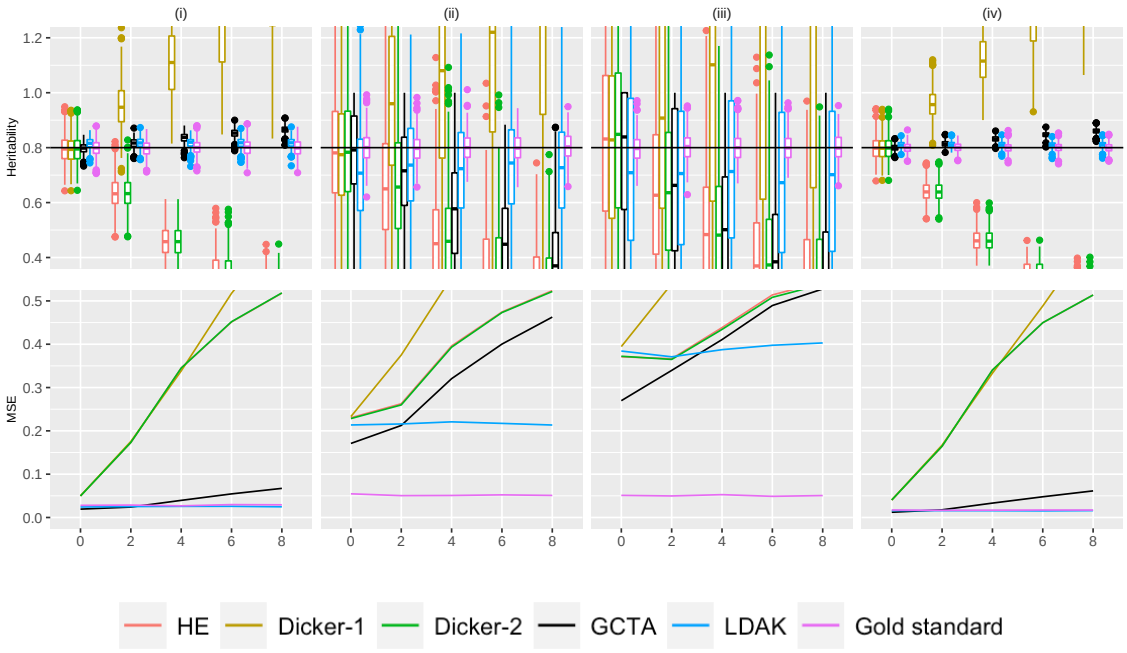

Figure 2.3: Simulation Study 1C (repeated markers). On the top row, the X-axis plots the parameter r, the number of times that 10% of the markers are being repeated as described in Supplementary Section S4. Estimates of h^2 using different estimators are plotted along the Y- axis. The value n refers to the number of individuals simulated. The value m is the total number of causal markers simulated, as described in Section 2.4.5. We consider (i) $n = 1000, m = 200$ (ii) $n = 200, m = 1000$, (iii) $n = 200, m = 3000$ (iv) $n = 2000, m = 1000$. 500 data sets were simulated for each condition. A horizontal line is shown at $h^2 = .8$, the simulated truth. On the bottom row, the X-axis is the parameter r, and the MSE of each of the estimators is plotted on the Y-axis.

and LDAK estimators were lower than the variance of the moments estimators, but this difference is less pronounced in the cases where $n = 200$ (columns (ii) and (iii)), which may suggest that the number of individuals affects the likelihood based estimators more than the moments based estimators. The lower variance resulted in lower MSE for GCTA for all conditions with $\rho = 0$, but the bias in LDAK caused it to have comparable MSE to the moments estimators when $n = 200$ (Figures 2.1, 2.2, and 2.3).

We can also compare cases when the number of individuals is kept constant while the number of markers is increased by comparing $n = 200, m = 500$ in column (ii) vs $n = 200, m = 1500$ in column (iii). In Supplementary Section S2.1, we found that with unrelated individuals and independent markers, the standard deviation of heritability should be asymptotically proportional to $\sqrt{M}/n$ in the case of the Haseman Elston estimator. Accordingly, since $M = 2m$ in the simulations, when the number of causal markers increases, the standard deviation of the heritability estimates increased as well. This is shown in both Figures 2.1, 2.2, and 2.3, where MSEs were higher in column (iii) compared to column (ii). This trend appeared to hold true for both the likelihood based estimates and the moments based estimates.

We can compare cases when the number of individuals increased while holding the number of markers constant by comparing $n = 200, m = 500$ in column (i) vs $n = 2000, m = 500$ in column (iv). The variance and MSE of the heritability estimates decreased for all estimators, which agreed with the theoretical result for the Haseman Elston estimator.

Finally, some of the biases in the behavior of LDAK may be that the LDAK model does not match our generative model. LDAK reweights their genotypes using $X_{ij} = (G_{ij} - 2f_j) \times [2f_j(1 - f_j)]^{\alpha}$, and α is recommended to be 1.25 [Speed et al., 2012, 2017]. More details can be found in Supplementary Section S1. Our model doesn't explicitly simulate phenotypes in this manner, however. To investigate this, we also chose α in LDAK to be -1, which matches our simulated phenotypes due to our normalization scheme (Equation 2.1). Results (not included) were largely similar, although the estimated heritability was slightly closer to the simulated truth in the repeat case.

2.5.2 Simulation Study 1: Impact of marker LD

(a) Autocorrelation Structure: Data were simulated using the autocorrelation structure as described in Section 2.4.5, and a representative set of moments and likelihood estimators are evaluated on these simulated data. The estimated variance and bias of different estimators is shown in Figure 2.1.

The HE estimator and the Dicker-1 estimator do not explicitly account for LD structure, and because the Dicker-1 estimate was developed for the no-LD case, it shows bias when LD is present. In the top row of Figure 2.1, the Dicker-1 estimator shown in gold showed an increase in bias as ρ increased for all of (i)-(iv). Consequently, the MSE of the Dicker-1 estimator increases rapidly compared to all of the other estimators as we increase ρ (bottom row of Figure 2.1). In contrast, there was no increase in the MSE of the HE estimator when markers were autocorrelated, agreeing with Section 2.4.3. In the top row of Figure 2.1, the estimates of h^2 from the HE estimator did not appear to visually differ significantly from the true value of 0.8. This estimate behaved very similarly to the Dicker-2 estimator, despite the Dicker-2 estimator explicitly attempting to correct for LD. This is analytically shown in Section 2.4.3

The likelihood estimators in Figure 2.1 showed generally lower MSE and no obvious bias. The GCTA estimator is shown in black and the LDAK estimator is shown in light blue. Both of these estimators seemed to have lower MSE across all values of ρ than the moments estimators, as seen in the bottom row.

When $n = 200, m = 3000$, as ρ increased, there was a decrease in the MSE in all the estimators except the Dicker-1 estimator. In Figure 2.1, it can be seen that as ρ increases, the first and third quartiles of the estimates of h^2 decrease. It has previously been shown that fewer causal markers leads to decreased variance [Dicker, 2014], and hence this effect may be driven by a decrease in the effective number of markers as LD increases.

(b) Block Structure: Figure 2.2 shows the estimated variance and bias in different estimators when the genotypes were simulated from the block structure with parameter ρ,

as described in Section 2.4.5. Similarly to the autocorrelation structure, the Dicker-1 estimator had significant bias and high MSE, although this is expected because Dicker-1 as implemented here relates to the no LD case. The HE and Dicker-2 estimators were not as affected by the LD. In contrast to the autocorrelation, however, LDAK underestimated h^2 in Figure 2.2, columns (ii), (iii), and (iv). In the bottom row of Figure 2.2, this resulted in an MSE that was comparable to that of HE and Dicker-1. GCTA estimates appeared to still largely be unbiased and produced MSEs that were lower than the other estimators. Again, it was observed that there are cases when the MSE decreases as ρ increases, similarly to the autocorrelation case.

(c) Repeat Structure: Figure 2.3 shows the variance and bias patterns when the genotypes were simulated from the repeat structure with parameter r, as described in Section 2.4.5. As r increases, the number of times that 10% of the markers were simulated increased. There were m causal markers simulated and n individuals. For example, when $m = 1000$ and $r = 3$, there were 1000 causal markers that were simulated, and the first 100 markers were repeated 3 times, leading to a total of 1300 markers that were entered into the analysis. An increased value of r indicates more markers that are in perfect LD with the original causal markers. We also examined behavior when repeated markers had a small probability of not being exact duplicates, and results were similar but less pronounced (results not shown).

As in Figures 2.1 and 2.2, the estimates for Dicker-1 increase rapidly as r increases, agreeing with analytical calculations from Section (2.4.3). In contrast to Figures 2.1 and 2.2, in Figure 2.3, the estimates for HE and Dicker-2 decrease as r increases, corresponding to Equation (2.14) and to results in Supplementary Figure (2.9), where those equations were verified through simulation. The MSE of these two estimators also increase as r increases (Figure 2.3) and further produce very similar estimates, agreeing with analytical calculations from Section 2.4.3.

In Figure 2.3, the GCTA estimator produces estimates that are greater than $h^2 = 0.8$ when $n = 1000, m = 200$ and when $n = 2000, m = 1000$, but produces estimates that are lower than 0.8 when $n = 200, m = 1000$, and when $n = 200, m = 3000$. In other words, if

$n > m$, then the GCTA estimator is underestimating, and when $n < m$, the GCTA estimator is overestimating.

The LDAK estimator shows the same pattern of bias as GCTA in that as r increases, h^2 is underestimated when $n > m$ and overestimated when $n < m$. This is bias is less pronounced than with GCTA, however. In the bottom panel of Figure 2.3, it can be seen that as r increases, the MSE of LDAK appears relatively constant, whereas the MSE of GCTA is increasing when $n = 200, M = 1000$ or $n = 200, M = 3000$, as seen in columns (ii) and (iii).

2.5.3 Simulation Study 2: Likelihood Surfaces

In Figure 2.3, GCTA displayed an upward bias when $n > m$, and a downward bias when $n < m$. We hence hypothesized that the likelihood would be different if $n > m$ versus if $m > n$. The likelihood surface captures the joint likelihood of σ_e^2 and σ_g^2. From the model in Equation (2.5), $Var(y_i) = \sigma_e^2 + \sigma_g^2$. Hence we expect that the maximum likelihood lies on a diagonal, as $\sigma_g^2 \approx Var(y_i) - \sigma_e^2$. This appears to be true when the number of individuals is much larger than the number of markers, but when the number of individuals much less the number of markers, the axis of the conditional maxima becomes more horizontal (Figure 2.4). An intuition for this result is that as the number of individuals improves, we have better knowledge of the total phenotypic variance.

The likelihood surfaces also demonstrate a faster rate of change in the likelihood surface when the number of individuals is increased, comparing Figure 2.4A and G where the range of the colors is greater than in C and E. This observation corresponds with simulation study 1, where as the number of individuals increased, the variance of the estimates of heritability decreased. Finally, on the right hand side of Figure 2.4, the surfaces are still either diagonal or horizontal, but the maxima (red dots) are shifted. This agrees with simulation study 1 results, where there was bias in the GCTA method when the number of repeats increased.

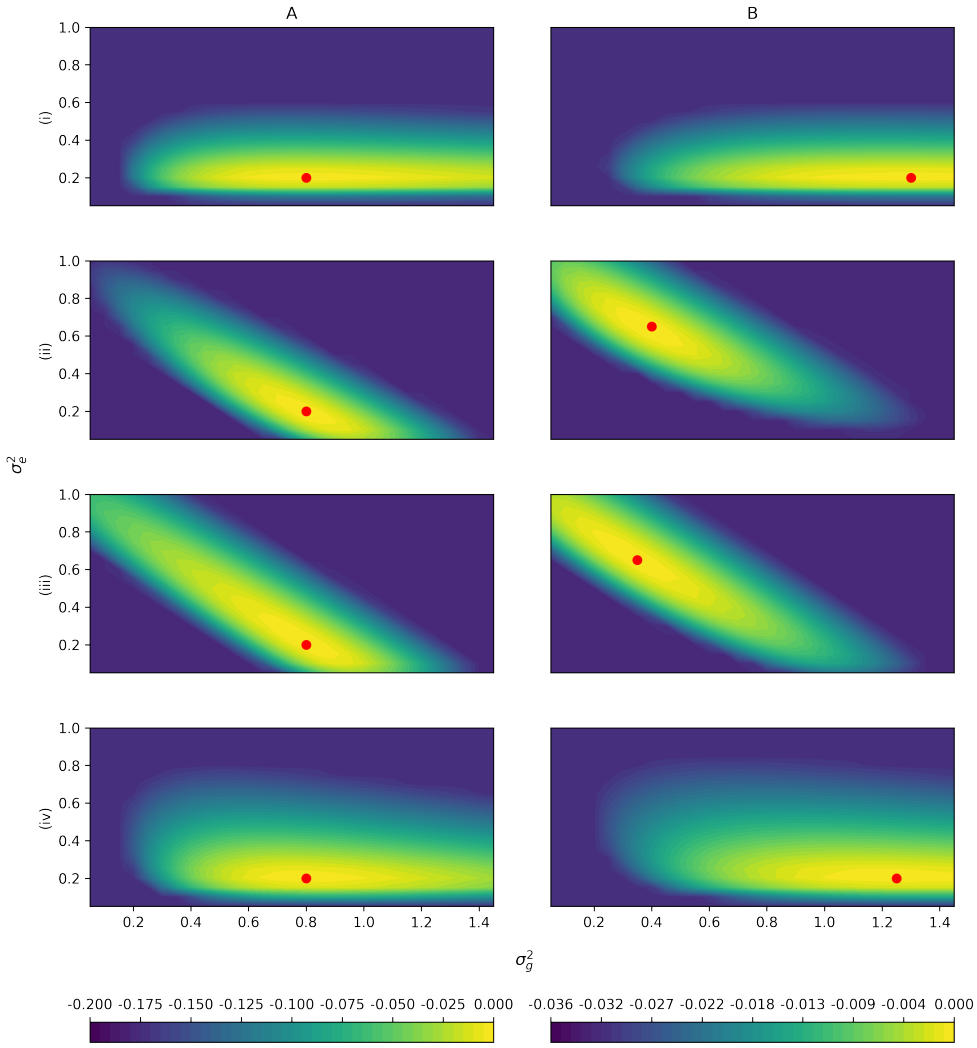

Figure 2.4: Simulation Study 2. The difference of the log likelihood from the maximum log likelihood is plotted. The colors depict the value of the difference from the maximum log likelihood. Likelihoods are truncated at the 60% quantile of B(i) for rows (i) and (iv), and at the 60% quantile of A(ii) for rows (ii) and (iii) for visibility. Row labels correspond with Figure 2.3. Column A has markers with no LD, and in column B, 10% of the markers are repeated 8 times, corresponding the the rightmost points in Figure 2.3. The average of 100 independent simulations using a grid with spacing 0.05 is plotted in each panel. Note that there is one color scale shared between (i) and (iv) on the left, and a different color scale shared between (ii) and (iii) on the right due to different ranges. The red point indicates the location of the maximum likelihood.

2.5.4 Simulation Study 3: Impact of relatedness in individuals

In simulation study 3, we studied the effect of familial structure on estimates of heritability using cousinships and found that an increase in the number of causal markers generally increased MSE unless relatedness was high.

The Dicker-2, HE, and GCTA estimators appeared unbiased for each of the relatedness structures (Figure 2.5). For GCTA and HE, we reasoned that because their model is conditional on the GRM, it took into account relatedness. Furthermore, because we've shown that HE and Dicker-2 are equivalent (Section 2.4.3), we can also explain the unbiasedness of Dicker-2. LDAK was also largely unbiased in the case of unrelated individuals, but in the case of 1st cousins, as the number of markers increased, we observed that LDAK started showing downward bias.

For the different relatedness structures (unrelated, full sibs, first cousins) we considered, we observed similar pattern in the change of MSE as we increased the number of markers. MSE was generally the lowest when the number of markers was closer to the sample size. However as we increased the number of markers, MSE for each estimator increased (Figure 2.6). For HE and Dicker-2, the unrelated individuals had the lowest MSE when the number of markers was 400, but increased as the number of markers increased. On the other hand, 1st cousins had MSE that remained steady (Figure 2.6). When the number of markers was 4000, the MSE of the unrelated individuals was larger than the MSE of the 1st cousins, agreeing with analytical calculations from Section 2.4.4.

For each of our estimators, unrelated individuals had the highest MSE and first cousins had the lowest MSE. Further, comparing the case of related to unrelated individuals, the MSE increased more slowly with the increase in the number of causal markers in related individuals.

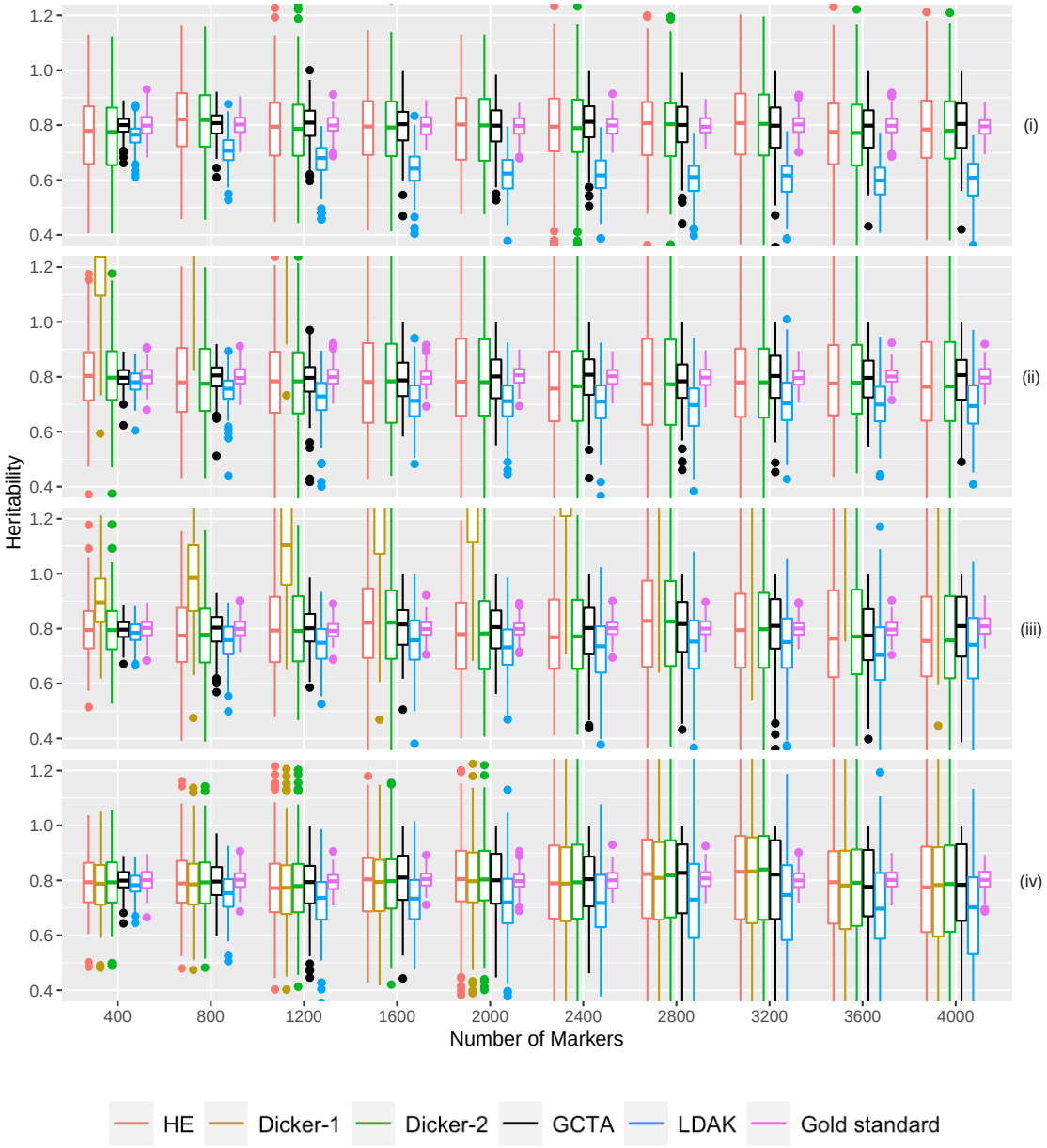

Figure 2.5: Simulation Study 3. Estimated h^2 from 500 sets of 10 groups of 40 related cousins plotted on the y-axis. The number of causal markers plotted on the x-axis. Data was simulated as described in Section 2.5.4 Different estimators are plotted in different colors. True heritability was set to be 0.8. Note that because of the chosen range of y values, Dicker-1 is sometimes not visible in the figure. Panels (i), (ii), (iii), and (iv) are first-, second-, third-cousins, and unrelated individuals respectively.

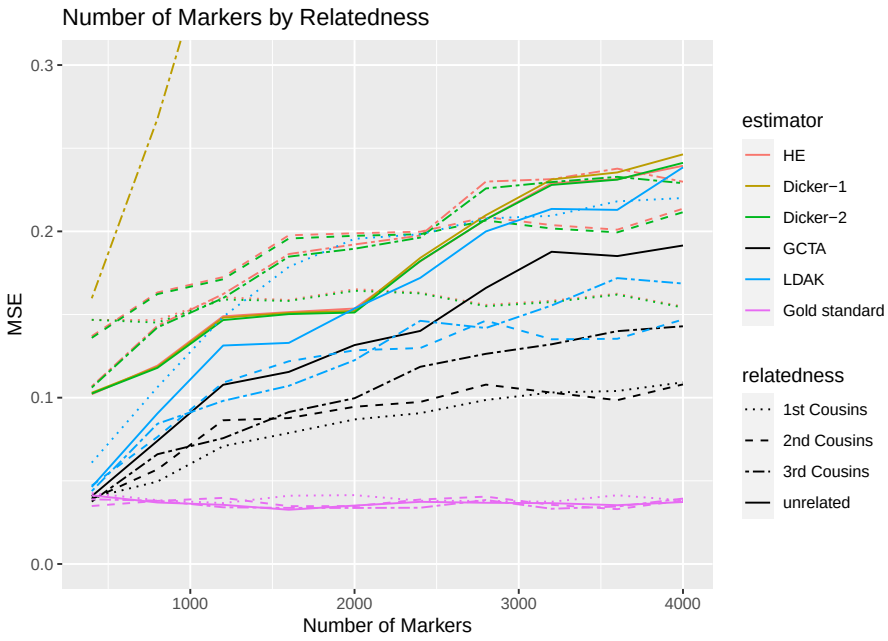

Figure 2.6: Simulation Study 3. MSE of estimates of h^2 from Figure 2.5. The X-axis indicates the number of markers in the simulation, and the Y-axis indicates the mean square error.

2.6 Discussion

The methods for SNP heritability estimation can be broadly classified into two groups; fixed-SNP-effects models and the random-SNP-effects models. The fixed-SNP-effect models [Dicker, 2014, Schwartzman et al., 2019] can more easily accommodate the LD structure among the genetic variants and can accommodate variants as both causal or non-causal. However, these approaches rely on independence among the individuals in the sample. On the other hand, the random-SNP-effect models [Yang et al., 2011] can accommodate and borrow power from related individuals, though it is generally recommended to exclude relationships with higher relatedness than 0.025 (this corresponds approximately to relatives second cousins or closer) to avoid confounding due to shared environments. These random-SNP-effects models assume all variants are causal and the majority of the methods do not accommodate LD among the markers in a statistically rigorous way. The asymptotic properties of these heritability estimators depend on model assumptions. In this paper, we have studied the impact of model misspecification on heritability estimation through extensive simulation studies. We have simulated data under various LD structures and have allowed a certain portion of the variants to be non-causal. We found little difference in the performance of a fixed-SNP-effect model method-of-moments estimator and a MOM estimator from a random-SNP-effect model under different model misspecification.

We have derived the analytic expression for the approximate bias of the HE estimator in presence of LD among markers. Section 2.3 considers various scenarios for the LD among causal and non-causal markers and analytically shows the impact of this correlation on the HE estimator. Our simulation studies and numerical results have also considered various LD scenarios to illustrate that the bias in heritability estimation depends on the underlying LD pattern and is often small. In many cases, the standard practice of pruning markers to reduce LD [Calus and Vandenplas, 2018] may be unnecessary.

In the case where Σ^{-1} can be computed ($M < n$), Dicker [2014] proposes a heritability estimator (Dicker-1) that can account for the LD among markers by rotating the genotypes.

The derivation of the consistency of the estimator, however, relies on the Normality assumption. In case of large n and M ($M >> n$), our simulation studies and analytical derivation in Section 2.3.3 show that the Dicker-2 estimator (fixed-SNP-effect model based estimator) and HE estimator (random-SNP-effect model based estimator), are essentially the same. Hence, in the situation $M >> n$, Dicker-2 estimator has limited ability to correct for the LD among markers because the HE estimator has bias under some forms of LD, as shown in Equation (2.13). This is a contradiction to the claim that the Dicker [2014] always provides an improved estimator of heritability in presence of LD among markers.

Further estimators have been proposed in, for example, Hou et al. [2019], which proposes the h^2_{GRE} estimator. This estimator's goal is to relax assumptions about the LD structure of the data by giving each causal effect its own SNP-specific variance, and has been shown to provide some robustness to LD structures. In the case that the LD matrix is estimable ($n > M$), if no binning is used, the h^2_{GRE} estimator is approximately equivalent to the Dicker-1-Σ estimator if $n \to \infty$ and M remains constant (Supplementary Section S5), but expands the scope of the Dicker-1-Σ estimator by using a pseudoinverse. This allows the h^2_{GRE} estimator to be used in cases when some markers are in perfect LD, which was not possible with the Dicker-1-Σ estimator. The h^2_{GRE} estimator also corrected bias in the Dicker-1-Σ estimator in our simulations for a finite number of individuals. Furthermore, we found that in some cases, the h^2_{GRE} estimator has lower variance than the Dicker-1 estimator even if there is no LD (Supplementary Figure S6). This is possibly due to the use of the empirical Σ in h^2_{GRE} estimator which may reduce the variance of the estimate. We note, however, that the h^2_{GRE} estimator is not defined if $q = n$, where q is the rank of $\mathbf{\Sigma}$ (Supplementary Section S5). This situation may arise in the case that $n < M$. We did not study h^2_{GRE} in detail because we aimed to analytically understand the simple estimators (estimators without any binning or weighting).

Another estimator that demonstrated robustness to some forms LD was proposed in Pazokitoroudi et al. [2020]. This estimator aimed to expand upon the HE estimator by allowing partitioning of heritability to multiple variance components. These partitioning

methods can be *ad hoc*, but have been shown to improve robustness of estimators to MAF and LD in some cases [Evans et al., 2018b]. In the case that the genome is not partitioned, this estimator reduces to the HE estimator (Supplementary Section 6). We did not consider partitioning in this paper so that we would be able to more easily understand the estimators analytically.

Fixed-SNP-effect model based estimators generally assume that sampled individuals are independent. These approaches do not accommodate related individuals in the heritability estimation. We demonstrate that even in the absence of LD, the Dicker-1 is severely biased in the presence of related individuals. However, because of its equivalence to the HE estimator, the Dicker-2 estimator generates consistent estimates of heritability with related individuals in the absence of LD.

The likelihood based approaches from the random-SNP-effects model category, especially the LDAK approach showed more bias under certain model misspecification as compared to the MOM estimators. Under different LD structures, the traditional GCTA approach showed more stability in terms of both bias and precision over the LDAK estimator. We did not observe any specific advantage of adjusting for LD by using the LDAK estimator.

Under the assumption of independence of individuals, the standard errors of the heritability estimator increases with the number of causal markers. This is an artefact of the assumption of complete independence (unrelatedness) of individuals. In any real sample, regardless of the extent of correction for population structure, there will always be variation in the degree of relatedness of individuals, and the extent of variation would depend on the nature of relatedness present in the sample. As shown in Simulation 3, the precision of the heritability estimators improve if we include relatives in the sample. The MSE of the estimators were generally lower when we had certain relatedness present in the sample. Moreover, the impact of increasing the number of markers on MSE was significantly less pronounced if we had relatedness in the sample. Hence, we highly recommend to at least include second cousins, if present in the study sample, in the SNP heritability estimation. If the study sample has substantial number of first cousins, it may be beneficial to assess the sensitivity

of the heritability estimate after inclusion of first cousins.

In general, MOM estimators had much larger standard errors compared to the likelihood-based estimators. However, the computational gain of these MOM estimators over the likelihood estimators is significant for large n and M and often outweighs limitation of large standard error. There was no apparent bias in these estimators besides the repeat structure in Simulation 1C. For repeat structures of the causal markers, we observed underestimation in HE regression and a small upward bias for GCTA estimator.

2.7 Supplementary material

S1 Likelihood-based Approaches:

The **GCTA REML** [Yang et al., 2011] estimator is derived by assuming that random-SNP-effects $\boldsymbol{\beta} \sim N(0, \sigma_g^2 \boldsymbol{I}_{m\times m})$ and that the normalized genotypes $\boldsymbol{\Gamma}$ are fixed. It assumes a random-SNP-effect model based approach for generation of phenotypes, and uses a Euclidean distance kernel for GRM calculation. Using the Normality assumption of $\boldsymbol{\beta}$, the GCTA REML estimator assumes that $y \sim N(0, \sigma_g^2 \Psi + \sigma_e^2 I)$ and uses a restricted maximum likelihood (REML) approach to estimate σ_g^2 and σ_e^2. Recently, binning methods, such as in GCTA-LDMS have been used to apply GCTA on markers binned for different linkage disequilibrium (LD) structures or for different allele frequencies [Yang et al., 2015]. However, such binning techniques are somewhat adhoc and are not incorporated in our simulation and analytical derivations.

The **LDAK** [Speed et al., 2012, 2017] estimator uses a similar approach to the GCTA REML estimator, also assuming fixed genotypes and random $\boldsymbol{\beta}$. The LDAK model tries to correct for uneven LD by computing a reweighted GRM as in Equation (2.19).

$$X_{ij} = (G_{ij} - 2f_j) \times [2f_j(1 - f_j)]^{\alpha} \tag{2.19}$$

The value $\alpha = -1.25$ is reported to generally work well with genomewide LD structure. Each of the raw genotypes is then weighted by substituting each column of G_j with $w_j G_j$, where w_j is chosen so that

$$w_j + \sum_{j}^{'} w_{j'} r_{jj'}^2 e^{-\lambda d_{jj'}} \tag{2.20}$$

is constant over j. The squared correlation coefficient between SNPs j and j' is denoted by $r_{jj'}^2$, the genomic distance is denoted by $d_{jj'}$, and λ is a constant. Note that $\alpha = -1$ corresponds with the GCTA REML estimator if all w_j are 1.

S2 Method of Moments Estimators: no-LD

In this section we derive basic moment properties of the random-SNP-effect Haseman-Elston (HE) estimator and the fixed-SNP-effects Dicker-1 estimator in the case of no LD. We see their differences, but also their similarity in practice. The more general case with LD is considered in Section 2.3 of the main paper.

S2.1 Haseman Elston Method of Moments Estimator:

The HE estimator is a second-order moments estimator based on a regression of products of phenotypes $y_i y_k$ for all pairs $i \neq k$ on the corresponding (i,k) terms of the $n \times n$ GRM matrix $\mathbf{\Psi} = M^{-1}\mathbf{\Gamma}_A\mathbf{\Gamma}'_A$. Given the standardized genotypes $\mathbf{\Gamma}$, the phenotypes depend only on the first m causal markers and $\mathbf{y} = \mathbf{\Gamma}_C\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the independent variables $\beta_j \sim N(0, \sigma_g^2/m)$, and $\epsilon_i \sim N(0, \sigma_e^2)$.

We first consider the estimator as a regression estimate conditional on Ψ. Noting $i \neq k$, so $\mathrm{E}(\epsilon_\mathrm{i}\ \epsilon_\mathrm{k}) = 0$ and that the β_j are independent, with mean 0 and variance σ_g^2/m,

$$\mathrm{E}(\mathrm{y_i y_k \Psi_{ik}}) = \mathrm{E}\left((\sum_{\mathrm{j}=1}^{\mathrm{m}} \Gamma_{\mathrm{ij}}\beta_\mathrm{j})\ (\sum_{\ell=1}^{\mathrm{m}} \Gamma_{\mathrm{k}\ell}\beta_\ell)\Psi_{\mathrm{ik}}\right) = \mathrm{E}\left(\sum_{\mathrm{j}=1}^{\mathrm{m}} \Gamma_{\mathrm{kj}}\Gamma_{\mathrm{ij}}\mathrm{E}(\beta_\mathrm{j}^2)\Psi_{\mathrm{ik}}\right) = \Psi_{\mathrm{ik}}^2\ \sigma_\mathrm{g}^2$$

Summing over all $n(n-1)/2$ pairs of distinct individuals, we have the method-of-moments equation

$$S_{Y\Psi} \equiv \sum_k \sum_{i<k} y_i y_k \Psi_{ik} = \sigma_g^2 \sum_k \sum_{i<k} \Psi_{ik}^2 \equiv \sigma_g^2 S_{\Psi\Psi}$$

so that σ_g^2 may be estimated as

$$\widetilde{\sigma_g^2} = \frac{S_{Y\Psi}}{S_{\Psi\Psi}} = \frac{\sum_k \sum_{i<k} y_i y_k \Psi_{ik}}{\sum_k \sum_{i<k} \Psi_{ik}^2} \tag{2.21}$$

Then an estimate of heritability is given by dividing by the empirical variance of $\mathbf{y}$.

Here we focus on the estimate of σ_g^2 and on the numerator and denominator denoted $S_{Y\Psi}$ and $S_{\Psi\Psi}$ respectively. We consider not only the conditional model, but also the variation in

Ψ over samples of genotypes from the population. Note that

$$\Psi_{ik} = M^{-1}\sum_{j=1}^{M}\Gamma_{ij}\Gamma_{kj} \quad \text{and} \quad \mathrm{E}(\Gamma_{\mathrm{ij}}) = 0, \;\; \mathrm{E}(\Gamma_{\mathrm{ij}}^2) = 1$$

So if individuals are independent, $\mathrm{E}(\Psi_{\mathrm{ik}}) = 0$, and if markers are independent,

$$\mathrm{E}(\Psi_{\mathrm{ik}}^2) = \mathrm{var}(\Psi_{\mathrm{ik}}) = \mathrm{M}^{-1}\mathrm{var}(\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{kj}}) \;=\; \mathrm{M}^{-1}(\mathrm{E}(\Gamma_{\mathrm{ij}}^2))^2 \;=\; 1/\mathrm{M}$$

and, under independence of individuals i, k and independence of markers j, w, ℓ,

$$\begin{aligned}
\mathrm{E}(\mathrm{y_i y_k}\Psi_{\mathrm{ik}}) &= M^{-1}\,\mathrm{E}\left((\sum_{\mathrm{j}=1}^{\mathrm{m}}\Gamma_{\mathrm{ij}}\beta_{\mathrm{j}} + \epsilon_{\mathrm{i}})(\sum_{\mathrm{w}=1}^{\mathrm{M}}\Gamma_{\mathrm{iw}}\Gamma_{\mathrm{kw}})(\sum_{\ell=1}^{\mathrm{m}}\Gamma_{\mathrm{k}\ell}\beta_{\ell} + \epsilon_{\mathrm{k}})\right) \\
&= M^{-1}\mathrm{E}\left(\sum_{\mathrm{j}=1}^{\mathrm{m}}\beta_{\mathrm{j}}^2(\sum_{\mathrm{w}=1}^{\mathrm{M}}\Gamma_{\mathrm{ij}}\Gamma_{\mathrm{iw}}\Gamma_{\mathrm{kw}}\Gamma_{\mathrm{kj}})\right) \\
&= M^{-1}\mathrm{E}\left(\sum_{\mathrm{j}=1}^{\mathrm{m}}\beta_{\mathrm{j}}^2\Gamma_{\mathrm{ij}}^2\Gamma_{\mathrm{kj}}^2\right) \;=\; \mathrm{M}^{-1}\mathrm{m}(\sigma_{\mathrm{g}}^2/\mathrm{m}) = \sigma_{\mathrm{g}}^2/\mathrm{M}
\end{aligned}$$

Hence $S_{\Psi\Psi}$ has expectation $n(n-1)/2M$ and $S_{Y\Psi}$ has expectation $\sigma_g^2 n(n-1)/2M$. Empirical simulations (not shown) showed that while the standard deviation of $S_{\Psi\Psi}$ is approximately n/M, that of $S_{Y\Psi}$ is of order $n/\sqrt{M}$, but both decrease to 0 as $M \to \infty$. Thus as $M \to \infty$ with n remaining fixed, both $S_{Y\Psi}$ and $S_{\Psi\Psi}$ converge in probability to 0. As the number of markers increases, the coefficient of variation of $S_{\Psi\Psi}$ remains constant, but that of $S_{Y\Psi}$ increases, and the empirical study shows the the standard deviation of the estimate of σ_g^2 to be of order $\sqrt{M}/n$. This result is in agreement with the theoretical equations for the estimator of Dicker [2014] in the case of no LD: see Lemma 2 and the Remarks following in that paper. That is, uncertainty in σ_g^2 and hence in h^2 increases as the number of markers M increases.

S2.2 The Dicker-1 fixed-SNP-effects model moments estimator

The Dicker-1 estimator [Dicker, 2014] is also a method of moments estimator, but starts from very different assumptions. The standardized genotypes Γ_{ij} are are assumed to be distributed

$N(0,1)$, independent over individuals i. The effects β_j are fixed effects, and in our case where only the first m markers are causal, $\beta_j \equiv 0$ for $j = (m+1), ...M$. The parameter to be estimated is $\sigma_g^2 \equiv \boldsymbol{\beta}'\boldsymbol{\Sigma}^*\boldsymbol{\beta}$ where here $\boldsymbol{\beta}$ is the m-vector of effects at causal markers augmented by $(M-m)$ zeros and $\boldsymbol{\Sigma}^*$ is the true LD matrix of correlations among all M markers. Because of the Normality assumption for genotypes, these can be rotated to orthonormality. This implies that the case of known $\boldsymbol{\Sigma}^*$ is mathematically equivalent to $\boldsymbol{\Sigma}^* = \boldsymbol{I}$. For simplicity we consider this case, then $\boldsymbol{\beta}'\boldsymbol{\Sigma}^*\boldsymbol{\beta} = \sum_{j=1}^m \beta_j^2$ and $m^{-1}\sum_{j=1}^m \beta_j^2 \equiv \sigma_g^2/m$, equivalent, for large m to the random-SNP-effects HE assumption $\beta_j \sim N(0, \sigma_g^2/m)$.

Dicker [2014] uses the quadratic forms $\|\mathbf{y}\|^2 = \mathbf{y}'\mathbf{y}$ and $\|\boldsymbol{\Gamma}'_A\mathbf{y}\|^2 = M\ \mathbf{y}'\boldsymbol{\Psi}\mathbf{y}$. Without making Normality assumptions, we can compute

$$\begin{aligned}
\mathrm{E}(\mathrm{M}\ \mathbf{y}'\boldsymbol{\Psi}\mathbf{y}) = &\sum_{i=1}^{n}\sum_{k=1}^{n}\mathrm{E}(\mathrm{M}\ \mathrm{y_i}\Psi_{\mathrm{ik}}\mathrm{y_k}) \\
= &\sum_{i=1}^{n}\sum_{k=1}^{n}\mathrm{E}\left((\sum_{\mathrm{j=1}}^{\mathrm{m}}\Gamma_{\mathrm{ij}}\beta_{\mathrm{j}}+\epsilon_{\mathrm{i}})(\sum_{\mathrm{w=1}}^{\mathrm{M}}\Gamma_{\mathrm{iw}}\Gamma_{\mathrm{kw}})(\sum_{\ell=1}^{\mathrm{m}}\Gamma_{\mathrm{k}\ell}\beta_\ell+\epsilon_{\mathrm{k}})\right) \qquad (2.22)
\end{aligned}$$

Under independence of Γ_{iw} and Γ_{kw} for $i \neq k$, the coefficient of σ_e^2 is seen to be Mn. Under independence of markers indexed by j, ℓ and w, the majority of terms in β_j and β_l in this expression disappear, leaving only a coefficient of $\sigma_g^2 = \sum_{j=1}^m \beta_j^2$. The remaining terms have $j = \ell \neq w$ (in which case $i = k$), or $j = \ell = w$ (in which case terms with both $i = k$ and $i \neq k$ remain). Grouping these two sets of terms this coefficient reduces to

$$(M-1)\ \mathrm{E}(\sum_{\mathrm{i}}\Gamma_{\mathrm{ij}}^2\Gamma_{\mathrm{iw}}^2)+\mathrm{E}(\sum_{\mathrm{i}}\sum_{\mathrm{k}}\Gamma_{\mathrm{ij}}^2\Gamma_{\mathrm{kj}}^2) \quad = \quad \mathrm{n(M-1)+Kn+n(n-1)} \ = \ \mathrm{n(M+n+K-2)}$$

where $K = \mathrm{E}(\Gamma_{\mathrm{ij}}^4)$. Combining the following two equations,

$$\begin{aligned}
\mathrm{E}(\mathrm{n}^{-1}\mathrm{M}\ \mathbf{y}'\boldsymbol{\Psi}\mathbf{y}) &= (M+n+K-2)\sigma_g^2 \ + \ M\sigma_e^2 \\
\mathrm{E}(\mathrm{n}^{-1}\mathbf{y}'\mathbf{y}) &= \sigma_g^2+\sigma_e^2
\end{aligned}$$

and assuming $K = 3$ we obtain the Dicker [2014] method-of-moments estimator of σ_g^2:

$$\tilde{\sigma}_g^2 \ = \ (n(n+1))^{-1}(M\ \mathbf{y}'\boldsymbol{\Psi}\mathbf{y} \ - \ M\mathbf{y}'\mathbf{y}) \ = \ (n(n+1))^{-1}(\|\boldsymbol{\Gamma}'_A\mathbf{y}\|^2 \ - \ M\|\mathbf{y}\|^2) \qquad (2.23)$$

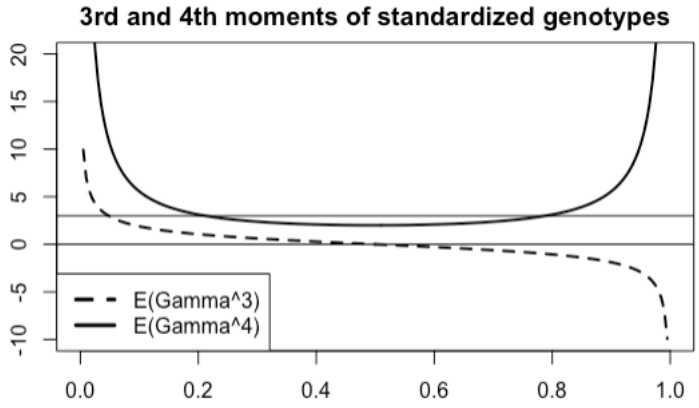

Figure 2.7: Skewness and kurtosis of the normalized genotypes as a function of allele frequency

Note that whereas the numerator and denominator of the HE estimator (2.6) always has the correct expectations, Equation (2.23) is only exact if $K = 3$. Since K appears only in the term $(M + n + K - 2)$ the impact will be small for large M and/or n, but it is worth noting that K can be quite large (> 100) for loci with rare alleles (see Figure 2.7). Under the $N(0, 1)$ assumption, Dicker [2014] gives also many other expressions for high-order moments of these estimators. However, these depend more critically on the higher-order moments of the Γ_{ij}, and hence his Normality assumption.

Although the assumptions underlying the MoM estimator (2.23) are very different from those of the HE estimator of Equation (2.6), operationally and in performance the estimators are quite similar, in the case of known or no LD. The key difference from the HE estimator is then that whereas the latter considers only Ψ_{ik} for $i \neq k$, the Dicker estimator uses the full $n \times n$ matrix $M\mathbf{\Psi} = \mathbf{\Gamma}_A\mathbf{\Gamma}'_A$. This use of the diagonal terms Ψ_{ii} permits an estimators of σ_g^2 and σ_e^2 that is linear in the relevant quadratic forms, rather than the ratio S_{YT}/S_{TT}, but strict correctness and moment properties are dependent on the Normality assumption for Γ_{ij}.

S3 Moment based estimators: LD case

S3.1 Biases in HE estimator in the presence of LD

In the presence of LD, the HE estimator may be biased. A formula for this bias, approximating the expectation of a ratio by the ratio of expectations, is derived in Section 2.3 of the main paper. We here include some further analyses of the theoretical predictions of Equation2.13, that LD changes the expectation of the HE estimator by a factor of $\frac{M}{m}\frac{R_{CC}+R_{CF}}{R_{CC}+2R_{CF}+R_{FF}}$.

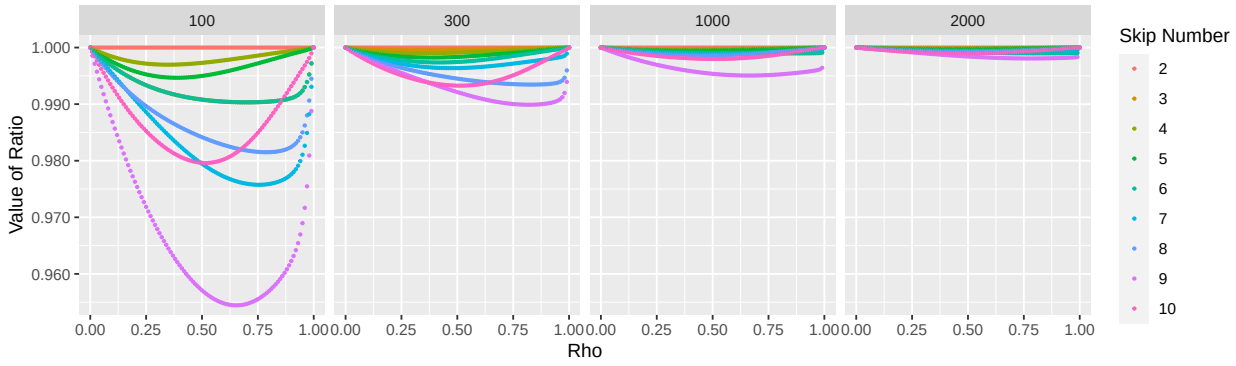

Figure 2.8: For the autocorrelation structure, values of the factor of Equation (Y-axis) 2.13 are plotted for different values of M (different panels), ρ (X-axis), and skip number (colors)

In Supplementary Figure 2.8, we calculated approximate theoretical biases of HE estimator in autocorrelated data for different values of ρ (x-axis), number of markers total markers M (different panels), and different "skip" numbers using equation 2.4.3. The skip number is the number of elements until a causal marker is seen. For example, if the skip number for is 2, then every second marker is causal, and all others are noncausal. The value of the ratio reported (Y-axis) indicates that the estimator is unbiased when the ratio is 1. We observed that as predicted in Section 2.4.3, no bias is observed when the skip number is 2, and furthermore, the bias is close to 1 whenever M is large for all skip numbers up to 10.

For the block structure, we can analytically show that for any number of blocks and any value of ρ, there is no bias resulting from Equation 2.13. We begin by computing for the

case that there is 1 block consisting of all M markers, with m of the markers being causal. The correlation between all of the markers in the block is ρ. Without loss of generality, we can assume that all of the causal markers are listed before the noncausal markers. When this is the case, we can calculate that $R_{CC} = m + m(m-1)\rho$, $R_{CF} = m(M-m)\rho$,and $R_{FF} = M - m + (M-m)(M-m-1)\rho$. Substituting these values, we reach

$$E(\tilde{\sigma}_g^2) \approx \sigma_g^2 \frac{M}{m} \frac{m + m(m-1)\rho + m(M-m)\rho}{m + m(m-1)\rho + 2m(M-m)\rho + M - m + (M-m)(M-m-1)\rho}$$

and upon simplifying this expression, we find that $E(\tilde{\sigma}_g^2) \approx \sigma_g^2$. To extend this to the case of multiple identical blocks, we note that each of R_{CC}, R_{CF}, and R_{FF} are multiplied by the number of blocks, and hence the bias is the same

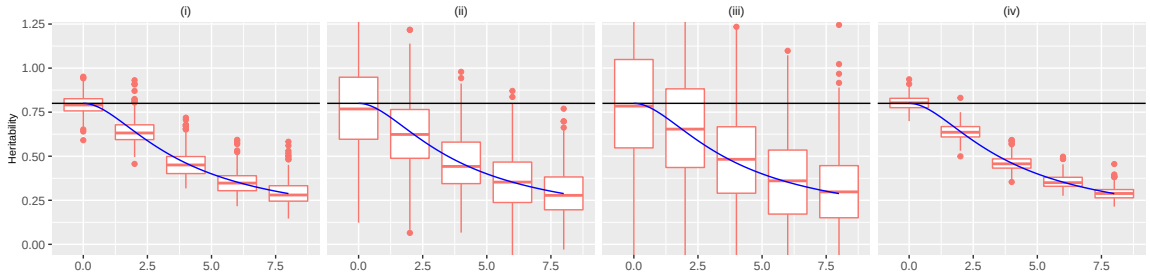

Figure 2.9: Estimates of h^2 (y-axis) for different values of r, the number of times that 10% of the markers are being repeated. Estimates are made using the HE estimator (red box plots) with data simulated from the repeat structure of simulation study 1. The true simulated heritability was 0.8 (solid black line). The solid blue line plots the theoretical estimates based on Equation (2.14). The set up is the same as in Figure 2.3, with (i) $n = 1000, m = 200$ (ii) $n = 200, m = 1000$, (iii) $n = 200, m = 3000$ (iv) $n = 2000, m = 1000$.

Of the simulation study examples of this paper, the bias is marked in the case of non-causal markers that repeat the genotypes of causal markers. Figure 2.9 aims to validate the formula for the bias and assess the variation in bias across realizations by taking estimates of heritability from the repeat structure simulated data in simulation study 1 and comparing

against the theoretical values from Equation (2.14). It is shown that there is close alignment of the theoretical values and the observed. We also note that the theoretical bias does not depend on the value of m, since if we multiply m by a constant c, then if we have set a fixed percentage of markers to be repeated, d is also multiplied by our constant c, and

$$\sigma_g^2 \frac{(cm + rcd)^2}{cm(cm + rcd(2 + r))} = \sigma_g^2 \frac{(m + rd)^2}{m(m + rd(2 + r))}$$

S3.2 Impact of LD on the Dicker-1 Estimator

We consider now the estimator of Equation (2.7) in the presence of LD. Although this estimator would likely not be used in practice because it does not attempt to adjust for LD and hence has a different estimand than σ_g^2 as defined here, it provides important motivation for the Dicker-2 estimator (Equation 2.9). We here provide justification for poor performance of the Dicker-1 estimator in simulation. Even in the absence of LD, this estimator of σ_g^2 is unbiased only if $\mathrm{E}(\Gamma_{\mathrm{ij}}^4) = 3$ (Supplementary Section S2.2) but the bias is negligible for large M or n.

Recall that for the Dicker-1 estimator (Equation 2.7), $\tilde{\sigma}_g^2 = (n(n+1))^{-1}(M\mathbf{y}'\mathbf{\Psi}\mathbf{y} - M\mathbf{y}'\mathbf{y})$. We begin by analyzing the term $\mathbf{y}'\mathbf{y}$. Note that $E(\Gamma_{ij}\Gamma_{il}) = \Sigma_{jl}^*$ so that

$$\mathrm{E}(\mathrm{n}^{-1}\mathbf{y}'\mathbf{y} \mid \boldsymbol{\beta}) = n^{-1}\mathrm{E}\left((\mathbf{\Gamma}_{\mathrm{C}}\boldsymbol{\beta} + \boldsymbol{\epsilon})'(\mathbf{\Gamma}_{\mathrm{C}}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \mid \boldsymbol{\beta}\right) \quad = \quad \boldsymbol{\beta}'\mathbf{\Sigma}^*\boldsymbol{\beta} + \sigma_{\mathrm{e}}^2$$

where the vector $\boldsymbol{\beta}$ contains only entries for causal markers.

The fixed-SNP-effects Dicker-1 estimator estimates $\tau^2 = \boldsymbol{\beta}'\mathbf{\Sigma}^*\boldsymbol{\beta}$ [Dicker, 2014], which may differ from the additive genetic variance $\sigma_g^2 \equiv \sum_{j=1}^m \beta_j^2/m$ in the presence of LD. However, in our simulation studies, each replicate uses β_j at causal loci $j = 1, ..., m$ that are independently generated $N(0, \sigma_g^2/m)$ and independent of the standardized genotypes Γ_{ij} (Section 2.4.5).

Thus, over replicate simulations

$$\begin{aligned}\mathrm{E}(\boldsymbol{\beta}'\boldsymbol{\Sigma}^*\boldsymbol{\beta}) &= E(\sum_{\ell=1}^{m}\sum_{j=1}^{m}\beta_\ell\Sigma^*_{\ell j}\beta_j)\\ &= (\sigma_g^2/m)\sum_{j=1}^{m}\Sigma^*_{jj}\\ &= \sigma_g^2\end{aligned}$$

and hence $\mathrm{E}(\mathrm{n}^{-1}\mathbf{y}'\mathbf{y}) = \sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{e}}$

We now analyze the $\mathbf{y}'\boldsymbol{\Psi}y$ term. Like with the $\mathbf{y}'\mathbf{y}$ term, we may consider expectations of the estimator over replicates assuming that β_j $(j = 1, ..., m)$ are independent $N(0, \sigma_g^2/m)$. First, we note that $\mathbf{y}'\boldsymbol{\Psi}\mathbf{y} = 2S_{Y\Psi} + \sum_i y_i^2\Psi_{ii}$ so that, from Equation (2.12)

$$\begin{aligned}\mathrm{E}(\mathrm{M}\ \mathbf{y}'\boldsymbol{\Psi}\mathbf{y}) =\ & \mathrm{E}(2\mathrm{MS}_{\mathrm{Y}\Psi} + \mathrm{M}\sum_{\mathrm{i}}\mathrm{y}_{\mathrm{i}}^2\Psi_{\mathrm{ii}})\\ &= n(n-1)m^{-1}\sigma_g^2\ (R_{CC} + R_{CF}) + M\mathrm{E}\left[\sum_{\mathrm{i}}(\sum_{\mathrm{j}}\Gamma_{\mathrm{ij}}\beta_{\mathrm{j}} + \epsilon_{\mathrm{i}})^2(\sum_{\mathrm{w}}\Gamma^2_{\mathrm{iw}})\right]\\ &= n(n-1)m^{-1}\sigma_g^2\ (R_{CC} + R_{CF})\ +\ nm^{-1}\sigma_g^2\sum_{j=1}^{m}\sum_{w=1}^{M}\mathrm{E}(\Gamma^2_{\mathrm{ij}}\Gamma^2_{\mathrm{iw}})\ +\ \mathrm{Mn}\ \sigma^2_{\mathrm{e}}\end{aligned}$$

In general $\mathrm{E}(\Gamma^2_{\mathrm{ij}}\Gamma^2_{\mathrm{iw}})$ is unknown, but a lower bound on the double-sum term is the no-LD value $mK + m(M-1)$ (see Supplementary Section S2.2) while a rough approximation might be $mK + (M-1)(R_{CC} + R_{CF})$, where again $K = \mathrm{E}(\Gamma^4_{\mathrm{ij}})$. This approximation gives the overall result

$$\mathrm{E}(\mathrm{n}^{-1}\mathrm{M}\mathbf{y}'\boldsymbol{\Psi}\mathbf{y}) \approx\ \sigma_g^2\ (K\ +\ \frac{(n+M-2)}{m}(R_{CC} + R_{CF}))\ +\ M\sigma_e^2$$

Combining the $M\mathbf{y}'\boldsymbol{\Psi}\mathbf{y}$ and $M\mathbf{y}'\mathbf{y}$ terms, we have

$$\mathrm{E}\left((\mathrm{n}(\mathrm{n}+1))^{-1}(\mathrm{M}\mathbf{y}'\boldsymbol{\Psi}\mathbf{y}\ -\ \mathrm{M}\mathbf{y}'\mathbf{y})\right) \approx \sigma^2_{\mathrm{g}}\ (\mathrm{n}+1)^{-1}(\mathrm{K}\ +\ \frac{(\mathrm{n}+\mathrm{M}-2)}{\mathrm{m}}(\mathrm{R_{CC}} + \mathrm{R_{CF}}) - \mathrm{M}).$$

Since the squared correlations $r^2_{j\ell}$ are non-negative, $(R_{CC} + R_{CF}) \geq m$ and the estimator will overestimate σ_g^2 and hence also heritability h^2. Unlike the HE estimator where the LD inflates both numerator and denominator (Equation 2.13), the form of the estimator (2.7) means that it can only be inflated by LD.

S3.3 Moment estimators designed to accommodate LD

In the case when LD must be estimated from the sample data, Dicker [2014] and Schwartzman et al. [2019] developed moment-based estimators of σ_g^2, σ_e^2, and h^2 under the fixed-SNP-effects framework.

Here we consider the estimator of Dicker [2014] in the case of LD. Again, the GRM $\Psi = M^{-1}\mathbf{\Gamma}_A\ \mathbf{\Gamma}'_A$, and LD matrix $\mathbf{\Sigma} = n^{-1}\mathbf{\Gamma}'_A\mathbf{\Gamma}_A$. If the standardized genotypes, Γ_{ij}, are marginally $N(0,1)$ and independent over i, and if $\mathbf{\Sigma}^*$ is the true positive definite correlation matrix of the Γ_{ij} over j, then $\mathbf{\Sigma}^{*^{-1/2}}\mathbf{\Gamma}'_A$ are independent $N(0,1)$ and the estimator (2.23) becomes

$$\begin{aligned}\tilde{\sigma}_g^2 &= (n(n+1))^{-1}((\mathbf{\Sigma}^{-1/2}\mathbf{\Gamma}'_A\mathbf{y})'(\mathbf{\Sigma}^{-1/2}\mathbf{\Gamma}'_A\mathbf{y}) - M\mathbf{y}'\mathbf{y}) \\ &= (n(n+1))^{-1}(\mathbf{y}'\mathbf{\Gamma}_A\mathbf{\Sigma}^{-1}\mathbf{\Gamma}'_A\mathbf{y} - M\mathbf{y}'\mathbf{y}) \qquad (2.24)\end{aligned}$$

and again $\sigma_g^2 + \sigma_e^2$ is estimated by the phenotypic variance $n^{-1}\mathbf{y}'\mathbf{y}$. More generally, as shown by Dicker [2014], if $n > M$ and $\mathbf{\Sigma}$ is a norm-consistent estimator of the true correlation matrix the properties and results of the non-LD estimator (2.23) apply also in the LD case to the estimator (2.24).

However, in most applications, M is much larger than n. and the estimator (2.24) breaks down, and as shown in Dicker [2014], In this case they propose to use lower-order moments of the trace of $\mathbf{\Sigma} = n^{-1}\mathbf{\Gamma}'_A\mathbf{\Gamma}_A$. Specifically they define

$$\mu_1 = \frac{tr(\mathbf{\Sigma})}{M} \text{ and } \mu_2 = \frac{tr(\mathbf{\Sigma}^2)}{M} - \frac{(tr(\mathbf{\Sigma}))^2}{Mn} \qquad (2.25)$$

The estimator of σ_g^2 becomes

$$\tilde{\sigma}_g^2 = \frac{\mu_1(\mathbf{\Gamma}'_A\mathbf{y})'(\mathbf{\Gamma}'_A\mathbf{y}) - M\mu_1^2\mathbf{y}'\mathbf{y}}{n(n+1)\mu_2} \qquad (2.26)$$

and again $\sigma_g^2 + \sigma_e^2$ is estimated by $n^{-1}\mathbf{y}'\mathbf{y}$. For more on the theory and properties of the estimator (2.26) see Dicker [2014]. For the current paper, we implement this estimator as "Dicker-2" in our simulations and results.

Schwartzman et al. [2019] proposed a method of moments estimator based on that of Dicker [2014]. They derive a form that depends only on summary statistics instead of the raw genotypic and phenotypic data and hence their estimator has wider applicability. However, in the basic form (not using only summary statistics) their estimator is essentially equivalent to the estimator (2.26), so we do not consider it further in this paper.

S4 Simulation of Genetic Marker LD Structures

Autocorrelated: we assume that for each individual, M markers are generated from a multivariate Gaussian with $AR1(\rho)$ covariance matrix. We generate the markers for each individual independently. In other words, we assume that for individual i, genotypes $\tilde{G}_i$ are generated from $\tilde{G}_i \sim N(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{M-1} \\ \rho & 1 & \rho & \dots & \rho^{M-2} \\ \vdots & & \vdots & & \vdots \\ \rho^{M-1} & \rho^{M-2} & \rho^{M-3} & \dots & 1 \end{pmatrix}$$

The continuous values $\tilde{G}_i$ are then converted to discrete genotypes G_i taking value 0, 1 or 2. For a marker with alternate allele frequency f, $G_{ij} = 0$, 1, or 2, depending on if $\tilde{G}_{ij}$ is less than $\Phi^{-1}(f^2)$, between $\Phi^{-1}(f^2)$ and $\Phi^{-1}(f^2 + 2f(1-f)) = \Phi^{-1}(2f - f^2)$, or greater than $\Phi^{-1}(2f - f^2)$, where $\Phi(\cdot)$ is the $N(0,1)$ distribution function. Note that this trichotomy gives the correct marginal genotype probabilities, but reduces the genotypic correlation (LD) between markers below that used in the simulation matrix Σ: compare panels A with D, or B with E in Figure 2.10.

Block: we generate block genotypes according to the same mechanism as the autocorrelated genotypes, except we choose that

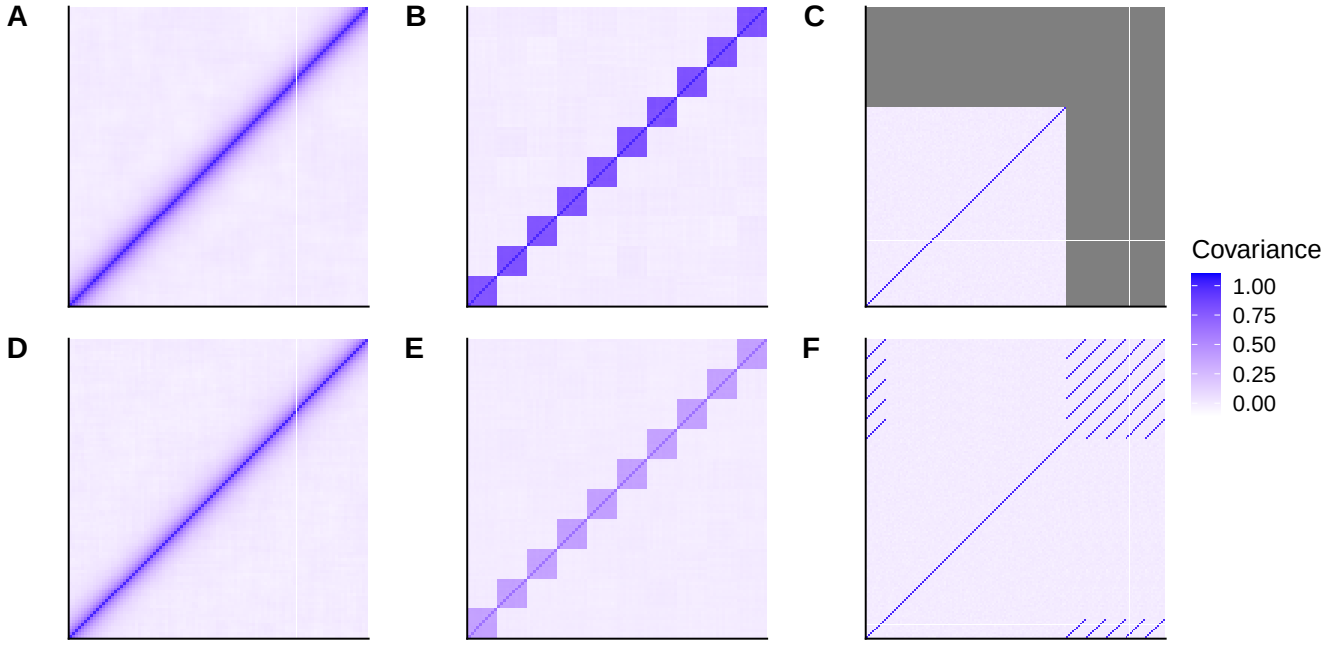

Figure 2.10: These panels plot the empirical covariance matrices for simulated genotypes from 10,000 individuals and $p = 100$ markers. The correlation between markers decreases after discretization but the pattern generally remains the same. (A) Autocorrelated markers were generated from the Gaussian model, i.e. plotting $Cov(\tilde{G})$ (B) Blocked markers were generated from the Gaussian model. (C) Independent markers were generated. (D) Autocorrelated markers were generated and then discretized and normalized, i.e. this is $Cov(\Gamma)$ (E) Blocked markers were discretized and normalized. (F) Repeated markers were generated with 10 markers being repeated 5 times.

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & & \vdots & & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix}$$

for each block. We assume that there are 10 blocks, each with $M/10$ markers.

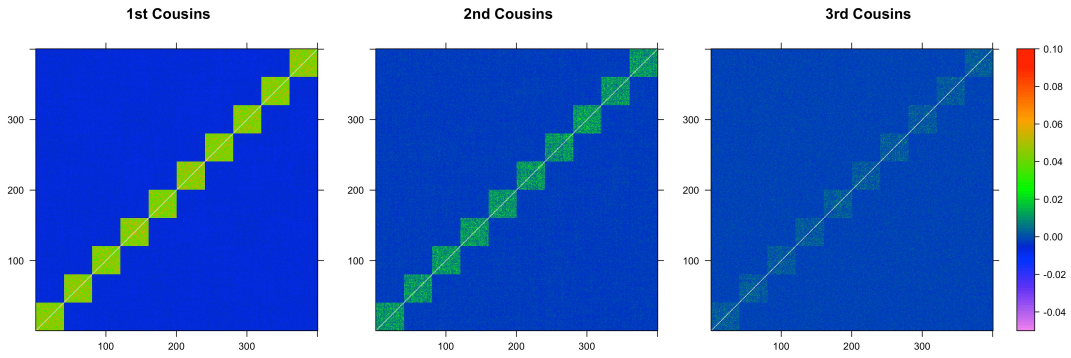

Figure 2.11: Colors represent values of the log of 1 plus the average of 100 GRMs generated from 400 individuals. The i, jth entry of the matrix corresponds to the relatedness between ith individual and the jth individual. Sets of cousins are adjacent in groups of 40. Colors are thresholded at 0.1, and set to white if it is above the threshold.

Repeat: In this case m marker genotypes are independently generated from the binomial distribution. That is, for a marker with alternate allele frequency f, $G_{ij} \sim Binomial(2, f)$. We designate a proportion of markers to be repeated. We repeat these markers r times.

Choice of causal markers

For the three simulation LD structures, we selected $\boldsymbol{G}_c$ to be a subset of $\boldsymbol{G}$. For the autocorrelation and block simulated genotypes, we chose alternating markers to be causal and non-causal markers. For the repeat structure the original m markers were chosen to be causal, while the repeat genotypes were non-causal. The genotypes were standardized to each have mean 0 and variance 1, using the empirical allele frequencies in the simulated sample of n individuals. The matrix $\boldsymbol{\Gamma}_A$ of standardized genotypes was formed as given in Equation (2.1), while $\boldsymbol{\Gamma}_C$ is the corresponding matrix for the m causal markers.

S5 Equivalence of a simplified h^2_{GRE} and Dicker-1-Σ

Recall that from Section 2.4.2, the Dicker-1 estimator can be expressed as $(n(n+1))^{-1}(\|\boldsymbol{\Gamma_A}'y\|^2 - M\mathbf{y}'\mathbf{y})$ if $\boldsymbol{\Sigma}^*$ is known to be the identity matrix. In the case that $\boldsymbol{\Sigma}^*$ is known or estimable but not

the identity, we have Dicker-1-Σ. We replace $\mathbf{\Gamma}$ by $\mathbf{\Gamma\Sigma}^{-1/2}$, and we have

$$\frac{M\left[(\mathbf{\Gamma\Sigma}^{-1/2})'\mathbf{y}\right]'\left[(\mathbf{\Gamma\Sigma}^{-1/2})'\mathbf{y}\right] - M\mathbf{y}'\mathbf{y}}{n(n+1)} = \frac{M\mathbf{y}'\mathbf{\Gamma\Sigma}^{-1/2}(\mathbf{\Sigma}^{-1/2})'\mathbf{\Gamma}'\mathbf{y} - M\mathbf{y}'\mathbf{y}}{n(n+1)} \tag{2.27}$$

$$= \frac{M\mathbf{y}'\mathbf{\Gamma\Sigma}^{-1}\mathbf{\Gamma}'\mathbf{y} - M\mathbf{y}'\mathbf{y}}{n(n+1)} \tag{2.28}$$

On the other hand, if we do not apply partitioning, the h^2_{GRE} estimator is expressed as

$$h^2_{GRE} = \frac{n\hat{\boldsymbol{\beta}}'\mathbf{\Sigma}^{-1}\hat{\boldsymbol{\beta}} - q}{n-q} \tag{2.29}$$

$$\approx \frac{\mathbf{y}'\mathbf{\Gamma\Sigma}^{-1}\mathbf{\Gamma}'\mathbf{y} - \mathbf{y}'\mathbf{y}q}{n(n-q)} \tag{2.30}$$

$$\approx \frac{\mathbf{y}'\mathbf{\Gamma\Sigma}^{-1}\mathbf{\Gamma}'\mathbf{y} - \mathbf{y}'\mathbf{y}q}{n(n+1)} \tag{2.31}$$

Here, $\hat{\boldsymbol{\beta}}$ is defined to be $\frac{1}{n}\mathbf{\Gamma}'\mathbf{y}$, as per [Hou et al., 2019]. Furthermore, q is the rank of $\mathbf{\Sigma}$. If $n > M$ and Γ is full rank, then $q = M$. We assume that $\mathbf{y}'\mathbf{y} \approx n$. Furthermore, for $n >> M$, we assume $n(n-1) \approx n(n-M)$. With these assumptions, Equation (S13) and Equation (S10) are the same, demonstrating the equivalence. Upon rescaling the Dicker-1-Σ estimator by $\frac{n-1}{n-M}$, the Dicker-1-Σ and the h^2_{GRE} estimator are essentially equivalent (Supplementary Figure 2.12)

We simulated data similarly to Section 2.4.5, but excluded cases where $n < M$ because if $n < M$ and $\mathbf{\Gamma}$ is rank n, then $\mathbf{\Sigma}$ can often have rank n, in which case the GRE estimator is not well defined. We excluded the repeat LD structure because we were unable to calculate the Dicker-1-Σ estimator since we were unable to calculate Σ^{-1}. We found that the h^2_{GRE} estimator was robust to the structures of LD that we presented here. Furthermore, even though when $r = 0$ or $\rho = 0$, we have that $\Sigma^* = I$, the h^2_{GRE} could have lower MSE than Dicker-1. This may be because including the empirical Σ may reduce the variance of the estimate.

S6 Equivalence of RHE-mc with one component to HE

The randomized Haseman Elston estimator with multiple components (RHE-mc) uses a system of normal equations to estimate σ^2_g and σ^2_e (Equation 7 in [Pazokitoroudi et al.,

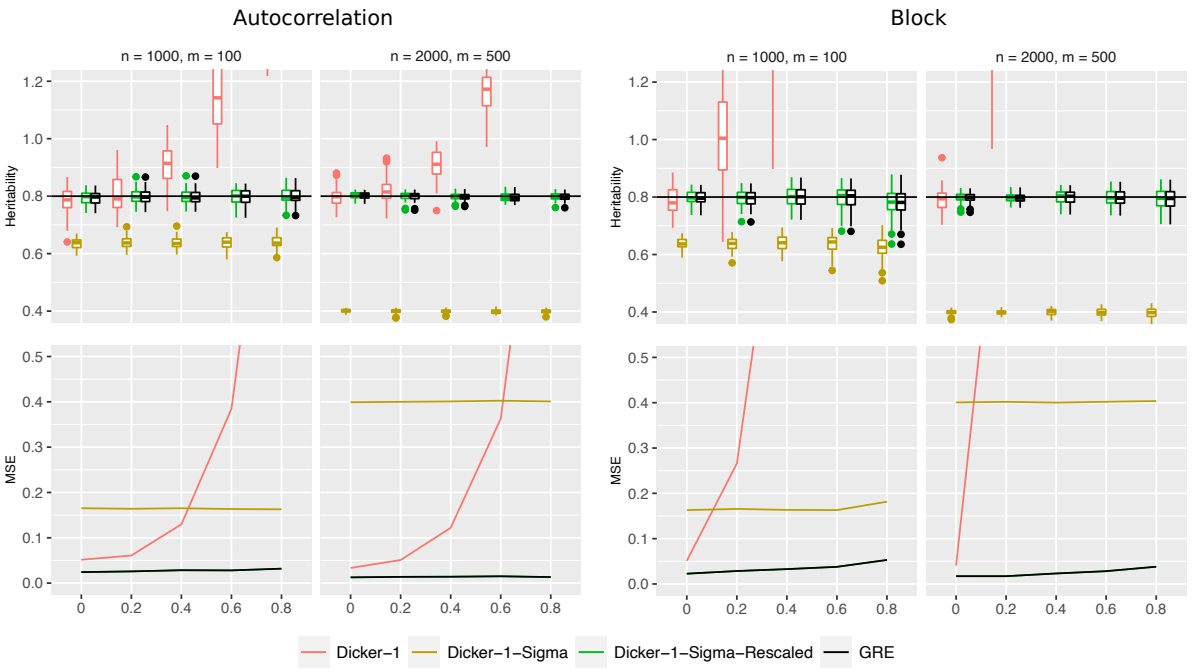

Figure 2.12: We simulated 50 data sets for each of autocorrelation, block, and repeat structures of each of the estimators, and including the h^2_{GRE} estimator (black). The X-axis plots ρ. A horizontal line is shown at $h^2 = .8$. On the top row, estimates of heritability are shown. On the bottom row, MSEs are shown.

2020]). If only one component is used in this estimator, then the equations become

$$\begin{pmatrix} tr(\mathbf{\Psi\Psi}) & n \\ n & n \end{pmatrix} \begin{pmatrix} \tilde{\sigma}_g^2 \\ \tilde{\sigma}_e^2 \end{pmatrix} = \begin{pmatrix} \mathbf{y'\Psi y} \\ \mathbf{y'y} \end{pmatrix} \tag{2.32}$$

Upon solving the system of equations for $\tilde{\sigma}_g^2$, we obtain the estimator

$$\tilde{\sigma}_g^2 = \frac{\mathbf{y'\Psi y} - \mathbf{y'y}}{tr(\mathbf{\Psi'\Psi}) - n} \tag{2.33}$$

Note that we used the fact that $\mathbf{\Psi}$ is symmetric, and hence $\mathbf{\Psi} = \mathbf{\Psi}'$. Because $\mathbf{y}$ is standardized, we have $\mathbf{y'y} = 1$. Then we note that Equation 2.33 is the same as Equation 2.15.

Chapter 3

MATRIX PRIOR FOR DATA TRANSFER BETWEEN SINGLE CELL DATA TYPES IN LATENT DIRICHLET ALLOCATION

This chapter is adapted with minimal modification from:

Alan Min, Timothy Durham, Louis Gevirtzman, and William Stafford Noble. Matrix prior for data transfer between single cell data types in latent Dirichlet allocation. *PLOS Computational Biology*, 19(5):e1011049, 2023

3.1 Author Contributions

TD and WSN conceptualized the idea of transferring data in LDA using a prior. AM proposed the use of the matrix prior approach to solve the data transfer problem, conducted simulation studies, and conducted validation studies of the matrix prior. AM and TD curated the data used for analysis, and determined validation approaches for the matrix prior. LG implemented the matrix prior approach in a Java distribution. TD and WSN provided critical feedback throughout all of the validation approaches. All authors contributed to reading the final manuscript.

3.2 Abstract

Single cell ATAC-seq (scATAC-seq) enables the mapping of regulatory elements in fine-grained cell types. Despite this advance, analysis of the resulting data is challenging, and large scale scATAC-seq data are difficult to obtain and expensive to generate. This motivates a method to leverage information from previously generated large scale scATAC-seq or scRNA-seq data to guide our analysis of new scATAC-seq datasets. We analyze scATAC-seq data using latent Dirichlet allocation (LDA), a Bayesian algorithm that was developed

to model text corpora, summarizing documents as mixtures of topics defined based on the words that distinguish the documents. When applied to scATAC-seq, LDA treats cells as documents and their accessible sites as words, identifying "topics" based on the cell type-specific accessible sites in those cells. Previous work used uniform symmetric priors in LDA, where the parameters of the Dirichlet prior were set to be equal, but we hypothesized that nonuniform matrix priors generated from LDA models trained on existing data sets may enable improved detection of cell types in new data sets, especially if they have relatively few cells. In this work, we test this hypothesis in scATAC-seq data from whole *C. elegans* nematodes and SHARE-seq data from mouse skin cells. We show that nonsymmetric matrix priors for LDA improve our ability to capture cell type information from small scATAC-seq datasets.

3.3 Introduction

Single cell genomics has emerged as a powerful method to characterize gene expression (scRNA-seq) and chromatin accessibility (scATAC-seq). The resulting data enables fine-grained identification of cell types. For example, in *Caenorhabditis elegans*, scRNA-seq and scATAC-seq have been used to measure genome-wide gene expression levels and chromatin accessibility for the majority of individual cells in the developing embryo and second-stage larval (L2) worms [Durham et al., 2021, Cao et al., 2017, Packer et al., 2019].

Several research groups have found that a Bayesian modeling approach called latent Dirichlet allocation (LDA) is an effective method for distinguishing different cell types in scRNA-seq and scATAC-seq data [González-Blas et al., 2019, Dey et al., 2017]. LDA was developed to model topics in text corpora using counts of words in each document, but when applied to scATAC-seq data, can be used to condense peaks into topics that describe cell types within the data. When applied to scATAC-seq data, the outputs of LDA are a cell-topic matrix, describing the topics assigned to each cell, and a topic-peak matrix, describing how strongly a peak contributes to the definition of each topic. LDA is also well-suited to model single cell genomics data because it expects a matrix of integers as input, and thus

can naturally operate on the raw count matrices generated by scATAC-seq or scRNA-seq.

Despite promising results, the challenges posed by scATAC-seq data motivated us to incorporate auxiliary data into the LDA algorithm. Single cell data is still expensive to gather, but there are large compendia of single cell ATAC-seq data available. We aim to use large reference sets of scATAC-seq data ("atlases") to improve the analysis of smaller datasets through the use of LDA with a nonuniform matrix prior. Specifically, we propose to use previously generated data to create a probabilistic prior for use by the scATAC-seq LDA model. We further investigate the possibility of using scRNA-seq data to transfer information to smaller scATAC-seq datasets. In general, the Bayesian prior methodology provides a principled and computationally lightweight way to incorporate auxiliary data.

We verified the utility of our approach via simulation and then applied the technique to a dataset from *C. elegans* produced using the sci-ATAC-seq assay [Durham et al., 2021] and a dataset from mouse skin cells produced using the SHARE-seq assay [Ma et al., 2020]. We first used simulated data to verify the feasibility of transferring information between two datasets with the same, known, underlying topics; and we show that the nonuniform matrix prior can increase the ability of LDA to identify true underlying topic structure within a given dataset. Next, for the *C. elegans* and mouse skin scATAC-seq data, we split each full data set into a larger "reference" subset and a smaller "target" subset, then applied LDA with a uniform symmetric prior to the reference subset and used the results of that LDA as a nonuniform matrix prior for an LDA model of the target subset. We report that in the mouse skin data, agreement with previously called cell types improved by using auxiliary scATAC-seq data, and that correlation of the output matrices from the "target" LDA with the output matrices from the LDA on the full data set is higher with the nonuniform matrix prior than with the uniform symmetric prior. For the *C. elegans* data, we also found increased correlation of the output matrices between the full data set LDA and the target LDA when using the matrix prior; however, unlike with the SHARE-seq data, we saw no improvement in the agreement with previously called cell types. Finally, we leveraged the paired nature of the scATAC-seq and scRNA-seq data in the mouse skin SHARE-seq dataset to attempt

to transfer information across single cell assays. We used the output from an LDA on the scRNA-seq data as a matrix prior for an LDA on the scATAC-seq data. In this case, we did not see a clear improvement when using the matrix prior, but the cross-assay matrix prior might still be improved with further hyperparameter tuning.

3.4 Approach

3.4.1 Background: latent Dirichlet allocation

LDA was originally developed to model text documents as a mixture of topics. For a fixed vocabulary, a document is described as the number of occurrences of each word in the vocabulary. In our case, instead of words, we are modeling scATAC-seq peaks or scRNA-seq genes. We assume that each cell is generated from a mixture of topics, and that each topic has a distribution of peaks or genes from the vocabulary associated with it. The parameters we hope to estimate are the distribution of topics for each cell and the distribution of peaks or genes for each topic, called the "cell-topic matrix" and the "topic-peak matrix" or "topic-gene matrix," respectively. A generative model is assumed for each of the cells, described below, following [Blei et al., 2003].

Let N be the number of sequencing reads or unique molecular identifiers (UMIs) in a cell. Let T be the number of topics. Let V be the number of peaks or genes in the vocabulary. Let U be the number of cells. Let the vector $\mathbf{w} = (w_1, \ldots, w_N)$ be a cell, where each w_i is the number of reads or UMIs associated with a particular peak or gene from the vocabulary. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)$ be the topic distribution of a given cell, where $\sum_{i=1}^{T} \theta_i = 1$. Let $\boldsymbol{\phi}_t = (\phi_{t1}, \ldots, \phi_{tV})$ be the peak or gene distribution for topic t such that $\sum_{w=1}^{V} \phi_{tw} = 1$, where ϕ_{tw} indicates the probability of observing a read or UMI from the peak or gene w given that its topic was t. Let $\boldsymbol{\alpha}$ be the "basis vector" of length T such that $\sum_i \alpha_i = 1$ and $\alpha_i > 0$ for each i, which parameterizes the Dirichlet prior distribution over the cell-topic matrix. Similarly, let $\boldsymbol{\beta}$ be a basis vector of length V such that $\sum_j \beta_j = 1$ and $\beta_j > 0$ for each j, which parameterizes the Dirichlet distribution over the topic-peak or topic-gene matrix.

Let c_α be the "concentration parameter" for $\boldsymbol{\alpha}$, and c_β be the concentration parameter for $\boldsymbol{\beta}$. Let ξ be the average number of peaks or genes in a cell.

We note that $\boldsymbol{\alpha}$ is a prior that describes the frequency at which topics are observed, and $\boldsymbol{\beta}$ is a prior that describes the frequency at which peaks or genes are observed, i.e. a large value of α_i or β_i indicates that the ith topic or ith peak or gene is more common, respectively.

Blei et al. [2003] first describe a variational Bayes approach to optimizing the posterior distribution of the topic assignments for each peak or gene. Following González-Blas et al. [2019] and Darling [2011], however, we use a Gibbs sampling algorithm for LDA for scATAC-seq data (Algorithm 1) .

In González-Blas et al. [2019], the last Gibbs sampling iteration is chosen to be the output of the topic-peak and cell-topic matrices. We modify this approach slightly by calculating the posterior likelihood of each of the iterations and choosing the iteration with the highest likelihood. This is the maximum a posteriori (MAP) estimate.

3.4.2 Data transfer using a matrix prior

To achieve our goal of leveraging large scale data to improve inference in small datasets, we extend Algorithm 1 to accept a matrix prior $\boldsymbol{B}$ to replace the vector prior $\boldsymbol{\beta}$ for the topic-peak or topic-gene distribution. Unlike the vector prior $\boldsymbol{\beta}$, which specifies the same prior distribution over vocabulary elements regardless of topic, the matrix prior $\boldsymbol{B}$ can specify a different prior distribution for each topic. The matrix prior $\boldsymbol{B}$ has elements B_{tw} such that for each t, $\sum_w B_{tw} = 1$, corresponding to the probability of observing peak or gene w in topic t. It has corresponding concentration parameter c_B. We modify the distribution of topics in Algorithm 1, $p(z_{ij} = k \mid \boldsymbol{z}^{(-i)}, \boldsymbol{w})$, the probability of topic z_{ij} being k given all other topic assignments and all documents, in the algorithm by instead writing that

$$p(z_{ij} = k \mid \boldsymbol{z}^{(-i)}, \boldsymbol{w}) \propto (n_{dk} + c_\alpha \alpha_k) \frac{n_{kw} + c_B B_{tw}}{n_k + \sum_{w=1}^{V} c_B B_{tw}} \tag{3.1}$$

An interpretation of $c_B B_{tw}$ is that we are adding pseudocounts of the peak or gene w to the topic t. This gives us finer control over the prior distribution of the topic-peak or topic-gene

Result: Topic-peak matrix n_{kw} estimating $\boldsymbol{\phi}$ and cell-topic Matrix n_{dk} estimating $\boldsymbol{\theta}$

Input: Cells $\{d_i\}$ each with peaks w_{ij}

Assign counts n_{dk}, n_{kw} and n_k

Initialize topic labels z_{ij}

for *iterations in 1:n_iterations*

  for *each cell* d_i

    for w_{ij} *in* d_i

      $topic = z_{ij}$

      $peak = w_{ij}$

      $n_{d,topic} = n_{d,topic} - 1$

      $n_{peak,topic} = n_{peak,topic} - 1$

      $n_{topic} = n_{topic} - 1$

      for k *in topics*

        $p(z_{ij} = k \mid \boldsymbol{z}^{(-i)}, \boldsymbol{w}) \propto (n_{dk} + c_\alpha \alpha_k) \frac{n_{kw} + c_\beta \beta_w}{n_k + \sum_{w=1}^{V} c_\beta \beta_w}$

      end

      Pick a new topic according to $p(z_{ij} = k \mid \boldsymbol{z}^{(-i)}, \boldsymbol{w})$

      $n_{d,topic} = n_{d,topic} + 1$

      $n_{peak,topic} = n_{peak,topic} + 1$

      $n_{topic} = n_{topic} + 1$

    end

  end

end

Algorithm 1: LDA Gibbs sampling algorithm to estimate topic assignments for each word, by calculating $p(z_{ij} = k \mid \boldsymbol{z}^{(-i)}, \boldsymbol{w})$, the probability of topic z_{ij} being k given all other topic assignments and all documents

matrix. A key feature of using a matrix prior is that we can use the inferred ϕ from one LDA model as the prior for another model with the same vocabulary and similar topics, thereby transferring information from one model to another. We note that this proposed method is similar to a method first proposed in [Wood et al., 2017], but the exact form of the prior is different due to the hyperparameter c_B.

We call the dataset that we use to generate the matrix prior the "reference dataset" and the dataset to which we apply the prior the "target dataset." In order to transfer information from the reference to the target, we first run the LDA algorithm on the reference, where both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are set to be the "uniform priors" described in Algorithm 1; in other words, $\alpha_i = 1/T$ and $\beta_j = 1/V$. This outputs an estimate of the topic-peak or topic-gene matrix, $\hat{\boldsymbol{\phi}}_{\boldsymbol{ref}}$, that contains information about the distribution of peaks or genes in each topic. We set the prior $\boldsymbol{B} = \hat{\boldsymbol{\phi}}_{\boldsymbol{ref}}$, with a concentration parameter c_B. As the concentration parameter increases, the LDA algorithm upweights $\hat{\boldsymbol{\phi}}_{\boldsymbol{ref}}$.

3.5 Methods

3.5.1 Data

Simulated data

To evaluate the performance of the matrix prior methodology, we first simulated scATAC-seq data and compared inferred topics to the true generative topics that we used to create the synthetic data according to the LDA generative process (Section 3.4.1). We fixed the true topic-gene distribution across all simulated datasets, with peak or gene distributions for each topic simulated with a Dirichlet distribution with parameters $c_\beta = 0.1$, $\beta_j = 1/V$, resulting in V-element vectors. We simulated a new cell-topic matrix for each new dataset, also using a Dirichlet distribution, with parameters $c_\alpha = 0.3$, $\alpha_i = 1/T$ resulting in a T-element vector. We set the number of peaks/genes $V = 8000$, the number of topics $T = 30$, and the number of reads/UMIs per cell to be on average $\xi = 4000$.

In the *true matrix simulation*, the goal was to evaluate the matrix prior when the true

topic-gene matrix was used as the matrix prior in LDA of simulted data. We simulated target datasets with 1000, 2000, 4000, and 8000 cells, all with the same topic-gene matrix. For each dataset size, we simulated five datasets. In the *inferred matrix simulation*, instead of providing the true topic-gene matrix, we inferred the topic-gene matrix by performing an LDA of simulated reference data using a uniform symmetric prior, then provided that inferred matrix as the prior to LDA of simulated target data sets. We generated four synthetic reference datasets: one for each of 1000, 2000, 4000, or 8000 cells, and we simulated four target datasets of 1000 cells. We performed LDA with a uniform symmetric prior on each reference dataset, and used each resulting topic-gene output matrix as a prior for LDA of each target dataset.

C. elegans *scATAC-seq data*

Recently published *C. elegans* scATAC-seq data demonstrated increased resolution of cell types compared to bulk ATAC-seq data [Durham et al., 2021]. The cells were collected from animals in larval stage 2, and the authors used LDA followed by clustering to identify cell types, which are the labels we use here. The dataset has 30,764 cells and 13,734 peaks. We split the cells uniformly at random into a "reference dataset" of 27,764 cells and a "target dataset" of 3,000 cells to investigate the performance of the matrix prior.

SHARE-seq mouse skin data

SHARE-seq is a co-assay that generates both scATAC-seq and scRNA-seq data from the same single cells simultaneously [Ma et al., 2020]. For our analysis we selected the mouse skin data set from the original publication, which has 34,774 cells from 22 cell types. The scATAC-seq data for these cells consist of chromatin accessibility across 344,592 peaks, while the scRNA-seq data contain the expression measurements of 22,813 genes. First, we repeated our experiments from the *C. elegans* dataset, testing the effectiveness of the matrix prior for transferring information from a larger reference dataset to a smaller target dataset. Second, we leveraged the co-assay data, in which we know the ground-truth pairing of scATAC-seq

and scRNA-seq measurements for each cell, to investigate whether we could use the matrix prior to transfer information between scRNA-seq and scATAC-seq datasets.

3.5.2 LDA analysis

LDA analysis of simulated data

For our analysis of simulated data, we used the implementation of the LDA algorithm that employs the Gibbs sampling scheme described in Algorithm 1 [Durham et al., 2021]. We set the number of Gibbs sampling iterations to be 1000. The iteration with the highest posterior probability was used to infer the cell-topic and topic-gene matrices.

In the true matrix simulation, we used the true generative topic-gene matrix as the matrix prior for the target data LDA. We compared the quality of the inferred cell-topic and topic-gene matrices both using the matrix prior and using the uniform prior. For the uniform prior, we supplied LDA with the parameters $c_\alpha = 0.3$, and $c_\beta = 0.1$, matching the simulation Dirichlet parameters. When we used the matrix prior, we supplied LDA with the concentration parameter $c_\alpha = 0.3$, which matched the simulation Dirichlet parameter, and we set the matrix prior concentration parameter to $c_B = 1000$.

In the inferred matrix simulation, we used LDA with a uniform prior to infer a topic-gene matrix $\hat{\boldsymbol{\phi}}$ using each of the four reference datasets with between 1000 and 8000 cells. We trained these uniform prior LDA models with $c_\alpha = 0.3$, and $c_\beta = 0.1$. We then used each resulting $\hat{\boldsymbol{\phi}}$ as the matrix prior for each of the four target datasets. We again compared using the matrix prior to using a uniform prior. We used the same settings for the matrix prior LDA models as in the true matrix simulation.

LDA analysis with experimental data: transfer from scATAC-seq to scATAC-seq

To assess whether using a matrix prior would improve LDA performance compared to a uniform prior on small, sparse data, we first trained an LDA model with a uniform prior on all available cells to compare to using our prior on a small dataset. We will refer to this

model as the "joint model". Next, we split the data into a larger "reference set" of 31,774 cells and a smaller "target set" of 3000 cells, and trained a uniform prior LDA model on each of these sets. The output from the target data when the uniform prior was used set the floor for the expected performance of our prior. Last, we used the gene-topic probabilities from the reference model as a matrix prior for a new LDA model of the target set of cells, and we compared the results of this model with those of the uniform models to evaluate the effectiveness of the matrix prior.

Hyperparameter search

We conducted a hyperparameter search to inform our selection of values for the number of topics to use, T, and the two concentration parameters c_α and c_B. We used a grid-search strategy, testing seven values for T (2, 3, 4, 5, 10, 15, and 20), four values for c_α (0.03, 0.3, 3.0, and 30.0), and eight values for c_B (10, 50, 75, 150, 250, 1000, 2000, and 4000); and we employed a likelihood-based measure, perplexity, as our evaluation metric, following Wallach et al. [2009]. Perplexity is the negative exponent of the likelihood, and is calculated for each cell as $\exp(-\frac{\mathcal{L}(\mathbf{w}|\mathbf{z},\phi,c_\alpha,c_B)}{N})$, where N is the number of reads in a cell. A lower value of perplexity is better. To calculate the likelihood required for the perplexity measure, we use the Chib-style estimation procedure [Chib, 1995], which is a method using a Markov chain to evaluate $\mathcal{L}(\mathbf{w} \mid \mathbf{z}, \phi, c_\alpha, c_B)$.

Due to our two-tiered method, in which we train a uniform prior LDA on a reference subset of the data and then use the output as a matrix prior for an LDA on the target subset of the data, we also required two corresponding tiers for our hyperparameter search. Thus for the first tier, we generated a set of matrix priors by training, for each value of T in our grid search, a uniform prior LDA model on the reference data set. All of these models used $c_\alpha = 3$ and $c_\beta = 800$, which we chose based on the hyperparameter values reported in Durham et al. [2021]; note that we lowered c_β compared to the published value of 2000 to allow the data to have a greater role in determining the topic-gene matrices we would use as matrix priors.

In the second tier of the hyperparameter search, we conducted the full grid search on the target data set with a range of values for T, c_α, and c_B. We split the target data into ten different training/test splits of 2700 and 300 cells, and then trained each of the ten splits on each of the hyperparameter combinations, with the LDA using the matrix prior from tier 1 that corresponded to each value of T. Then, we evaluated the performance of each hyperparameter combination by computing the perplexity on the held out test set cells.

We found that the optimal number of topics was 10, that the perplexity was relatively insensitive to the choice of c_α, and that perplexity dropped with increasing values of c_B. Following Durham et al. [2021], which suggested that LDA models were robust to extra topics, we increased the number of topics to add some flexibility to the model, and set T to be 15 in our experiments. The results of the hyperparameter search and further comments are in Figures S1 and S2. Note that supplementary materials are found at the end of this chapter. UMAP plots of the results of LDA for different numbers of topics are shown in Figure S3.

We conducted the hyperparameter search for our LDA modeling of the SHARE-seq mouse skin data in similar fashion to the hyperparameter search for the *C. elegans* data. However, the SHARE-seq scATAC-seq dataset contains about ten times more reads per cell than the *C. elegans* dataset, leading to much longer LDA training times per cell and higher memory usage. To make the hyperparameter search more efficient, we limited the SHARE-seq analysis to the 20,000 peaks with the highest variance amongst cells, and randomly downsampled the dataset to 7,000 cells. We split the 7000 cells into 6300 cells for the reference dataset, and 700 cells for the target dataset. We then made ten train/test splits of the target data set by sampling 630 training cells and using the remaining 70 held-out test cells for the Chib method. As with the *C. elegans* data, our hyperparameter search results show that perplexity was not very sensitive to the c_α parameter, and that perplexity decreased as the value of the concentration parameter c_B increases, suggesting that the matrix prior was able to improve the quality of the our inference (Figure S2). However, surprisingly, our hyperparameter search achieved the lowest perplexity with just $T = 2$ topics. We conjectured that this may

be because of the low number of peaks included in the analysis, the low number of cells, or the peaks with highest variance may not be able to distinguish the cells. It is also possible that without a normalization for the mean accessibility, we might not select for peaks with the most meaningful differences in accessibility. We hence opted to instead continue using 15 topics as in the *C. elegans* case. UMAP plots of the results of LDA for different numbers of topics are shown in Figure S4, and we found that 10-15 topics visually had good separation of the previously called cell types.

LDA analysis with experimental data: transfer from scRNA-seq to scATAC-seq

To use scRNA-seq data to generate a matrix prior for the LDA of the scATAC-seq data, our method requires that the scATAC-seq and scRNA-seq data share the same vocabulary. Hence, we translated the scATAC-seq data from a vocabulary of peaks to one of genes by counting the number of ATAC-seq cut sites per cell that overlapped each gene and its promoter (defined as the 2kb region immediately 5' of the transcription start site). We defined cut sites as reported in [Durham et al., 2021]; briefly, they are 60 bp regions centered on the mapping locations of the 5' ends of the paired end reads (which define the extent of the original DNA fragment that was cut out of the genome). We investigated whether this translated scATAC-seq data retained similar information to the native scATAC-seq data by qualitatively comparing UMAP plots of the LDA output of the raw peaks versus the summed cut sites and saw that the plots were qualitatively similar (Figure S5). Furthermore, we investigated the similarity between the scRNA-seq data and the translated scATAC-seq data by testing the correlation between the two data sets based on the number of counts per cell and counts per gene (Figure S6). We found that there was a moderate amount of correlation between the scRNA-seq data and the scATAC-seq data, suggesting that generating a matrix prior from scRNA-seq data and applying it to an LDA of a translated scATAC-seq data set may be useful. See Supplementary note 1 for results and discussion of our work in this direction.

3.5.3 Evaluation

Evaluation of LDA in simulated data

We used the mean squared error (MSE) to evaluate our inferred cell-topic matrix $\hat{\boldsymbol{\theta}}$ and topic-gene matrix $\hat{\boldsymbol{\phi}}$. Then, we calculated the MSEs for the cell-topic matrix and topic-gene matrix using

$$\frac{1}{UT}\sum_{i=1}^{U}\sum_{j=1}^{T}(\hat{\theta}_{ij}-\theta_{ij})^2 \text{ and } \frac{1}{TV}\sum_{i=1}^{T}\sum_{j=1}^{V}(\hat{\phi}_{ij}-\phi_{ij})^2 \tag{3.2}$$

where U is the number of cells, and T is the number of topics.

In addition to MSE, we calculated Pearson's r and Spearman's r after flattening the cell-topic and topic-gene matrices.

One complication in simulation is that the order of topics inferred by LDA will not necessarily match the order of topics in the simulated ground truth; i.e. topic 1 from the output of LDA is not necessarily semantically the same as topic 1 in the simulated true cell-topic matrix. Therefore, prior to calculating any performance measure, we must match topics. We do this by using a greedy approach on the topic-peak or topic-gene matrix, considering each pair of topics in sorted order by Euclidean distance and allowing only one-to-one matches. For each topic, we match it to the true topic that is closest in MSE.

We used this greedy topic matching algorithm when evaluating the uniform prior LDA models in both the true matrix simulation and the inferred matrix simulation, and when evaluating the matrix prior LDA model for the inferred matrix simulation. We did not need to match topics for the matrix prior LDA in the true matrix simulation, because providing the true topic-gene distributions as the prior already imparts the proper topic semantics to the target LDA, making the inferred topic-gene matrix and the true topic-gene matrix directly comparable.

Evaluation of LDA in experimental data

Unlike our simulation experiments, in which we know the true underlying topic distributions, we do not have a ground truth to compare against for our analyses of datasets derived from biological experiments. Instead, to evaluate the performance of our LDA models, we compared the output of each LDA on a target subset to the LDA output when training on the full data set (i.e. the "joint model"). This has the interpretation that if our matrix prior allows the smaller target data set LDA to infer similar topics to the joint model, then our matrix prior has succeeded. The joint model results in topic assignments for all cells, inclusive of both cells assigned to be reference cells and those assigned to be target cells. Hence, to compare the joint model with the target LDA, we selected the rows in the cell-topic matrix of the joint model that corresponded to the target dataset cells. We then applied our greedy topic matching algorithm to account for any topic-switching that occurred in the LDA of the reference data set compared to the joint model. Finally, we evaluated the similarity between the outputs of the matrix prior LDA and the joint model using Pearson's correlation, Spearman's correlation, and MSE. We focus on Pearson correlation in the main text but report other metrics in the Supplementary Material.

In addition to our quantitative evaluation of the matrix prior approach, we qualitatively evaluated its performance by visualizing the cell-topic matrix using UMAP [McInnes et al., 2018]. For the *C. elegans* data, we took the cell-topic matrix LDA output from one of the splits of the target training set, computed a two dimensional UMAP embedding, and represented the embedding as a scatter plot in which we colored the cells based on their published cell type labels [Durham et al., 2021]. We quantitatively measured how well the inferred cell-topic matrices could separate cells of different types (both at the level of broader cell types and more specific ones, such as neuron subtypes) by using the silhouette coefficient. The silhouette coefficient is a distance-based measure of cluster cohesiveness that has previously been used as a performance metric when comparing the performance of different single cell analysis methods Luecken et al. [2022], and it is computed using Equation

3.3. For a cell with index i, and C_i the set of indices of cells with the same cell type label as cell i, $a(i) = \frac{1}{C_i - 1} \sum_{j \in C_i, i \neq j} d(i,j)$ is the average Euclidean distance to all other cells of the same cell type. We similarly define $b(i) = \min_{k \neq i} \frac{1}{C_k} \sum_{j \in C_k} d(i,j)$ to be the lowest average Euclidean distance to a different cell type from cell i.

$$\frac{b(i) - a(i)}{\max(b(i), a(i))} \tag{3.3}$$

The resulting silhouette coefficent is bounded between -1 and 1. Values close to 1 indicate that a point is closer to points in the same cluster than to points in other clusters, and a high silhouette coefficient indicates that the LDA topics reflect the published cell types well.

3.5.4 Data availability

The code to run LDA with a matrix prior is available at https://github.com/gevirl/LDA using the -betaFile option. There are no primary data in the paper. A vignette of running this software is available at https://github.com/Noble-Lab/lda_matrix_prior

3.6 Results

3.6.1 Matrix prior improves topic inference in simulation

We began testing the matrix prior by running LDA on simulated data derived from known cell-topic and topic-gene matrices. In the true matrix simulation, we tested the hypothesis that an LDA given the true topic-gene matrix as a prior would yield results more similar (by MSE) to the ground truth than the uniform prior LDA would. We also tested the hypothesis that increasing the number of target cells would reduce the performance advantage of the matrix prior over the uniform prior. In the inferred matrix simulation, we tried using LDA output on a reference data set as our matrix prior for the target LDA and tested the effect of varying the reference data set size on the performance of the target LDA.

In the true matrix simulation, we found that using the matrix prior instead of the uniform prior led to more accurate inference of the topic-gene and cell-topic matrices (Figure 3.1a).

We kept the weight of the matrix prior, c_B, constant and varied the number of cells in the target dataset to understand effects of target dataset size. The MSE was higher when we used a uniform prior than when we used the matrix prior regardless of target dataset size (we tested 1000, 2000, 4000, and 8000 target cells), although the difference in performance decreased as the cell number increased, suggesting that with more cells the data began to overwhelm the prior. To better understand how the matrix prior improves the LDA results, we analyzed some representative uniform prior LDA results in more detail. When the number of cells was low, the uniform prior LDA underestimated the weights of high probability topics, and even at 8000 cells, it incorrectly predicted topic assignments in some cells (Figure S7, top row). On the other hand, for the matrix prior LDA, the inferred and true cell-topic matrices agreed (Figure S7, bottom row). We found similar results for the topic-gene matrices (Figure S8).

In the inferred matrix simulation, we found that the matrix prior improved the inference of the cell-topic and the topic-gene matrices compared to the uniform prior, as measured by MSE against the ground truth (Figure 3.1b). We also trained a series of LDA models on a 1000 cell target data set, each with a matrix prior generated from a uniform LDA trained on a reference dataset with 1000, 2000, 4000, or 8000 cells. As the number of reference cells increased, the MSE continually improved (x-axis), and we note that even the matrix prior generated from a reference dataset of only 1000 cells improved LDA performance compared to a uniform prior (Figure 3.1b). In representative simulated datasets, the cell-topic and topic-gene matrices more closely followed the $y = x$ line as the number of reference cells increased (Figure S9). These two simulations suggest that under ideal conditions, the matrix prior method is able to improve the quality of inferred topics in LDA.

3.6.2 Whole worm scATAC-seq prior improves concordance with joint model

We used the *C. elegans* data to validate the ability of our matrix prior to improve LDA inference on real data. We randomly split the 34,764 cells into a 3,000 cell target dataset and a 27,764 cell reference dataset, and used the split data to train a matrix prior LDA on the

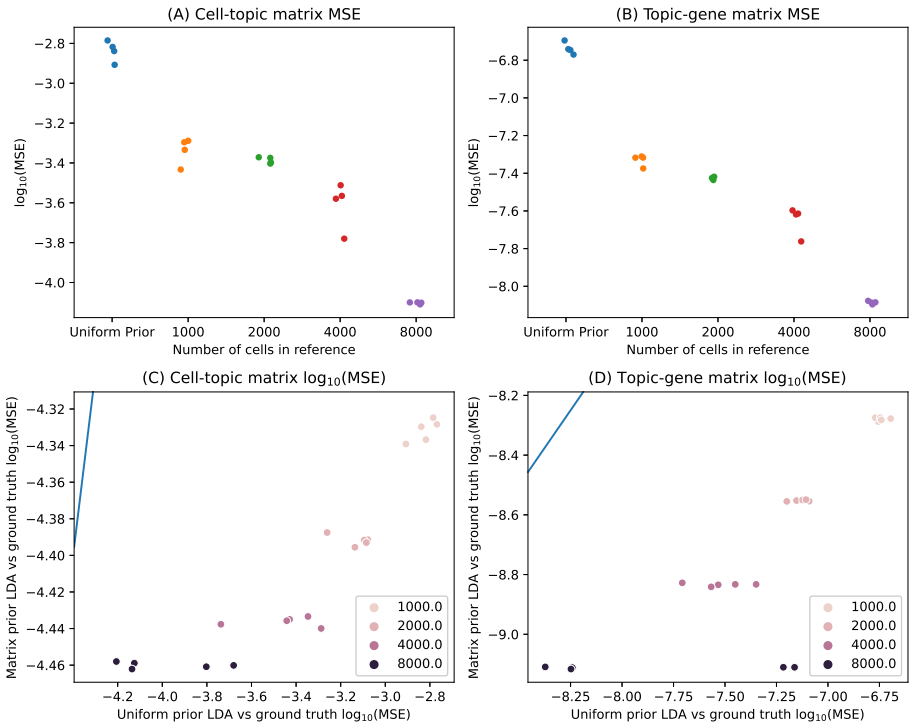

Figure 3.1: Simulation experiments show that the matrix prior improves the concordance between inferred topics and the ground truth compared to the uniform prior. Experiments from the true matrix simulation and the inferred matrix simulation are shown here for different numbers of cells in the target dataset (different colors). (a) MSE from the ground truth to the LDA with a ground truth matrix prior (y-axis) is plotted against the MSE from the ground truth to the uniform symmetric prior LDA (x-axis), for both the cell-topic matrix (left) and the topic-gene matrix (right). The blue line is the line $y = x$. Each point represents one independently simulated dataset, with a unique true cell-topic matrix and topic-gene matrix. (b) MSE to the ground truth for the LDA with a matrix prior inferred from a simulated reference dataset is shown for different reference data set sizes. MSE is plotted for both the cell-topic matrices (left) and the topic-gene matrices (right).

target dataset. Then, we trained a separate uniform prior LDA on the full *C. elegans* dataset

(the "joint model"), and compared the inferred topics between the matrix prior LDA and the joint model. We also trained a uniform prior LDA on the target dataset, and evaluated whether the matrix prior LDA results were more similar to the the joint model than the uniform prior LDA results.

We found that the Pearson correlation between the target LDA and the joint model was higher when we used the matrix prior than when we used the uniform prior (Figure 3.2, left). The correlation increased with increasing c_B, but eventually reached a saturation point. We also found that the matrix prior outperformed the uniform prior in Spearman correlation and MSE (Figure S10). When plotting the values of the cell-topic and topic-gene matrices, the points near the $y = x$ line for the cell-topic matrix tightened as the concentration parameter c_B increased (Figure S14). We also additionally note that in our hyperparameter search, we found evidence that the use of the matrix prior improved the quality of the topic-gene matrix, since the perplexity value in the held out test set improved as the matrix prior concentration parameter increased (Figure S1)

3.6.3 SHARE-seq scATAC-seq matrix prior improves concordance with the joint model

We used data from mouse skin cells analyzed using the SHARE-seq assay [Ma et al., 2020] to further validate the ability of a matrix prior to improve inference in LDA. We split the 34,774 cells into a 31,774 cell reference dataset and a 3000 cell target dataset, and the same analyses were applied as in the *C. elegans* data (Section 3.6.2).

We note that increasing the value of c_B more consistently improved the correspondence between the target LDA and the joint model for the topic-gene matrix than the cell-topic matrix. This difference is most likely due to the fact that the matrix prior is specified as a prior on the topic-gene distribution, and thus only influences the cell-topic distribution indirectly through the training of the LDA model.

In addition to serving as another dataset to validate our scATAC-seq prior, the SHARE-seq co-assay data allowed us to evaluate whether a matrix prior generated from one data modality, scRNA-seq, could improve LDA performance on another data modality, scATAC-

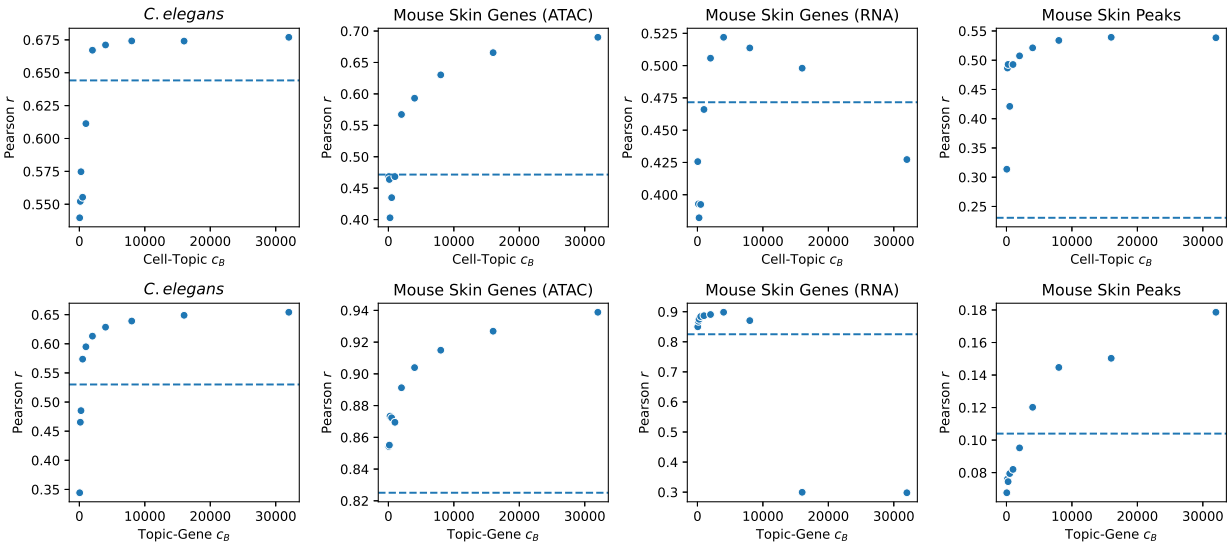

Figure 3.2: The Pearson correlation between LDA results on the target dataset for the matrix prior LDA and the full dataset uniform symmetric LDA (the "joint model") increases as c_B increases. Pearson r values are plotted as a function of c_B for the cell-topic (top row) and topic-gene (bottom row) matrices for LDA experiments on four different datasets: *C. elegans* scATAC-seq data (first column), SHARE-seq mouse skin scATAC-seq data with the peak vocabulary translated to genes (second column), SHARE-seq mouse skin scRNA-seq data (third column), and SHARE-seq mouse skin scATAC-seq data using the peak vocabulary (fourth column). The dotted horizontal lines indicate the correlation between the uniform prior LDA and the joint model.

seq. Because the SHARE-seq scATAC-seq and scRNA-seq data were generated from the same cells, we were able to directly assess the agreement between the scRNA-seq LDA and the scATAC-seq LDA, with and without a matrix prior derived from the scRNA-seq data (Section 3.5.2). See Supplementary Note 1, where we report that the scRNA-seq prior was able to improve inference for moderate values of c_B but worsened inference for larger values. Note that although scRNA-seq and scATAC-seq produce data on different scales, this does not affect the matrix prior because it is based on the topic-gene probabilities (which sum to

1 for each topic) and not the counts.

3.6.4 Mouse skin cell types are more clearly separated with the use of the matrix prior

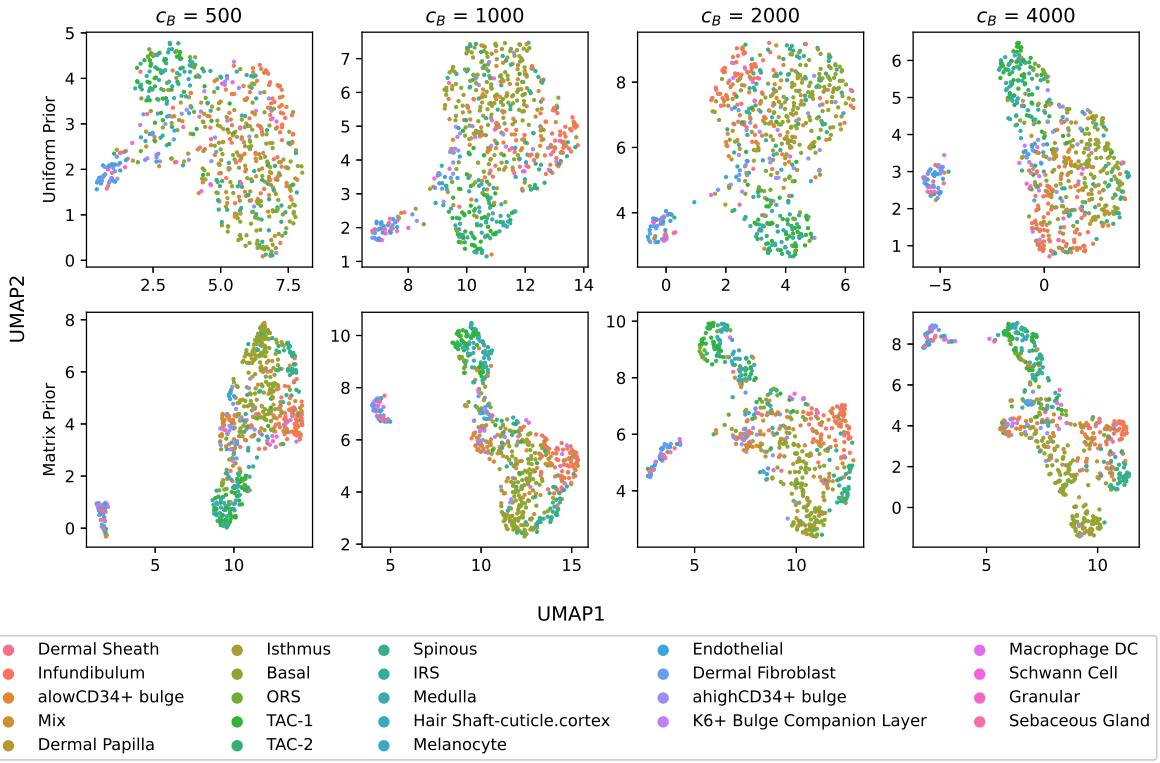

Figure 3.3: Increasing the weight of the matrix prior (bottom row) shows a qualitative improvement in the ability of the target dataset LDA to discriminate among cell types compared the uniform prior (top row). UMAP embeddings of the cell-topic matrices from SHARE-seq mouse skin scATAC-seq data using the peak vocabulary are trained with different values of c_B (different columns). Scatter points representing cells are colored by their published cell type annotations.

We next aimed to assess whether using a matrix prior would improve the ability of LDA to distinguish the cell types in a small subset of the SHARE-seq mouse skin data set [Ma

et al., 2020]. We split the data into a reference data set and a target data set (see Methods), and then we analyzed the target data set using both a uniform prior and a matrix prior derived from LDA on the reference data set. After applying UMAP to our two models to reduce the 15-dimensional topic space into a two-dimensional UMAP space, we qualitatively observed that the matrix prior LDA resulted in cell clusters that better agreed with the published cell type labels than the uniform prior LDA, and this improvement became more marked as we increased the weight of the prior, c_B (Figure 3.3). We used the silhouette score to quantitatively measure how well the cells clustered by their cell type labels (Figure S16), but the silhouette values did not improve with use of the matrix prior. We conducted a similar experiment in *C. elegans data* (Supplementary Note 2), but did not see qualitative improvement of cell clusters in *C. elegans*. A possible reason for this is that even in the case of the uniform prior, LDA created separation in the *C. elegans* cell types, and hence no further improvement was possible by using the matrix prior.

We also used the perplexity measure to quantitatively measure how well the model was doing when different weights of the prior were used (Figure 3.4). This is a method that is similar to that of the hyperparameter search (Section 3.5.2). We found that as we trained matrix prior models with increasing values of c_B, the perplexity decreased. This was true both for *C. elegans* and SHARE-seq. For both cases, the uniform prior resulted in greater (worse) values of perplexity. For *C. elegans*, perplexity values for both the uniform prior and matrix prior decreased as c_B increased. On the other hand, for the SHARE-seq data, the perplexity values decreased for the matrix prior, but not for the uniform prior.

3.7 Discussion

We have shown through both simulation study and through analysis of real data that the matrix prior we propose is able to capture information from a larger reference dataset and impart the semantics of the topic-gene or topic-peak matrix onto a smaller target dataset. In our simulation studies, we found that when the true topic-gene and cell-topic matrices were known, we were able to recover those matrices both by directly inputting the truth as the

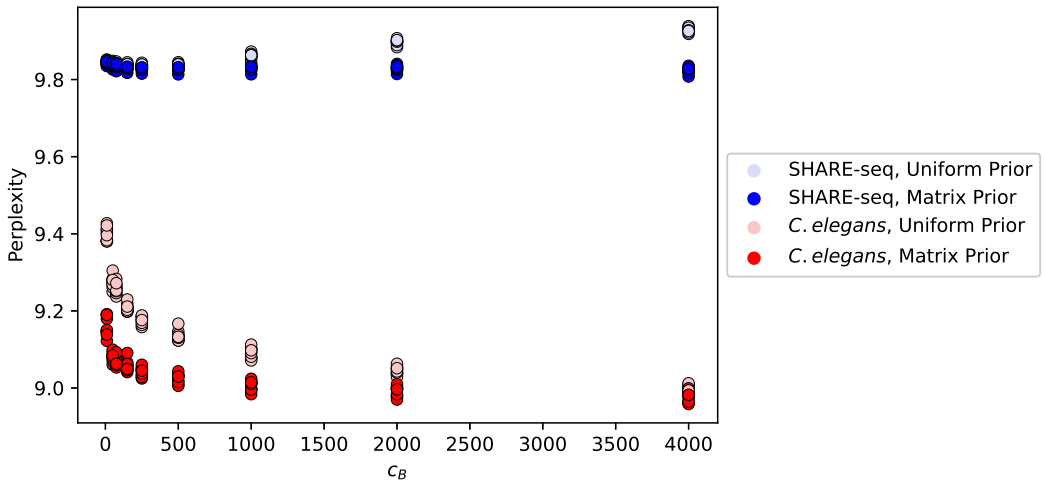

Figure 3.4: Perplexity values (y-axis) demonstrate quantitative improvement of the LDA model after using the matrix prior (darker colors) compared to the uniform prior (lighter colors) for various values of the weight of the prior (x-axis). The same procedure was used for both the SHARE-seq data set (blue) and the *C. elegans* data set (red). Each point is a separate split of the target data into a test and a training set.

prior and by inferring the topic-gene matrix from a reference data set (Figure 3.1). These simulations demonstrated that in ideal conditions, the matrix prior can greatly improve performance of LDA. In our real data examples, we examined *C. elegans* and mouse skin cell data. We saw promising results when we used scATAC-seq data to generate a prior for analyzing a target scATAC-seq dataset. We found that LDA results on a small target dataset were more concordant with a model of the full data set when we used a matrix prior derived from a larger reference dataset than when we used a uniform prior (Figure 3.2). Furthermore, in the case of the mouse skin data, we showed qualitative improvements in cell type discrimination for the matrix prior LDA compared to the uniform prior LDA (Figure S16).

We also attempted to transfer information across single cell data modalities by deriving a matrix prior from scRNA-seq data and applying it to a target scATAC-seq dataset. We

found that with moderate values of the concentration parameter, the agreement between the outputs of LDA based on the target dataset and LDA based on the full data set improved when using the matrix prior compared to the uniform prior (Figure S12). Leveraging information from multiple single cell data modalities is an area of active research. Some popular tools, like Seurat [Butler et al., 2018], scVI Tools [Gayoso et al., 2021], or LIGER [Liu et al., 2020], take multiple data sets as input and use embedding techniques to analyze them jointly. These approaches are powerful, but require manipulating potentially very large datasets every time one wants to add new data into the model. In contrast, our matrix prior approach requires just a single large upfront compute task to train an LDA on a reference dataset, which yields a compact gene-topic matrix that can be used as a matrix prior for training comparatively lightweight LDA models in all subsequent analyses of new datasets. We also anticipate that new approaches, such as Polarbear [Zhang et al., 2022] and BABEL [Wu et al., 2021], that use deep learning models to translate data from one single cell modality to another will improve our ability to generate cross-modality matrix priors, not only between scATAC-seq and scRNA-seq data, but also between other pairs of modalities.

3.8 Supplementary Material

Supplementary note 1: SHARE-seq scRNA-seq matrix prior may have limited use

We investigated the feasibility of using scRNA-seq data to construct a matrix prior for scATAC-seq analysis. This required that we map the scRNA-seq data and the scATAC-seq data onto a shared feature axis. To accomplish this mapping, translated the scATAC-seq data from a vocabulary of peaks to one of genes by simply counting the number of scATAC-seq cut sites overlapping each gene and its promoter (see Methods). We then proceeded with a similar analysis to Sections 3.6.2 and 3.6.3. To evaluate the performance of the matrix prior, we leveraged the fact that the scATAC-seq data and scRNA-seq data were generated from the same set of cells, and compared the inferred cell-topic and topic-gene matrices from the matrix prior LDA that was trained on the target scATAC-seq dataset to the output

matrices generated by a joint model on the full scRNA-seq dataset.

We found that using the scRNA-seq data to construct the matrix prior did not consistently improve inference of the cell-topic and topic-gene matrices, although we saw some improvement at moderate values of c_B. For the cell-topic matrices, we found that as the concentration parameter c_B increased beyond 4,000, the Pearson correlation to the joint model output tended to decrease. Similarly, Spearman correlation and MSE both got worse as c_B increased beyond 4,000. For the topic-gene matrices, however, Spearman correlation improved as c_B increased (Figures 3.2, S12). When plotting the raw values of the cell-topic matrix, we observed more points along the x- and y-axes. The reference model and joint model assigned low probability topics to different cells more frequently as c_B increased (Figure S14). For the topic-gene matrix, the diagonal had increased density, meaning that the two analyses agreed more as c_B increased (Figure S15)

We note that the read counts per gene and per cell in scRNA-seq was only moderately correlated with the number of cut sites summed over the gene body in the scATAC-seq data, with a Pearson correlation of 0.661 for the signal per cell, and 0.765 for the signal per gene (Figure S6). This makes sense because, although chromatin accessibility and gene expression are highly related biological phenomena, the information encoded by scATAC-seq and scRNA-seq count data is not the same. This could be one reason why the matrix prior only improves results compared to the uniform prior at moderate values of c_B. If a prior derived from a different data modality is given too much weight, then the discrepancies between the modalities are more likely to lead to a poor model fit.

Supplementary Note 2: C. elegans silhouette values did not improve through use of the prior

The *C. elegans* data was previously labeled with a cell type based on marker genes through a clustering method that took into account all scATAC-seq peaks using all the cells together. We hypothesized that we would be better able to recover these cell type labels in the target dataset by incorporating information from the matrix prior. To this end, we analyzed the target dataset using both the matrix prior derived from a reference subset of cells and the

uniform prior, and then evaluated how well the clusters in the cell-topic output agreed with the published cell type labels.

We used UMAP [McInnes et al., 2018] to reduce the 15-dimensional topic space to a two-dimensional representation and then colored each cell according to its cell type label. We compared these UMAP plots of the results of our LDA analysis with the matrix prior and with the uniform prior, as well as with different weights of c_B (Figure S17). The UMAP plots show that all of the LDA models produce reasonable agreement with the cell type labels, although we note that quantifying the cell type discrimination in each LDA output with the silhouette score shows a slight increase in the mean silhouette value, from 0.216 with the uniform prior to 0.220 with the matrix prior and $c_B = 4000$ (Figures S16, S18).

Despite the similarity of the UMAP plots, we noted that at $c_B = 4000$ the neurons appeared to split into two clusters, and the average silhouette score for the neurons decreased from 0.254 with the uniform prior to 0.109 with the matrix prior and $c_B = 4000$. This observation suggested the hypothesis that the use of the prior might help to resolve fine grained cell types, so we repeated our cell type label analysis specifically on the neurons, this time using the published neuron subtype labels (Figure S19). The UMAP plots show that as c_B increased the cells were clustered more tightly together, however there was little change in how well the cell types were separated compared to the LDA with the uniform prior. In addition, the mean silhouette scores for the neuron subtypes (Figure S20), did not increase as the value of c_B increased. Overall, these analyses do not suggest that the neuron subtypes were further resolved by the use of the matrix prior.

Variable	Definition
N	The number of reads in a cell
T	The number of topics
V	The number of peaks or genes in the vocabulary
U	The number of cells
$\boldsymbol{w}$	A vector of genes or peaks in a cell
$\boldsymbol{\theta}$	A vector $(\theta_1, \ldots, \theta_T)$; the topic distribution of a given cell
$\hat{\boldsymbol{\theta}}$	Inferred $\boldsymbol{\theta}$
$\boldsymbol{\phi}$	A matrix such that ϕ_{tw} is the probability of observing a peak or gene w for a topic t
$\hat{\boldsymbol{\phi}}$	Inferred $\boldsymbol{\phi}$
$\boldsymbol{\alpha}$	A basis vector of length T that parameterizes the Dirichlet prior distribution over the cell-topic matrix.
$\boldsymbol{\beta}$	A basis vector of length V that parameterizes the Dirichlet prior distribution over the topic-peak/gene matrix.
c_α	The concentration parameter for $\boldsymbol{\alpha}$
c_β	The concentration parameter for $\boldsymbol{\beta}$
$\boldsymbol{z}$	A vector of topic assignments corresponding to $\boldsymbol{w}$
ξ	The average number of peaks or genes in a cell
$\boldsymbol{B}$	A $T \times V$ matrix that parameterizes the matrix prior for the topic-peak/gene matrix.
c_B	A concentration parameter for $\boldsymbol{B}$
$\hat{\boldsymbol{\phi}}_{ref}$	The output inferred $\boldsymbol{\phi}$ of an LDA analysis on the reference dataset.

Table S8: Glossary of variables used

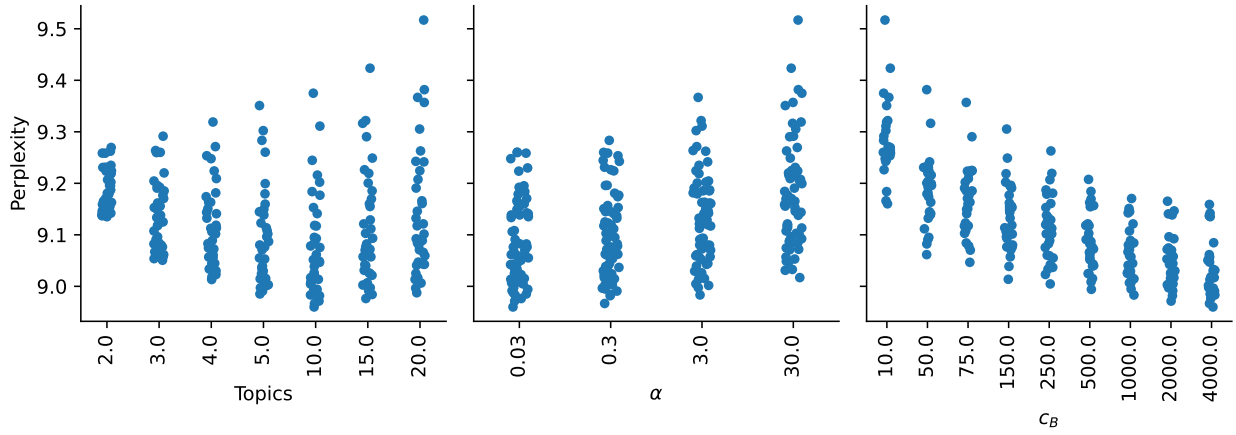

Figure S1: Hyperparameter search for *C. elegans* data was used to optimize the parameters. Each point is the average of 10 folds of the perplexity value of the test set. The x-axis is the value of the hyperparameter, and the y-axis is the perplexity value.

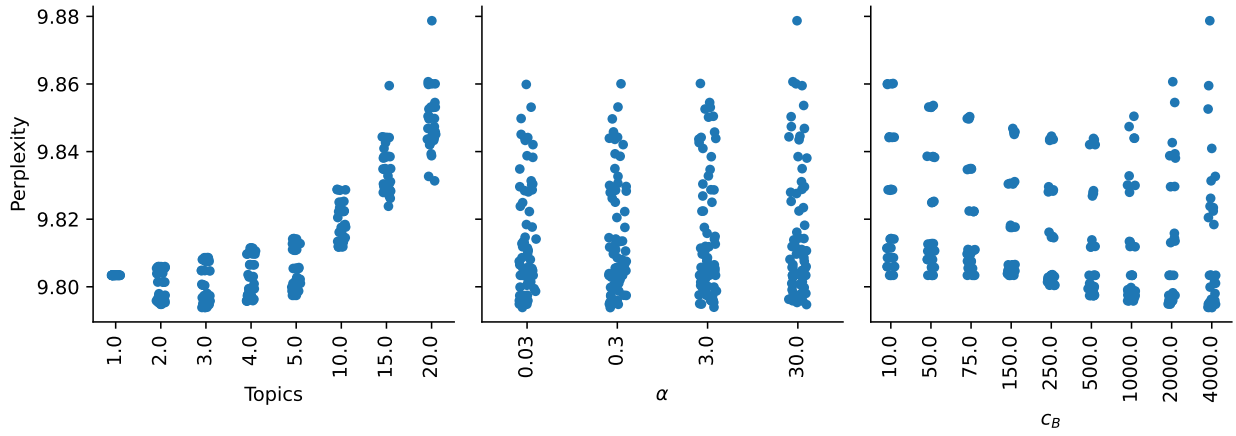

Figure S2: Hyperparameter search for SHARE-seq data with peaks was used to optimize the parameters. Each point is the average of 10 folds of the perplexity value of the test set. The x-axis is the value of the hyperparameter, and the y-axis is the perplexity value.

Cell Type	Count
Basal	159
Infundibulum	74
TAC-1	56
Spinous	52
Mix	46
alowCD34+ bulge	32
Hair Shaft-cuticle.cortex	29
Endothelial	22
ahighCD34+ bulge	22
Medulla	19
Dermal Fibroblast	18
ORS	18
Isthmus	13
IRS	12
Dermal Sheath	11
TAC-2	10
Dermal Papilla	8
Macrophage DC	8
K6+ Bulge Companion Layer	7
Granular	6
Schwann Cell	3
Melanocyte	3
Sebaceous Gland	2

Table S8: Table of counts of mouse skin cell types used in the target dataset.

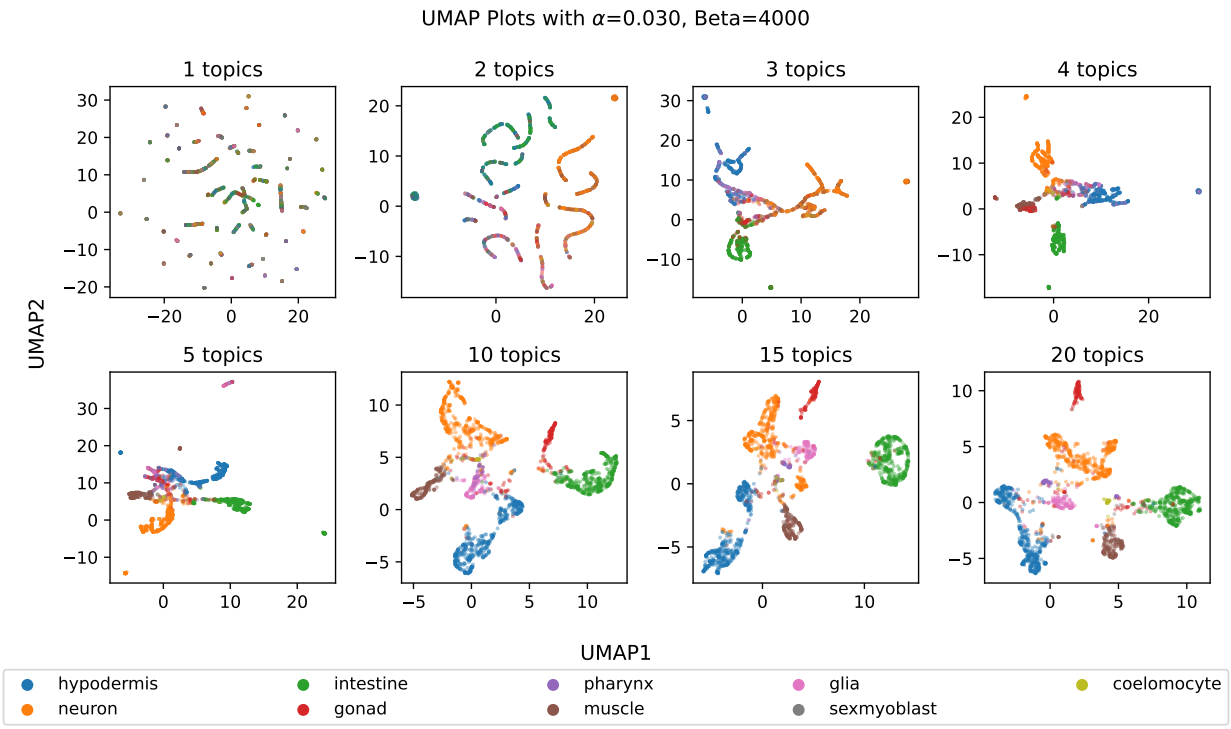

Figure S3: UMAP embeddings for different numbers of topics are shown to provide intuition on the effect of the number of topics. Embeddings were made for cell-topic matrices from matrix prior LDA on *C. elegans* scATAC-seq data using a fixed value of $c_B = 4,000$ and 13,734 peaks, while varying the number of topics.

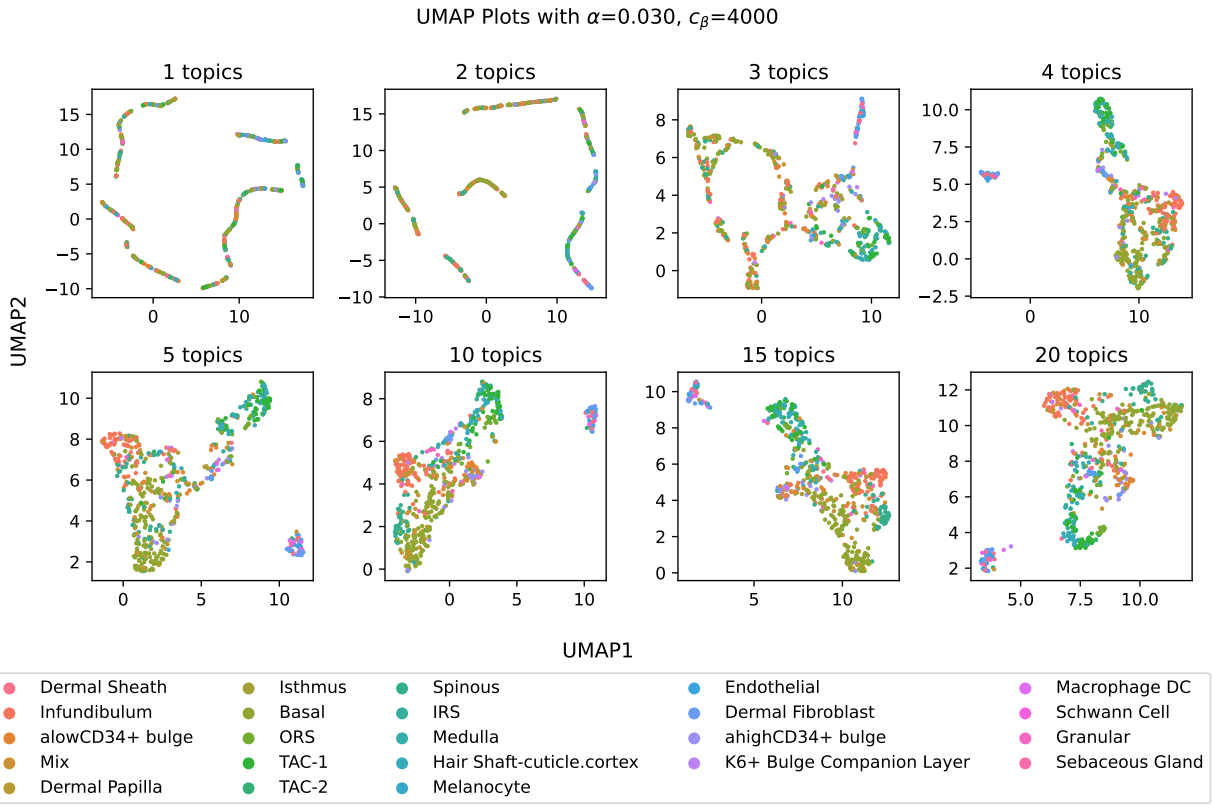

Figure S4: UMAP embeddings for different numbers of topics are shown to provide intuition on the effect of the number of topics. Plots were made with 630 target cells from the mouse skin SHARE-seq peak data subsetted to 7,000 cells and 20,000 most variable peaks. LDA was run with $c_\beta = 4,000$ using the uniform prior, while varying the number of topics.

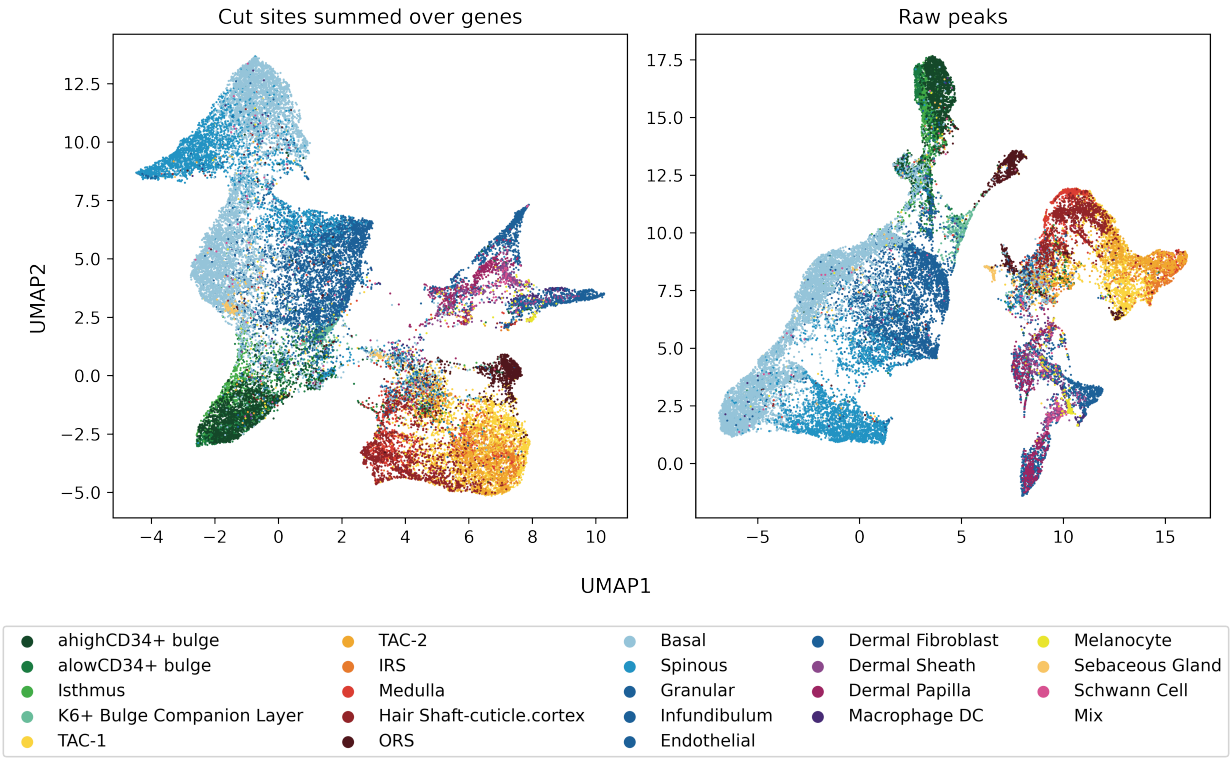

Figure S5: UMAP was applied to the cell-topic matrices output from LDA joint model to qualitatively compare cut sites summed over genes versus peaks. LDA was run with 15 topics, $c_\alpha = 3$, and $c_\beta = 4000$. On the left, the raw data fed into LDA are the cut sites summed over the 22,813 genes, as described in Section 3.5.2. On the right, the data fed into LDA are the 344,592 raw peaks.

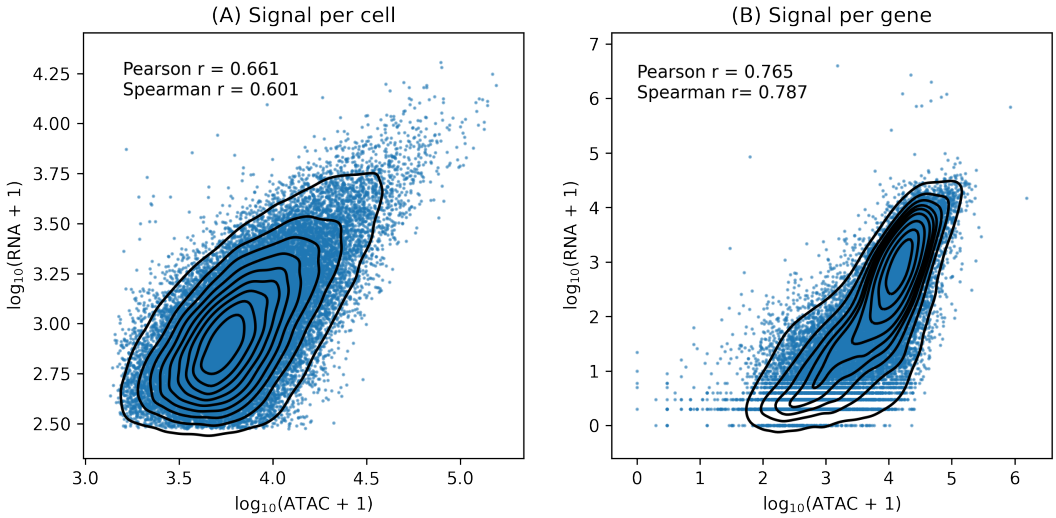

Figure S6: RNA reads and scATAC-seq cut sites summed over gene bodies are compared to determine the shared information between the data modalities. On the left, the signal per cell is the log plus one of the total number of counts for each cell (i.e. summing across all the genes in a cell). On the right, the signal per gene is the log plus one of the total number of counts for each gene (i.e. summing across all the cells for a gene). The Pearson correlation is reported in each plot. A kernel density estimator is overlayed on the data. The x-axis shows the score for the scATAC-seq data, and the y-axis shows the score for the scRNA-seq data.

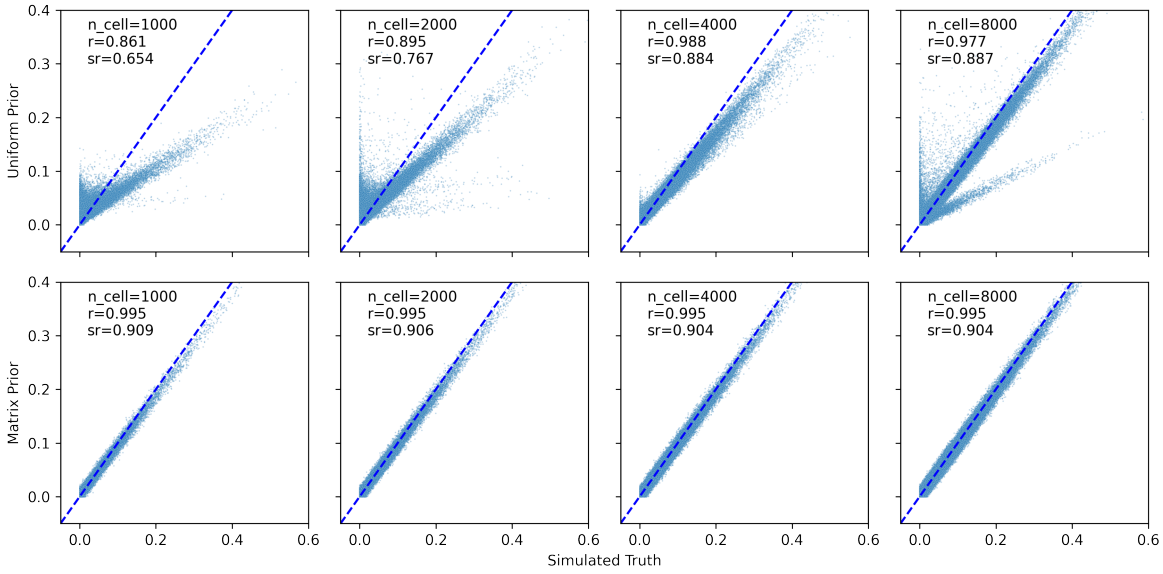

Figure S7: Scatter plots of cell-topic matrix values demonstrate the improvement of the matrix prior over the uniform prior in the true matrix simulation and further show that the performance of the uniform prior approaches that of the matrix prior as the number of cells increases. Plots show simulated true values (x-axis) of the cell-topic matrix against inferred values using LDA (y-axis). Pearson r (r) and Spearman r (sr) are reported for each plot. We compared different numbers of cells in the target dataset (different columns). We compared LDA with a uniform prior (top row) with a matrix prior generated from the true topic-gene matrix (bottom row). The blue dotted line is the line $y = x$.

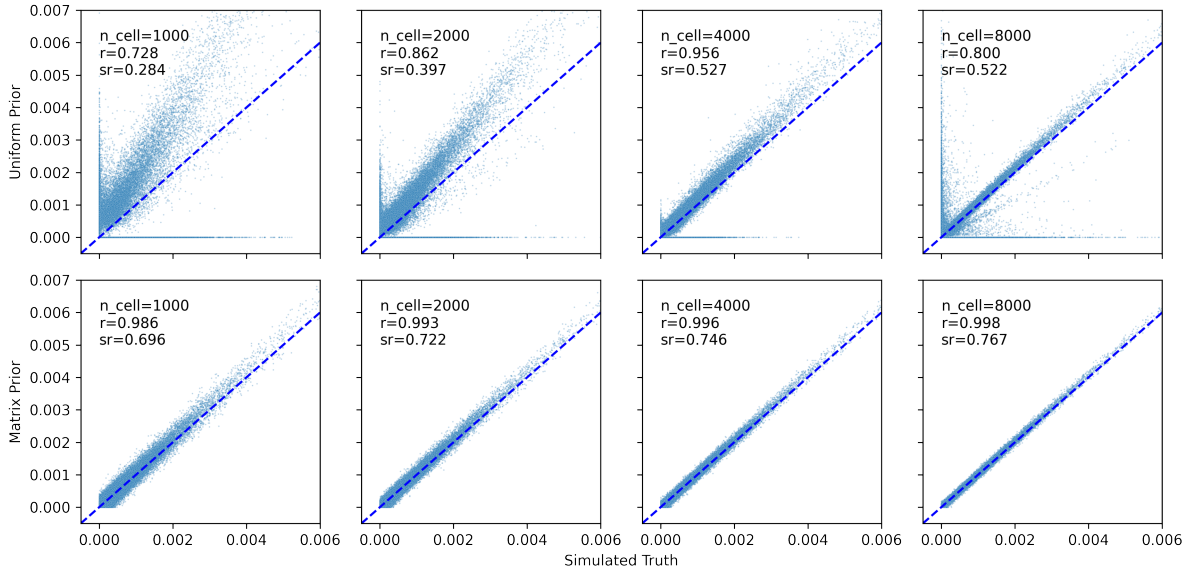

Figure S8: Scatter plots of topic-gene matrix values demonstrate the improvement of the matrix prior over the uniform prior in the true matrix simulation and further show that the performance of the uniform prior approaches that of the matrix prior as the number of cells increases Plots show simulated true values (x-axis) of the topic-gene matrix against inferred values using LDA (y-axis). Pearson r (r) and Spearman r (sr) are reported for each plot. We compared different numbers of cells in the target dataset (different columns). We compared LDA with a uniform prior (top row) with a matrix prior generated from the true topic-gene matrix (bottom row). The blue dotted line is the line $y = x$.

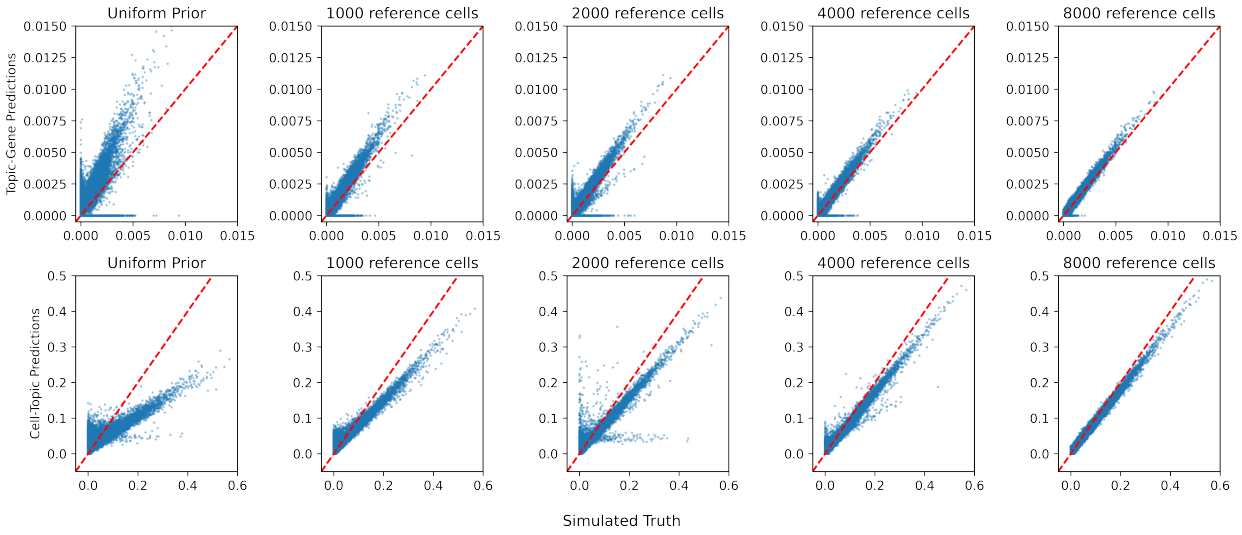

Figure S9: Scatter plots demonstrate the improvement of the matrix prior in both the cell-topic and topic-gene matrices as the number of reference cells increases in the inferred matrix simulation. 1000 simulated cells were analyzed using a uniform prior (left-most column) and a matrix prior. The dotted red line is the $y = x$ line. True simulated values (x-axis) and inferred values (y-axis) are plotted for both the topic-gene matrices (top) cell-topic matrices (bottom).

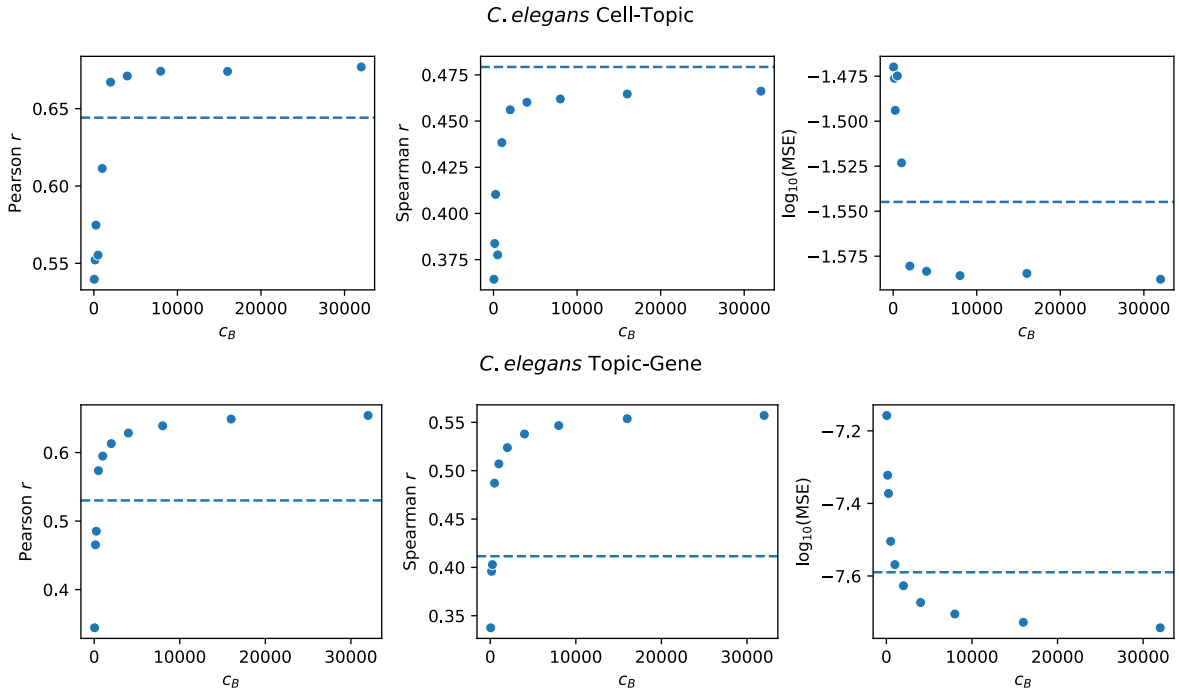

Figure S10: Summary statistics (y-axis) between output matrices of the joint model and LDA with matrix prior, both trained on the *C. elegans* data, show that as the weight of the prior increases, agreement between the matrix prior and joint model also increases. Each plot shows the matrix prior LDA results (points) for increasing values of c_B (x-axis) versus the uniform prior (blue dotted line). The top row of plots shows summary statistics for the cell-topic matrix, and the bottom row of plots shows summary statistics for the topic-gene matrix.

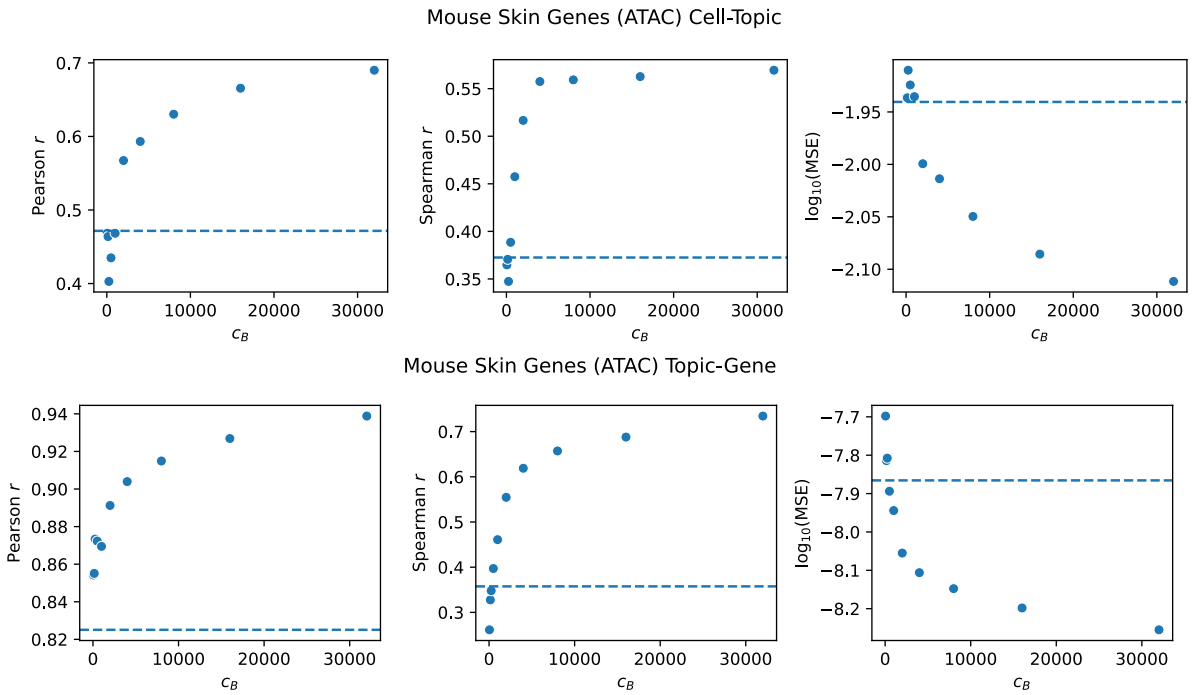

Figure S11: Summary statistics (y-axis) between output matrices of the joint model and LDA with matrix prior, both trained on the SHARE-seq mouse skin scATAC-seq data with cut sites summed over genes (i.e. using the genes vocabulary), show that as the weight of the prior increases, agreement between the matrix prior and joint model also increases. Each plot shows the matrix prior LDA results (points) for increasing values of c_B (x-axis) versus the uniform prior (blue dotted line). The top row of plots shows summary statistics for the cell-topic matrix, and the bottom row of plots shows summary statistics for the topic-gene matrix.

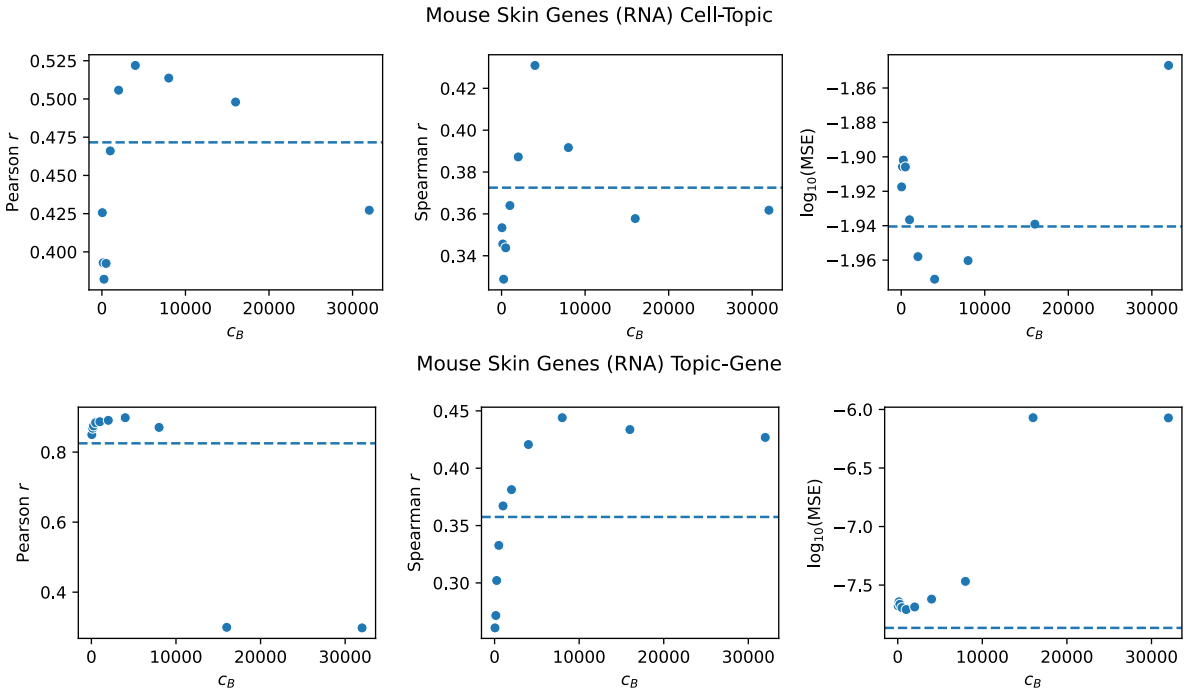

Figure S12: Summary statistics (y-axis) between output matrices of the joint model and LDA with matrix prior, both trained on the SHARE-seq mouse skin scRNA-seq data, show that as the weight of the prior increases, agreement between the matrix prior and joint model also increases with moderate weights and declines with higher weights. Each plot shows the matrix prior LDA results (points) for increasing values of c_B (x-axis) versus the uniform prior (blue dotted line). The top row of plots shows summary statistics for the cell-topic matrix, and the bottom row of plots shows summary statistics for the topic-gene matrix.

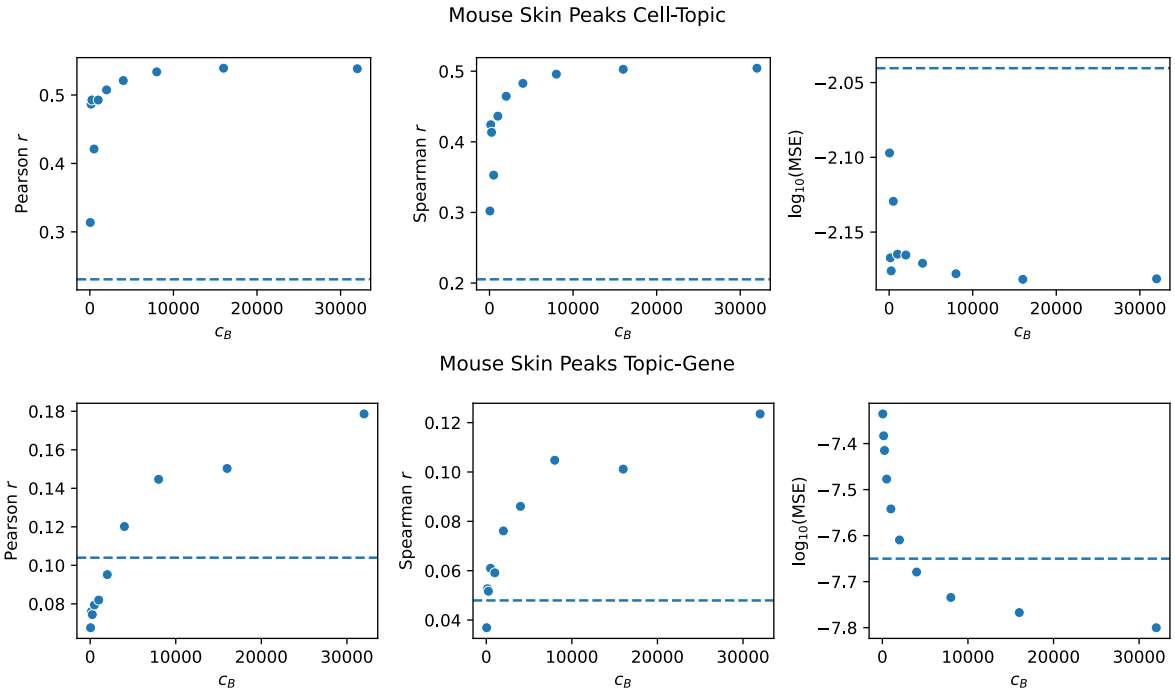

Figure S13: Summary statistics (y-axis) between output matrices of the joint model and LDA with matrix prior, both trained on the SHARE-seq mouse skin scATAC-seq data (i.e. using the peaks vocabulary), show that as the weight of the prior increases, agreement between the matrix prior and joint model also increases. Each plot shows the matrix prior LDA results (points) for increasing values of c_B (x-axis) versus the uniform prior (blue dotted line). The top row of plots shows summary statistics for the cell-topic matrix, and the bottom row of plots shows summary statistics for the topic-gene matrix.

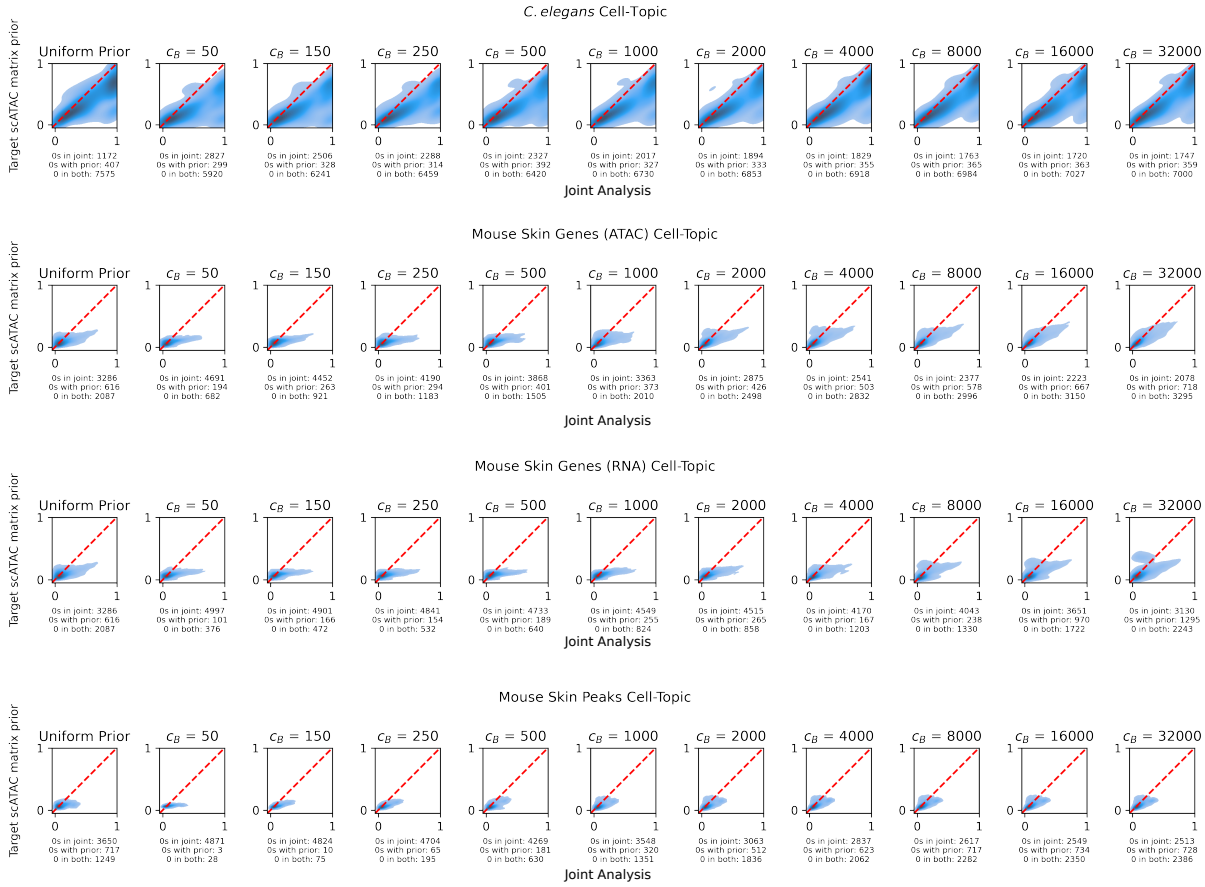

Figure S14: Comparing the cell-topic matrices of the joint model versus the LDA with the matrix prior reveals that as the weight of the matrix prior increases, the agreement between the models increases. The effect of different values of c_B were evaluated by comparing the cell-topic matrix using the matrix prior to the cell-topic matrix from the joint model. Different values of c_B are plotted across different columns, and different datasets are shown in different rows. We first flatten the cell-topic matrices so that they can be plotted. The cell-topic assignments from the joint model are shown on the x-axis, and the inferred cell-topic assignments from the matrix prior LDA are shown on the y-axis. A dotted red line is drawn to indicate the line $y = x$. Zero values are omitted from the plots, but the number of zeros exclusively in the cell-topic matrix of the joint model, exclusively in the cell-topic matrix of the LDA with matrix prior, and number of zeros in both is noted below each plot.

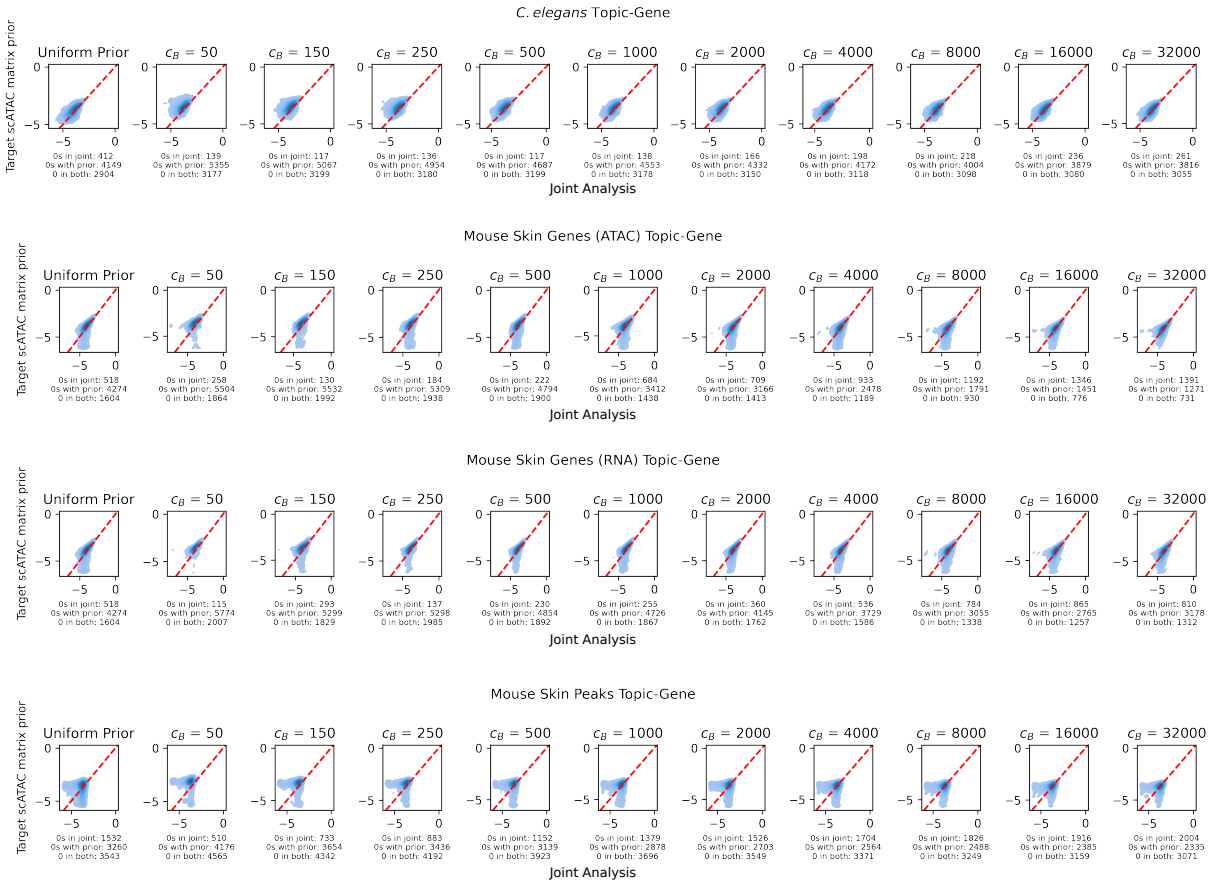

Figure S15: Comparing the topic-gene matrices of the joint model versus the LDA with the matrix prior reveals that as the weight of the prior increases, the agreement between the models increases. The effect of different values of c_B were evaluated by comparing the topic-gene matrix using the matrix prior to the topic-gene matrix from the joint model. Different values of c_B are plotted across different columns, and different datasets are shown in different rows. We first flatten the topic-gene matrices so that they can be plotted. The topic-gene assignments from the joint model are shown on the x-axis, and the inferred topic-gene assignments from the matrix prior LDA are shown on the y-axis. A dotted red line is drawn to indicate the line $y = x$. Zero values are omitted from the plots, but the number of zeros exclusively in the topic-gene matrix of the joint model, exclusively in the topic-gene matrix of the LDA with matrix prior, and number of zeros in both is noted below each plot.

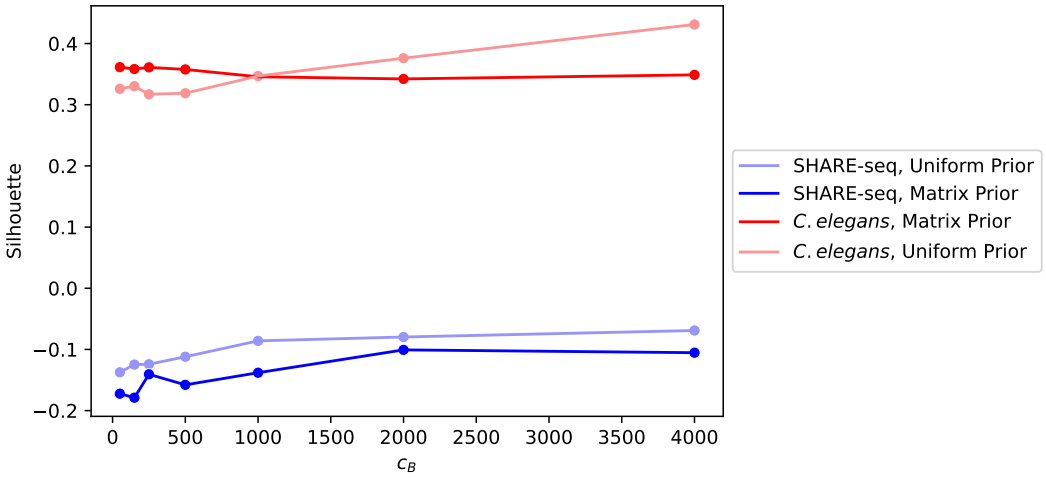

Figure S16: The matrix prior and uniform prior show similar performance as a function of c_B. Average silhouette values for the published cell type annotations (y-axis) are plotted against increasing values of c_B (x-axis). Different colored lines indicate whether the SHARE-seq data set (red) or the *C. elegans* data set (blue) was used. Each data set was analyzed using a uniform prior (lighter colors) and the matrix prior (darker colors)

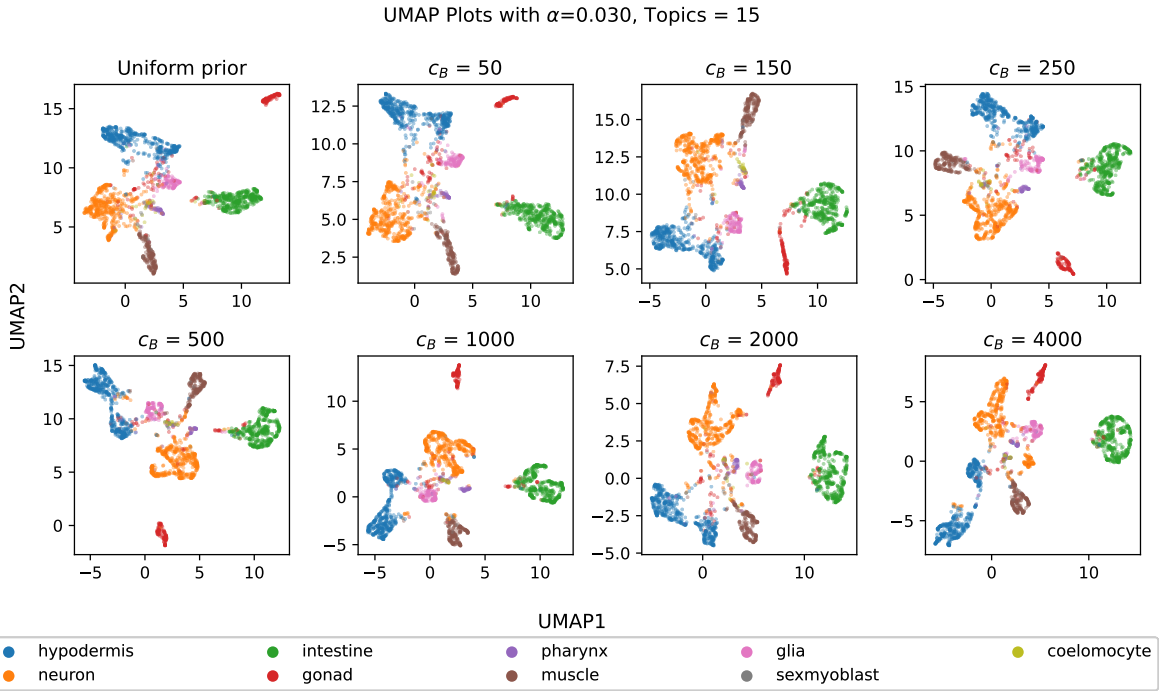

Figure S17: UMAP plot of all the *C. elegans* data reveal cell type structure from LDA analysis with different weights c_B of the matrix prior. Cells are colored based on their published cell type annotations.

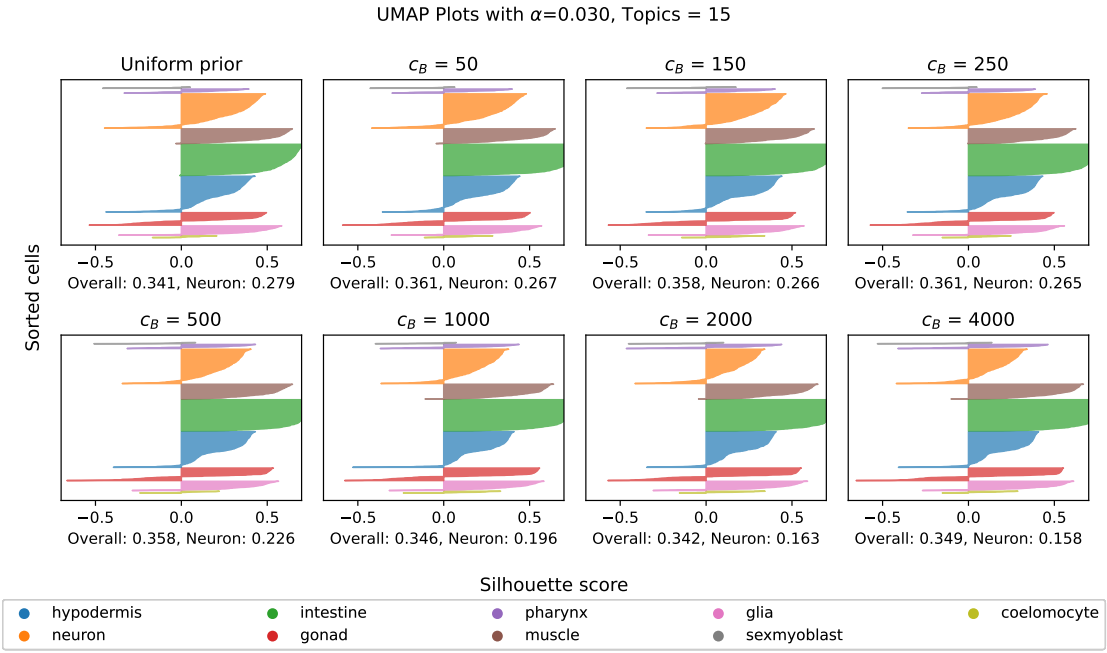

Figure S18: Silhouette plots demonstrate that in *C. elegans*, the silhouette value did not improve with increased weight of the prior. Silhouette values are shown for *C. elegans* cell types plotted for results from a uniform prior LDA model and matrix prior LDA models trained with increasing values of c_B using 15 topics and the scATAC-seq data translated into the genes vocabulary (ATAC cut sites summed over the promoter and gene body for 13,734 genes). Each row in each plot represents one cell, and the silhouette value of the cell is the length of the line. The mean silhouette value for all of the cells is shown as "Overall", and the mean silhouette value for only the neurons is shown as "Neuron."

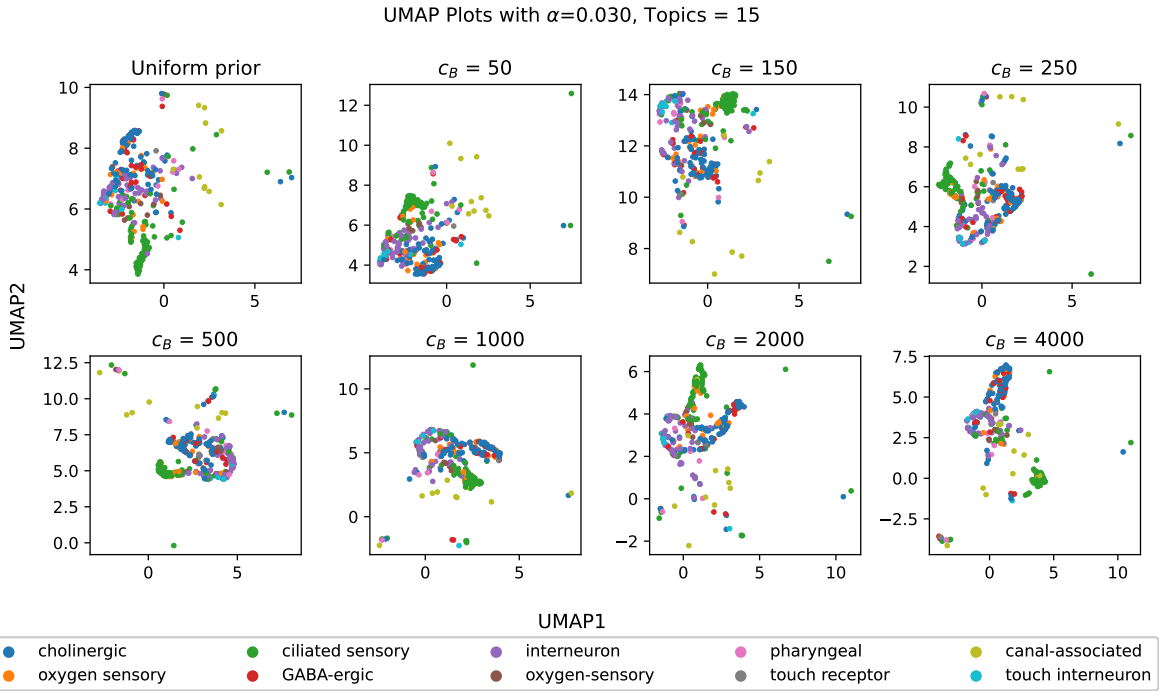

Figure S19: A subset of Figure S17 that includes only neurons, demonstrating that increased values of c_B have little effect on the ability of the matrix prior LDA to distinguish among published cell types. Cells are colored by published neuron subtype labels.

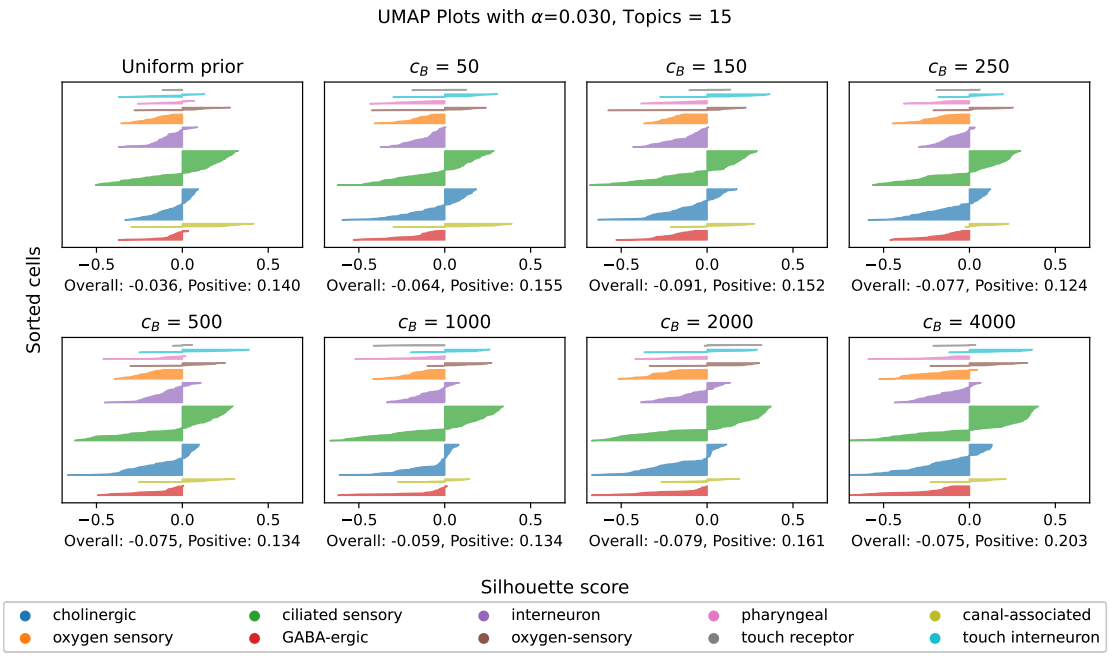

Figure S20: Silhouette plots of the neurons in the *C. elegans* dataset. The overall mean silhouette values and the mean positive silhouette values are reported.

Chapter 4

DATA LEAKAGE IN SEQUENCE-BASED MULTITASK GENOMICS MODELS

This chapter is work done with the following authors:

Alan Min, Surag Nair, Jacob Schreiber, Areeb Gani, Anshul Kundaje, and William Stafford Noble.

4.1 Author Contributions

AK, JS, SN conceptualized the idea of better understanding data leakage in these models. AM ran simulations and conducted theoretical analysis of the problem. AM and SN curated datasets and models to use as examples. AG conducted initial studies of the problem. All authors provided feedback about the experiments. AM wrote the chapter.

4.2 Abstract

A grand challenge in computational biology involves building computational models capable of predicting various types of genomic activity—such as mRNA expression levels, patterns of histone modifications, and regions of chromatin accessibility—solely on the basis of the genomic DNA sequence. Methods such as DeepSEA, Bassett, Basenji, Enformer and BPNet frame this as a multitask learning problem. In this setting, each task involves predicting one type of genomic activity in one particular cell or tissue type, and all of the tasks start from a common input, the DNA sequence. In this work, we demonstrate that this multitask learning setup can lead to inaccurate models, when genomic features that are irrelevant for one task are erroneously assigned significance in a related task. We illustrate the problem using a simple example, via a more sophisticated simulation, and in empirical results from several

published models. We hypothesize that the problem arises because of two reasons. The first is that many computational models use peak selection, the process of selecting regions of high signal to create training sets. We show through simulation and theory that peak selection may lead to biased results. The second reason is that many models use a shared latent structure, which we hypothesize to lead to inaccurate results. We show through simulation that the shared latent structure may lead to biologically implausible results. These problems are particularly problematic in the setting of *in silico* design, where such models are used to design sequence elements to achieve a particular patterns of genome activity.

4.3 Introduction

Numerous sequence-based deep-learning models have been developed to predict, from a given DNA sequence, genome-wide profiles of protein-DNA binding, chromatin accessibility, splicing, long-range chromatin contacts, and gene expression in diverse cellular contexts [Zhou and Troyanskaya, 2015, Avsec et al., 2021a,b, Kelley et al., 2018, 2016, Zhou et al., 2018]. These models often use multitask learning, a general framework that shares parameters between various tasks that use the same input Ruder [2017]. Multitask learning is purported to offer improved generalization, data augmentation in cases with limited training data, regularization, and other benefits compared to traditional single task learning. In the case of sequence-based models, the multiple tasks are often different genomic tracks of interest. For example, a model may simultaneously model the relationship between sequence and several types of chromatin profiling data such as ATAC-seq [Buenrostro et al., 2015], DNase-seq [Song and Crawford, 2010], and CHiP-seq [Robertson et al., 2007]. The goal of these large sequence models is to improve the quality of prediction for each of the different tasks simultaneously through the inclusion of the other tasks.

Because these data are often sparse, and because the inclusion of many tasks can cause computation to be expensive, a common practice in training large sequence-based models is to oversample regions with high signal in order to train efficiently. Often, this process involves a "peak calling" step, which identifies the regions with high signal. Many methods

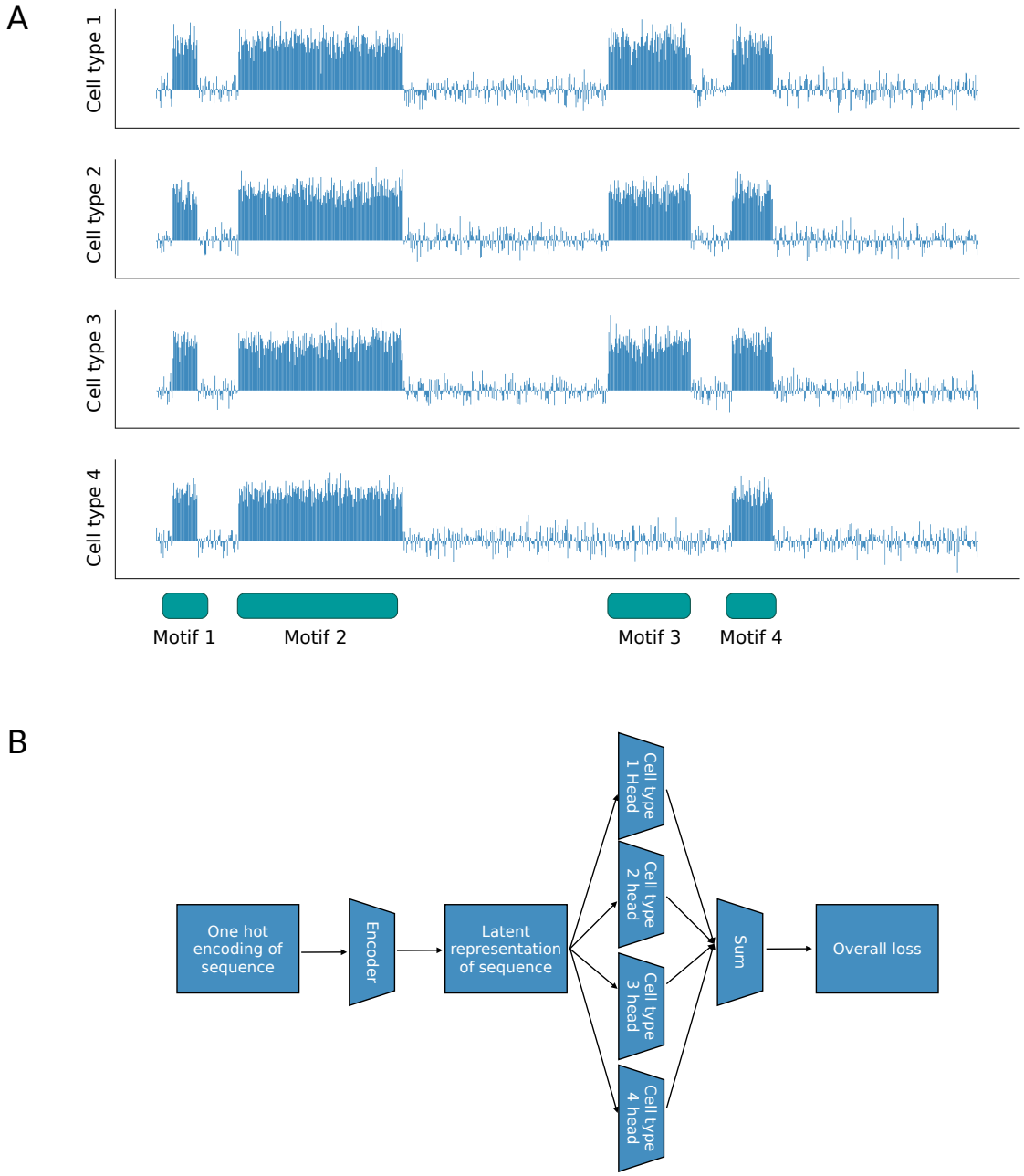

Figure 4.1: (A) Simulated example of DNase-seq data data across 4 cell types and 4 motifs. In this simulated example, cell types 1-3 are related, each having peaks for each of motifs 1-4. However, cell type 4 does not show binding for motif 3. For simplicity, we assume that the expression is associated with motifs directly on top of the motifs, although this may not be the case in biological data. (B) Architecture of stereotypical multitask learning setup.

have been developed for peak calling in CHiP-seq, DNase-seq, and ATAC-seq, among others [Zhang et al., 2008, Tarbell and Liu, 2019, Mortazavi et al., 2008, Qin et al., 2010, Fejes et al., 2008]. Peak calling methods use hidden Markov models, window-based methods, or more sophisticated methods, but at a basic level all of these methods find regions of the genome for which (smoothed) signal is greater than some given threshold [Kruczyk et al., 2013]. It is common practice then to either undersample background regions of the genome or to oversample peak regions in an effort to both limit computation time and improve the accuracy of the model [Yoon and Kwek, 2005]. In a multitask setting, furthermore, peak regions may be determined jointly across each of the tasks.

These standard approaches to sequence-based prediction have shown promising results, but there has been some doubt regarding whether state-of-the-art deep learning models are capturing important causal signals in data [Karollus et al., 2023]. In this work, we further investigate the performance of large sequence-based models in a multitask setting. Specifically, we present simulations and theoretical results which demonstrate that multitask learning can lead to bias in sequence-based models.

To illustrate some of the possible sources of the bias, we created a simple simulation (Figure 4.1). We consider a scenario in which the goal is to predict a genomic track, for example, gene expression or transcription factor binding, based on the presence or absence of a sequence motif, for example, DNase-seq signal (Figure 4.1A). For the remainder of our examples in this paper, we assume that the genomic data of interest is chromatin accessibility, analyzed through the use of DNase-seq, but similar arguments could be made for other types of genomic data. In our example, we simulate four cell types, which we label C1–C4, and four motifs, which we label M1–M4. In three of the cell types (C1–3), all four motifs (M1–4) are associated with increased chromatin accessibility. In contrast, C4 is a distinct cell type for which M3 is not associated with chromatin accessibility. The issue arises when these cell types are analyzed jointly. The joint analysis can lead to bias through either (1) oversampling of particular regions of the genome based on multiple cell types or (2) the joint training of a multitask deep learning model with a shared latent space.

We first summarize how bias arises from oversampling in a multitask learning setting. In large sequence-based genomic models, a common practice is to select windows of sequence associated with peaks to balance the data. These windows can be selected based on various strategies. For example, windows may be selected based on regions that exhibit TF binding [Zhou and Troyanskaya, 2015], by oversampling positive regions [Alipanahi et al., 2015], or based on peak regions [Kelley et al., 2016]. This selection of sequence windows to use for training is often conducted using data for all available cell types jointly. In other words, if a region is selected for one of the cell types, then it is included in the training set which is used to train the model for all of the cell types. The problem is that the distribution of relevant regions may not be equivalent for all cell types, and selecting based on the output of one of the cell types may bias the model toward certain motifs and peaks. For example, in Figure 4.1A, motif 3 is not present in cell type 4 but is present in the remaining cell types. A training regime that includes oversampling and peak selection may introduce bias in the predictions for cell type 4. We examine this oversampling and peak selection issue in detail in Section 4.5.1.

We next summarize the second source of bias from multitasking learning with a shared latent space. A latent space in this context is a hidden layer in the neural network that represents the input sequence in a lower dimension, and may be shared between all of the tasks. In Figure 4.1B, we outline the architecture for a stereotypical multitask learning neural net model that begins by taking a window of sequence and applies a one-hot encoding to it. This is followed by an encoder layer, finally resulting in a latent representation of the sequence. The latent representation is input into separate model "heads," each of which predicts the expression profile for a single cell type. At training time, the loss from each of the individual heads contributes to the training for the parameters within that head and for the parameters in the encoder. This results in a latent representation of the sequence that takes into account all of the tasks. The issue then arises when there are some cell types that are distinctive from the rest of the cell types, i.e. cell type 4 in our example. For this example, the latent representation may be suitable for cell types 1–3, which respond to motif

3. However, this representation could lead the model for cell type 4 to assign an inaccurate association to motif 3.

Through the rest of this chapter, we examine both of these issues through simulation using simplified examples, followed by analysis with real data and published models.

4.4 Methods

4.4.1 Peak selection simulation

We first describe our approach to simulating peak selection (Section 4.5.2). We simulate X, the hypothetical sequence similarity between a hypothetical sequence (Figure 4.2A) and a hypothetical motif in our first simulation. Note that we refer to collections of these sequence similarities and abstract sequence similarities using uppercase X and use lowercase x when referring to a specific sequence similarity. We do not simulate the sequence or the motifs. Although we do not specify a sequence similarity function, we imagine that this is a function that takes a motif and a sequence as input and outputs a scalar value which is greater for motifs and sequences that are more similar. The sequence similarity between motif and sequence can be thought of as a dimensionality reduction of the raw sequence. We chose to simulate this similarity value directly to avoid simulating sequence and enabling theoretical analyses.

We simulated X for five motifs and for two cell types, which we call iPSC cells and fibroblast cells. We then simulated DNase-seq signal by using a linear model

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + a_i + \epsilon_i \tag{4.1}$$

$$z_i = \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} + \gamma_5 x_{i5} + b_i + \epsilon'_i \tag{4.2}$$

against X, with $\beta = (10, 1, 0.1, 0.1, 0.1)^T$ for the iPSC cells and $\gamma = (0, 1, 0.1, 0.1, 0.1)$ for the fibroblast cells. The linear model also includes what we call a "latent chromatin state" a and b for iPSC and fibroblast cells, respectively. The chromatin states were simulated with $\begin{pmatrix} a \\ b \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix}\right)$, with $\rho = 0.5$ and $\sigma_a = \sigma_b = 3$. We simulated 1000

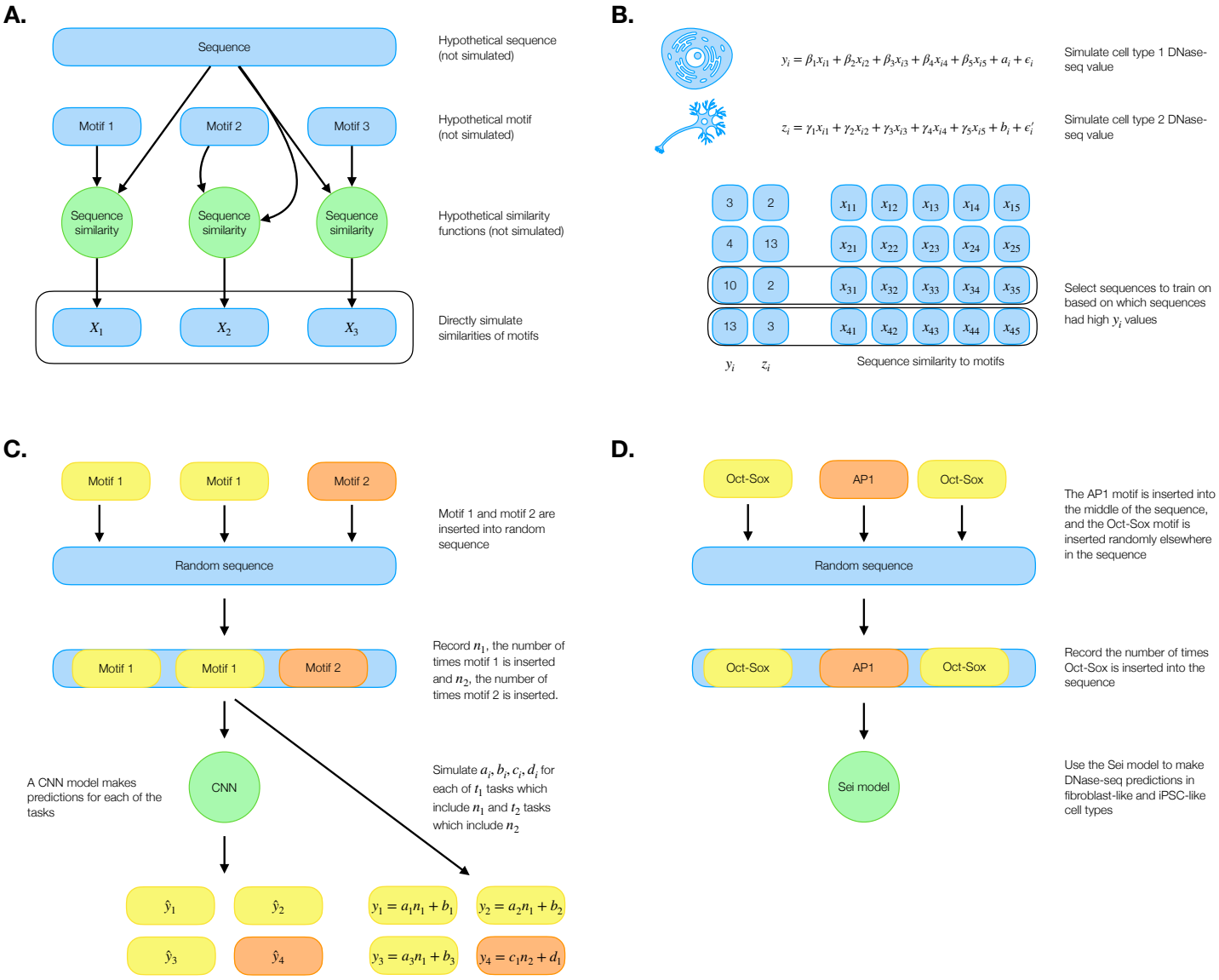

Figure 4.2: (A) Explanation of the simulated motif similarities. The sequences, motifs, and sequence-similarity functions are not simulated. We instead directly simulate values that would have resulted from computing sequence similarity between motif and sequence. (B) The peak selection process and DNase-seq data simulation is described. We select peaks based off of only y_i, but we are interested in the γ_i coefficients. (C) The simulation process for inserting two motifs into random sequence. We are interested in what happens when t_1 and t_2 vary. (D) Schematic of the analysis of the Sei model, where the AP1 motif is inserted into the center of the random sequence, and the Oct-Sox motif at other locations.

sequences. We included an error term simulated from a standard Normal distribution. Peaks were selected by choosing the top 100 highest DNase-seq values and choosing the associated sequences. We estimated β and γ using the lm package in R. Analyses and simulation were done in R version 4.2.3. We repeated the simulation 1000 times and averaged the resulting coefficient estimates.

For the second, more systematic simulations (Figure 4.5), we first simulated $x_i \sim N(0,1)$, then simulated a and b the same as earlier in this section with $\sigma_a = \sigma_b = 1$. We simulated $\epsilon, \epsilon' \sim N(0,1)$. Then, finally we simulated the DNase-seq signal using

$$y_i = \beta x_i + a_i + \epsilon_i \tag{4.3}$$

$$z_i = \gamma x_i + b_i + \epsilon_i' \tag{4.4}$$

where y_i is the signal for the first cell type and z_i is the signal for the second celltype. Note that x_i is shared between both y_i and z_i. We note that although X is meant to represent the similarity of a sequence to a motif, we simulated it using a Normal distribution to enable theoretical analyses when conditioning on peak selection. We then computed y and z using the simulated values of x, a, b, ϵ, and ϵ' with $\beta = \gamma = 1$. We consider ρ from -1 to 1 through 100 equally sized steps, and we simulated x 10,000 times. Let X be the set of all x_i that we simulated. In the peak selection step, we construct a subset $X_p = \{x_i : y_i \text{ is in the top } p \text{ highest values of } y\}$ consisting of the x associated with the highest values of y. We considered the top 500 to top 10,000 sequences in increments of 500 (Figure 4.2B). That is, when selecting the top s sequences, we find a y_0 such that only s sequences have corresponding $y_i > y_0$ We estimate $\hat{\gamma}$ using X_P as the sole covariate for y where $y_i > y_0$ in a simple linear regression model both with no intercept term (Figure 4.5A) and with an intercept term (Figure 4.5B). In both approaches, we recorded the estimated $\hat{\gamma}$.

4.4.2 Predicting linear functions of motif counts

For our task of predicting linear functions of motif counts, the goal was to simulate a sequence with n_1 copies of Motif-1 (CATCATCATCAT) inserted, and n_2 copies of Motif-2

(GCCGCCGCCGCC) inserted. We then aimed to train a CNN to learn linear functions of n_1 and n_2, namely $a_i n_1 + b_i$ or $c_i n_2 + d_i$, where in this section a_i and b_i are not referring the chromatin state as in Section 4.4.1. The CNN is to learn t_1 tasks of the form $a_i n_1 + b_i$ and t_2 tasks of the form $c_i n_2 + d_i$. The goal was to represent a scenario where the CNN has tasks to learn based on two unrelated motifs, and then vary t_1 and t_2 (Figure 4.2C).

For the simulation of the sequences, and training task, we used both packaged software and custom functions. We used the simDNA package (https://github.com/kundajelab/simdna) to simulate the sequences. For simplicity, we opted to use a zero-order background sequence. That is, each position had an equally likely chance of being any base pair. Also for simplicity, we opted to consider only exact motifs, that is, we considered a position weight matrix with weight in only one base per position. SimDNA required a small pseudocount, so we included a small probability for each of the positions of 10^{-11}. We then simulated a training set of 50,000 sequences and a validation set of 10,000 sequences. For each of these sequences, we then varied t_1 from 5, 15, 25, 35, and 45, and we sampled a_i from a Normal distribution with mean 0 and variance 1 and b_i with mean 0 and variance 4. We set $t_2 = 50 - t_1$ and simulated $c_i \sim N(0, 1)$ and $d_i \sim N(0, 4)$ in the same way, except that we always set $c_1 = 1$ and $d_1 = 0$. We simulated all the a_i, b_i, c_i and d_i once, and fixed them for all choices of t_1 and t_2 to maximize comparability.

To analyze the data, we implemented a simple CNN model. The model included three initial convolutional layers with a rectified linear unit activation function and pooling on the third layer. The kernel size was 16, and the stride was 1, with no padding for the model. After the three convolutional layers, we flattened the latent layer and continued with a separate head for each of the multiple tasks, each as a dense layer, resulting in a one-dimensional output. We trained the model using an Adam optimizer [Kingma and Ba, 2014] with initial learning rate 10^{-4}. We used a batch size of 10,000 and trained up to 10,000 epochs. We used Python version 3.7.12, and PyTorch version 1.13.1. We trained the model for 10 different random initializations.

We followed prior work in BPNetLite (`https://github.com/jmschrei/bpnet-lite`) us-

ing DeepLiftShap [Shrikumar et al., 2017] for attribution of the model. Modifications to the default implementation of DeepLiftShap were needed to make attributions for a sequence-based model, because when we replace a base pair, there is an addition of the new base pair but also the subtraction of an old base pair. To specify a background sequence for DeepLiftShap, we opted to use a dinucleotide shuffle of the sequences that we had, again following the BPNetLite example. We conducted the attribution for each of our models on the Motif-2 task.

4.5 Results

4.5.1 Oversampling leads to bias when chromatin states are correlated between cell types

In Section 4.3, we described how oversampling some regions of the genome in a joint setting can potentially lead to bias in a multitask learning setting. Here we will consider simple linear regression for TF binding as an illustrative example of how such a bias can occur. To do so, we introduce a variable representing latent chromatin state, which we assume affects TF binding in addition to the presence or absence of sequence motifs. The intuition for why latent chromatin state may result in biased estimates of the impact of motif size is that when region selection occurs, we may be inadvertently selecting loci that are associated with open chromatin state. In the presence of multiple cell types, when we select regions based on one cell type, we may also see spurious associations from motif to DNase-seq signal if chromatin state is correlated or anticorrelated between the two cell types. Here, we will consider the case of two cell types, which we refer to as iPSC and fibroblast.

We now introduce some notation and assumptions for our analysis. For simplicity, we assume that we observe scores for a list of previously determined motifs. We assume that these scores characterize the strength of signal for a certain motif within the given window,

and we use these scores as a proxy for sequence in our analysis. Specifically, we have

$$\begin{aligned} x_1 &= (x_{11}, \ldots, x_{1R}) \\ &\vdots \\ x_G &= (x_{G1}, \ldots, x_{GR}) \end{aligned}$$

where G is the number of genomic positions and R is the number of genomic motifs of interest. Although in general we could have large R, for our examples we use $R = 1$ for simplicity. Each x_{ij} is the continuous score representing the signal strength of motif j within a window surrounding genomic position i of pre-specified size.

Critically, we assume that each cell type has an associated latent variable representing chromatin state, and that these variables are not independent of one another. Specifically, we assume that a_i and b_i are distributed jointly as

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix} \right)$$

We then assume that the ChIP-seq signals for the iPSC and fibroblast cells are respectively governed by

$$y_i = \beta x_i + a_i + \epsilon_i \tag{4.5}$$

$$z_i = \gamma x_i + b_i + \epsilon_i' \tag{4.6}$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and $\epsilon_i' \sim N(0, \sigma_\epsilon^2)$ are random noise. Note that, for simplicity, we have dropped the intercept term. Finally, we assume that when creating a model we do not have access to a_i or b_i, and hence we fit the model using only our observed motif x_i, which we assume to be fixed and known. Under these assumptions, we can show that training genome-wide results in no bias, even when the latent variables are included (Proposition 1).

Proposition 1. Assuming that the model as defined above (Equations 4.5 and 4.6) is trained genome-wide, then the simple linear regression estimate γ is unbiased.

Proof. The least squares solution to a simple linear regression with one variable is given by $\hat{\gamma} = \frac{n\sum x_i z_i - \sum x_i \sum z_i}{n\sum x_i^2 - (\sum x_i)^2}$. Plugging in Equation 4.6, removing elements with a single ϵ_i because those will be 0 in expectation, and computing, we have

$$E(\hat{\gamma}) = E\left(\frac{n\sum x_i(\gamma x_i + b_i) - \sum x_i \sum(\gamma x_i + b_i)}{n\sum x_i^2 - (\sum x_i)^2}\right) \tag{4.7}$$

$$= E\left(\gamma + \frac{n\sum_i x_i b_i - \sum x_i \sum b_i}{n\sum x_i^2 - (\sum x_i)^2}\right) \tag{4.8}$$

We now take the expectation with respect to b_i, and we have that

$$E(\hat{\gamma}) = \gamma.$$

□

Next we consider the case where the training regions are selected in one cell type but used to train models for both. For example, we assume that the model is trained using genomic regions selected from iPSC cells, and we show that this procedure leads to a bias in the the estimate of γ We believe that this is a meaningful example because in many real world training scenarios, there may be many similar cell types and one distinct cell type. This may result in a scenario where few peaks are ultimately chosen from the distinct cell type.

For simplicity, we model genomic region selection by conditioning the linear regression on the event that $y_i > y_0$, where y_0 is a fixed threshold. While in many real scenarios, more sophisticated calling software may be used [Zhang et al., 2008], this simple approach allows us to calculate the bias term analytically (Proposition 2).

Proposition 2. Assume the model as defined above (Equations 4.5 and 4.6) is trained only in cases where the iPSC peaks $y_i > y_0$, where y_0 is a peak-calling threshold. We assume X is fixed. Then the expectation

$$E(\hat{\gamma} \mid y > y_0) = \gamma + (X_P^T X_P)^{-1} X_P^T \sigma_b \rho z_i$$

where X_P denotes a subset of X including only the rows of X where $y_i > y_0$, and

$$z_i = \frac{\phi(\frac{y_0 - \beta X_{P,i}}{\sqrt{\sigma_a^2 + \sigma_e^2}})}{1 - \Phi(\frac{y_0 - \beta X_{P,i}}{\sqrt{\sigma_a^2 + \sigma_e^2}})}$$

where ϕ and Φ are the probability density function and cumulative density function of a Normal distribution respectively.

Proof. We know that the least squares solution to a linear regression of $z = \gamma X$ is

$$\hat{\gamma} = (X^T X)^{-1} X^T z$$

Recalling that that $z = X\gamma + b + \epsilon'$, we have

$$\begin{aligned} \hat{\gamma} &= (X^T X)^{-1} X^T (X\gamma + b + \epsilon') \\ &= \gamma + (X^T X)^{-1} X^T b + (X^T X)^{-1} X^T \epsilon'. \end{aligned}$$

We condition on the event $y_i > y_0$ and take the conditional expectation. We note that ϵ' has mean 0 and is independent of all other variables, and we have

$$E(\hat{\gamma} \mid y > y_0) = \gamma + (X_P^T X_P)^{-1} X_P^T E(b \mid y > y_0). \tag{4.9}$$

We can calculate that the joint distribution

$$\begin{pmatrix} b_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \beta X_i \end{pmatrix}, \begin{pmatrix} \sigma_b^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_a^2 + \sigma_e^2 \end{pmatrix} \right)$$

by calculating that $E(b_i) = 0$, $E(y_i) = \beta X_i$, and $Cov(b_i, y_i) = E(b_i y_i) - E(b_i)E(y_i) = \rho\sigma_a\sigma_b$. Then from the conditional distribution properties of the Normal distribution, we have that

$$b_i \mid y_i \sim N\left(\sqrt{\frac{\sigma_b^2}{\sigma_a^2 + \sigma_e^2}} \rho (y_i - \beta X_i), (1 - \rho^2)\sigma_b^2 \right). \tag{4.10}$$

For notational simplicity, let $c = \sqrt{\frac{\sigma_b^2}{\sigma_a^2 + \sigma_e^2}} \rho$. From Equation 4.5.1 we need to calculate that

$$E(b_i \mid y_i > y_0)$$

$$= \int_{y_i=y_0}^{\infty} E(b_i \mid y_i = y_0) p(y_i \mid y_i > y_0) dy_i \qquad \text{Conditional expectation definition}$$

(4.11)

$$= \int_{y_i=y_0}^{\infty} \left[c y_i p(y_i \mid y_i > y_0) - c\beta X_i p(y_i \mid y_i > y_0) \right] dy_i \quad \text{Fill in } E(b_i|y_i = y_0) \text{ (Equation 4.10)}$$

(4.12)

$$= c \left(\beta X_i + \frac{\sqrt{\sigma_e^2 + \sigma_a^2}\phi(\frac{y_0 - \beta X_i}{\sqrt{\sigma_a^2 + \sigma_e^2}})}{1 - \Phi(\frac{y_0 - \beta X_i}{\sqrt{\sigma_a^2 + \sigma_e^2}})} \right) - c\beta X_i \qquad \text{See below}$$

(4.13)

$$= c \left(\frac{\sqrt{\sigma_e^2 + \sigma_a^2}\phi(\frac{y_0 - \beta X_i}{\sqrt{\sigma_a^2 + \sigma_e^2}})}{1 - \Phi(\frac{y_0 - \beta X_i}{\sqrt{\sigma_a^2 + \sigma_e^2}})} \right) \qquad \text{Subtraction}$$

(4.14)

$$= \sigma_b \rho \frac{\phi(\frac{y_0 - \beta X_i}{\sqrt{\sigma_a^2 + \sigma_e^2}})}{1 - \Phi(\frac{y_0 - \beta X_i}{\sqrt{\sigma_a^2 + \sigma_e^2}})}. \qquad \text{Definition of } c$$

(4.15)

To move from Equation 4.12 to Equation 4.13 we first note that y_i is distributed as a Normal distribution with mean βX_i and variance $\sigma_a^2 + \sigma_e^2$. We can then find the mean of $y_i \mid y_i > y_0$ using properties of the truncated Normal distribution. □

Although we saw in Proposition 2 that there is a bias if we condition on the signal, we can show that if we are only interested in one motif, then selection of regions based solely on sequence cannot cause a bias in the linear regression case. Assume that

$$z_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{4.16}$$

where $E(\epsilon_i) = 0$. We know that the solution to ordinary least squares is

$$\hat{\beta}_1 = \frac{n \sum x_i z_i - \sum x_i \sum z_i}{n \sum x_i^2 - (\sum x_i)^2}. \tag{4.17}$$

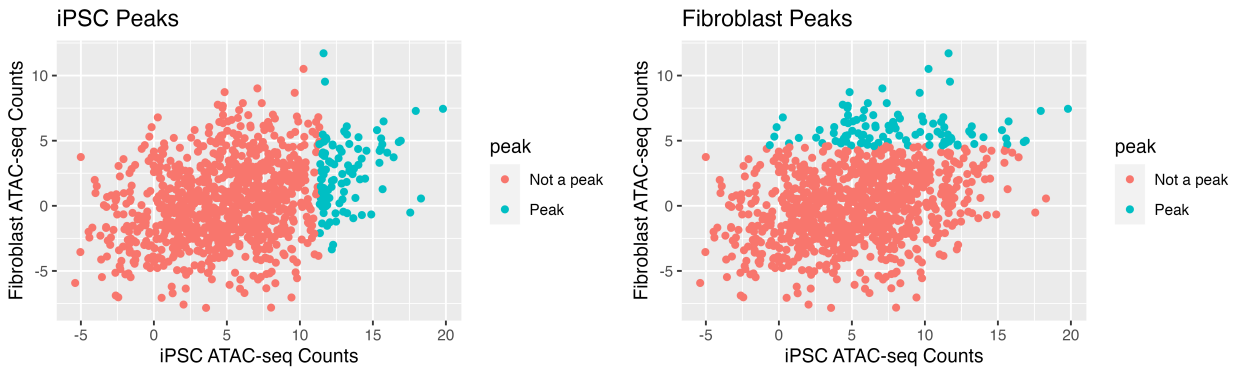

Figure 4.3: Each point represents a sequence within the genome, and the average TF binding value in IPSC cells (x-axis) compared to in fibroblast cells (y-axis) across that sequence. Peaks are selected using TF binding value for IPSC peaks (left) and using fibroblast peaks (right).

We can compute the expectation of $\hat{\beta}_1$ using Equations 4.16 and 4.17 as follows:

$$E(\hat{\beta}_1) = E\left(\frac{n\sum x_i(\beta_0+\beta_1 x_i+\epsilon_i) - \sum x_i \sum(\beta_0+\beta_1 x_i+\epsilon_i)}{n\sum x_i^2 - (\sum x_i)^2}\middle| x, \beta_0, \beta_1\right) = \frac{\beta_1(n\sum x_i^2 - (\sum x_i)^2)}{n\sum x_i^2 - (\sum x_i)^2}$$

This formula shows that, regardless of the choice of x_i, if the linear model is true and $E(\epsilon_i) = 0$, then $E(\hat{\beta}_1) = \beta_1$.

4.5.2 Simulation of linear regression with hidden chromatin state

Although we have successfully characterized the theoretical bias associated with peak selection (Proposition 2), we still aim to better understand the bias term through simulation. Accordingly, using a simulation that instantiates the bias from peak selection, we show that peak selection can impact the estimates in a simple linear regression model.

We simulated two cell types, iPSC and fibroblast cells, with the goal of predicting DNase-seq signal across the genome based on sequence. We then simulated a similarity value from each sequence to 5 motifs, resulting in a value from 0 to 1 for each motif. We assumed that the first motif contributes strongly to DNase-seq signal for the iPSC cells but contributes

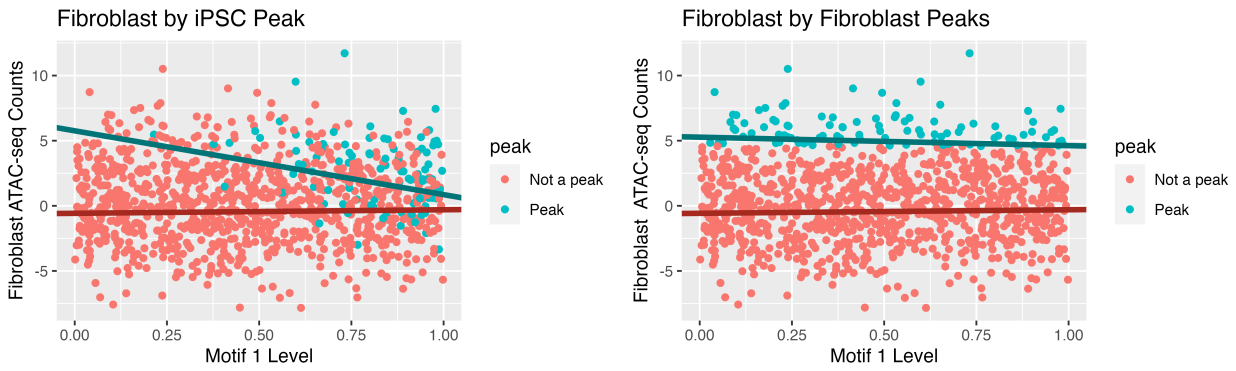

Figure 4.4: Each point represents a sequence within the genome. The level of motif 1 (x-axis) is plotted against the level of TF binding (y-axis), and sequences are colored by whether they were selected as a peak region based on iPSC TF binding (left) and fibroblast TF binding (right). The linear regression line based on only peak regions is shown (blue) and compared to the linear regression based on all simulated sequences (red).

nothing to fibroblast cells. The remaining motifs are assumed to contribute weakly to DNase-seq signal in both cell types. We then simulated the observed DNase-seq signal using the motif values for each cell with a linear model. Peaks are then selected based on their DNase-seq signal in either IPSC cells or fibroblast cells (Figure 4.3). We simulated 1000 sequences and selected the top 100. The sequences selected as peaks have some overlap but are largely different sequences in the different cell types. In addition to the motif determining the DNase-seq signal, we assumed that the DNase-seq signal is also impacted by chromatin state; i.e., DNase-seq signal is higher in open chromatin. In our simulation, we assumed that chromatin state was correlated between iPSC and fibroblast cells. This correlation is what causes the bias in our simulation because it connects the DNase-seq values from iPSC to fibroblast cells.

Next, we fit a linear regression model to identify the effects of each of the motifs on DNase-seq signal in our simulated example. We plotted the simulated DNase-seq level at each locus for fibroblast cells against the simulated sequence similarity of the first motif, which in fibroblast is simulated to have no effect. We then apply the peak selection process

on either the peaks from iPSC cells (Figure 4.4, left) or from fibroblast cells (Figure 4.4, right). Finally, we conducted linear regression based off of either only the peak data for all 5 motifs (blue lines) or all sequences (red lines). When we used only the peak data, we found that the linear regression erroneously showed a negative slope when the peaks were chosen from the iPSC cells. In contrast, the slope appeared to be approximately correct when we trained on fibroblast peaks, although it may not be generally true that the slopes are accurately estimated when using fibroblast peaks.

Model	Intercept	β_1	β_2	β_3	β_4	β_5
Fibroblast Peaks	6.08	0.01	0.18	0.01	0.04	0.02
iPSC Peaks	5.05	-3.58	0.65	0.08	0.08	0.08
Whole Genome	0.01	-0.01	0.99	0.11	0.11	0.09
Truth	NA	0	1	0.1	0.1	0.1

Table 4.5: Regression coefficients for fibroblast cells across 1000 replicate simulation experiments. The average estimated regression coefficient across the replicate experiments for each motif is shown, as well as the estimated intercept. The simulated data was simulated without an intercept, however the estimated models included an intercept to add some model misspecification. We considered the case when peak selection was done with the fibroblast data (Fibroblast Peak), with iPSC cells (iPSC Peak), and with all sequences (Whole Genome). The true values of the parameters are also shown (Truth).

We evaluated this effect more systematically by repeating the simulation process 1000 times, with the same true regression coefficients (Table 4.5). We found that the estimated regression coefficient for the first motif was almost -4 when we evaluated the fibroblast cells using peaks selected from iPSC cells, even though the true value was 0. This demonstrates that the effect size estimates based on selected peaks can be quite different from the true values. We recovered the correct regression coefficients when training genome-wide, as ex-

pected.

In the previous sections, we explored a simulation with a fixed correlation and selection criteria. We will now present a simulation that considers different correlation levels as well as different levels of selection. In brief, for two cell types, we first simulated x_i, the similarity of a motif to a sequence (Figure 4.2A). We then simulated a_i, b_i, ϵ_i, and ϵ_i' and simulated gene expression data, choosing $\beta = \gamma = 1$ (Equations 4.6 and 4.5). For the sake of this section, we assume that the two cell types are iPSC, associated with Equation 4.5, and fibroblast, associated with Equation 4.6. We selected peaks based on the value of y, the value of the DNase-seq data for the iPSC cells. We then used simple linear regression to estimate γ, and report the estimated $\hat{\gamma}$ for different values of ρ for different levels of selected peaks. We considered the linear regression both with and without an intercept term. More details regarding the simulation are available in Section 4.4.1.

We make several key observations based on this simulation. We find that when peak selection was not conducted, i.e. when the number of selected peaks was equal to 10,000, the number of total peaks, we observe no bias in the estimates of $\hat{\gamma}$. This is evidenced by the observation in Figures 4.5A,B that the magenta points, corresponding to no peak selection are on the horizontal line at $\hat{\gamma} = 1$, the true value of γ. These observations suggest that in this simulated setting, training the model genome-wide across all peaks results in no bias, regardless of the correlation between chromatin state of the two cell types. Next, we observe that when the correlation between the chromatin state ρ is 0, there is again no bias. In our example, this means that if there is no correlation between the latent chromatin state of iPSC and fibroblast cells, and peak selection is done using the iPSC data, then there will be no bias for the fibroblast cells. We further observe that as the number of selected peaks decreases, i.e. the peak selection becomes more stringent or y_0 increases, then the bias becomes more apparent for larger values of ρ. Intuitively, as we select the most extreme examples of peaks, then we exhibit a larger and larger bias. Finally, we note that the inclusion of an intercept term appears to change the direction of the bias. This is because the inclusion of the intercept constrains γ.

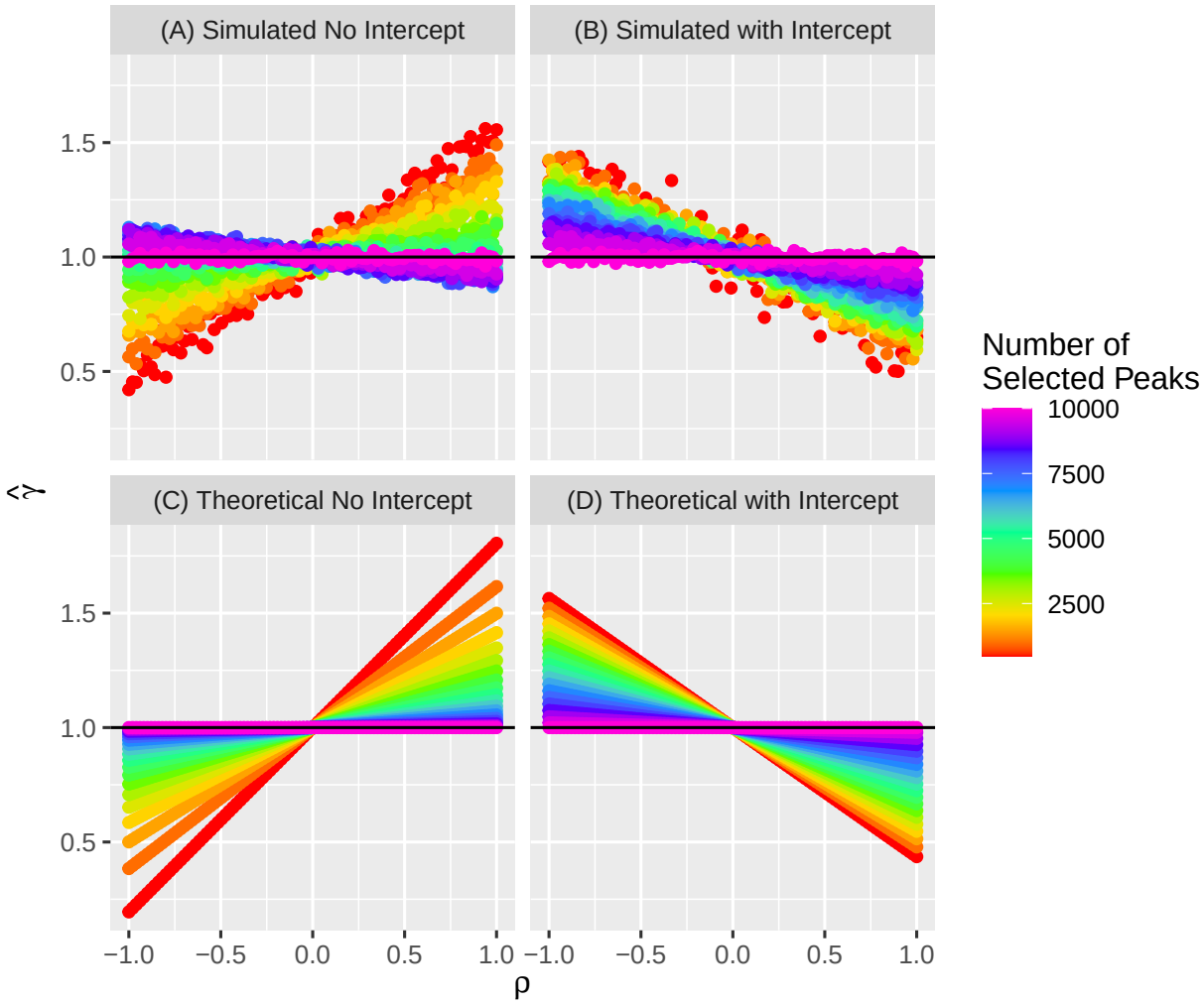

Figure 4.5: Simulated estimates of $\hat{\gamma}$ (y-axis) are plotted against values of ρ (x-axis) for various numbers of selected peaks (colors). Data are simulated, and a linear model is used either (A) without an intercept term or (B) with an intercept term to estimate $\hat{\gamma}$. Theoretical expected values of $\hat{\gamma}$ (Proposition 2) are plotted either (C) without an intercept term of (D) with an intercept term.

We then use Proposition 2 to compute the expectation of $\hat{\gamma}$ given x both without an intercept term (Figure 4.5C) and with an intercept term (Figure 4.5D). We found that the theoretical results matched largely with the simulation results. Together, these results demonstrate that bias can occur for γ under the simulation conditions.

4.5.3 A shared embedding layer leads to inaccurate models

In addition to peak selection, we have empirically observed that multitask training itself can lead to biases when there are a large number of related tasks, and a few additional unrelated tasks. This can be the case when training on multiple related cell types, and one distinct cell type. In the running example throughout this chapter, we have described a case where a sequence-based model is trained for iPSC cells and fibroblast cells. In this situation, there are many cell types that are input into the model, many of which are related to iPSC cells and there is a motif that is unique to these cells. On the other hand, this motif is not associated with the fibroblast cells.

We investigated this phenomenon in a simulation study. We first simulated completely random sequences, and then inserted two different motifs into the sequences (Motif 1 and Motif 2). The task was to predict linear functions of the number of times that Motif 1 and Motif 2 were inserted into the sequence (Figure 4.2C). We then varied the number of Motif-1 tasks t_1 and let the number of Motif-2 tasks t_2 be $t_2 = 50 - t_1$ so that there were always a total of 50 tasks. One of the Motif-2 tasks was always to predict n_2. We used a convolutional neural network (CNN) model with a latent layer, followed by different heads to predict each of the tasks. We trained 10 separate models with different initializations and averaged the results from the 10 models. More implementational details are in Section 4.4.2.

We found that as the number of Motif-1 tasks increased, training began to take longer for the Motif-2 task (Figure 4.6) because the validation MSE was higher for the same number of training epochs. This was apparent at both the short-training intervals, and at the longer training intervals. It appeared that prior to 3000 training epochs, there was a great difference between each of the validation curves, however even at later training epochs, upon

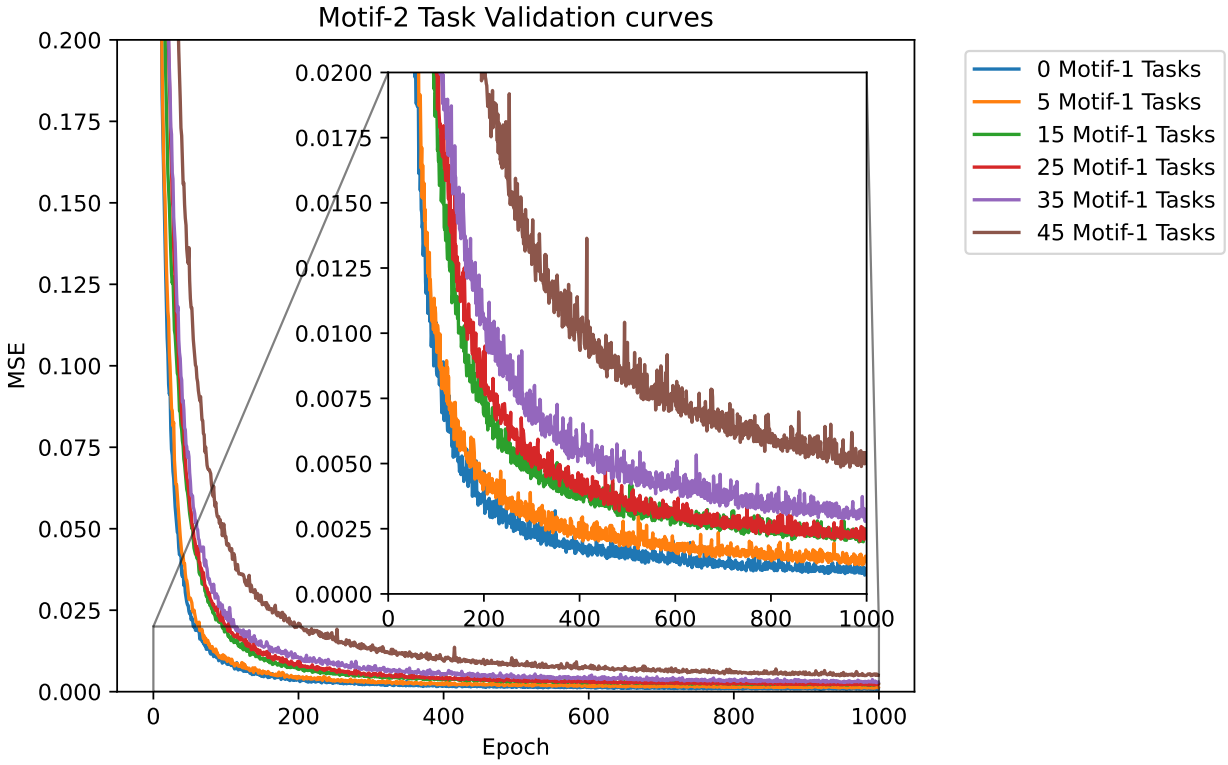

Figure 4.6: Training curves demonstrate that as the number of Motif-1 tasks increases, the training time for the Motif-2 tasks takes longer. The number of epochs (X-axis) is plotted against the validation MSE for the Motif-2 task (Y-axis), and the number of Motif-1 tasks is shown with the color of the curve. The inset plot shows a zoomed-in version of the outer plot.

zooming in, the same trend applies. We reasoned that as the number of Motif-1 tasks increased, the loss function focused more on capturing Motif-1 in the shared embedding space between, and hence capturing Motif-2 tasks became slower and slower. We theorize that the validation MSE does not stop decreasing in this example because we did not add noise to the experimental setup.

In addition to examining the validation curves, we also investigated the use of feature attributions in multitask models. We used the same tasks of predicting a linear combination of the number of Motif-1 and Motif-2 occurrences, and again considered the attributions for the head of the multitask model that was responsible for predicting the number of Motif-2 occurrences. We took the same models that we trained previously and used DeepLift to compute the feature attributions for the different bases in the sequence [Shrikumar et al., 2017]. We located a subsequence that contained both Motif-1 and Motif-2 and used the subsequence as a case study to understand how the number of Motif-1 tasks might influence the feature attributions when considering the Motif-2 head (Figure 4.7). In this example, we saw that the attributions for Motif-2 (GCCGCCGCCGCC) were greater than the attributions for Motif-1 (CATCATCATCAT), which indicates that the feature attribution method was able to capture some meaningful signal, since Motif-2 was the task that was supposed to be attributed for. Furthermore, there was evidence that DeepLiftSHAP correctly ignored background signal, as there was almost no visible attribution signal for the background sequence. This experiment also showed, however, that there was an inaccurate detection of Motif-1, since the attributions were higher than background sequence. Furthermore, the attribution assigned to Motif-1 increased as the number of Motif-1 tasks increased, which we speculate may be because the model gradually puts more weight on Motif-1 when there are more Motif-1 tasks.

We further examined the feature attributions more systematically by locating all instances of Motif-2 and Motif-1, and compared the attributions for these motifs to background random sequence (Figure 4.8). After locating the beginning of each instance of Motif-2 and Motif-1, we aggregated the attributions across motif position. We did this for different choices

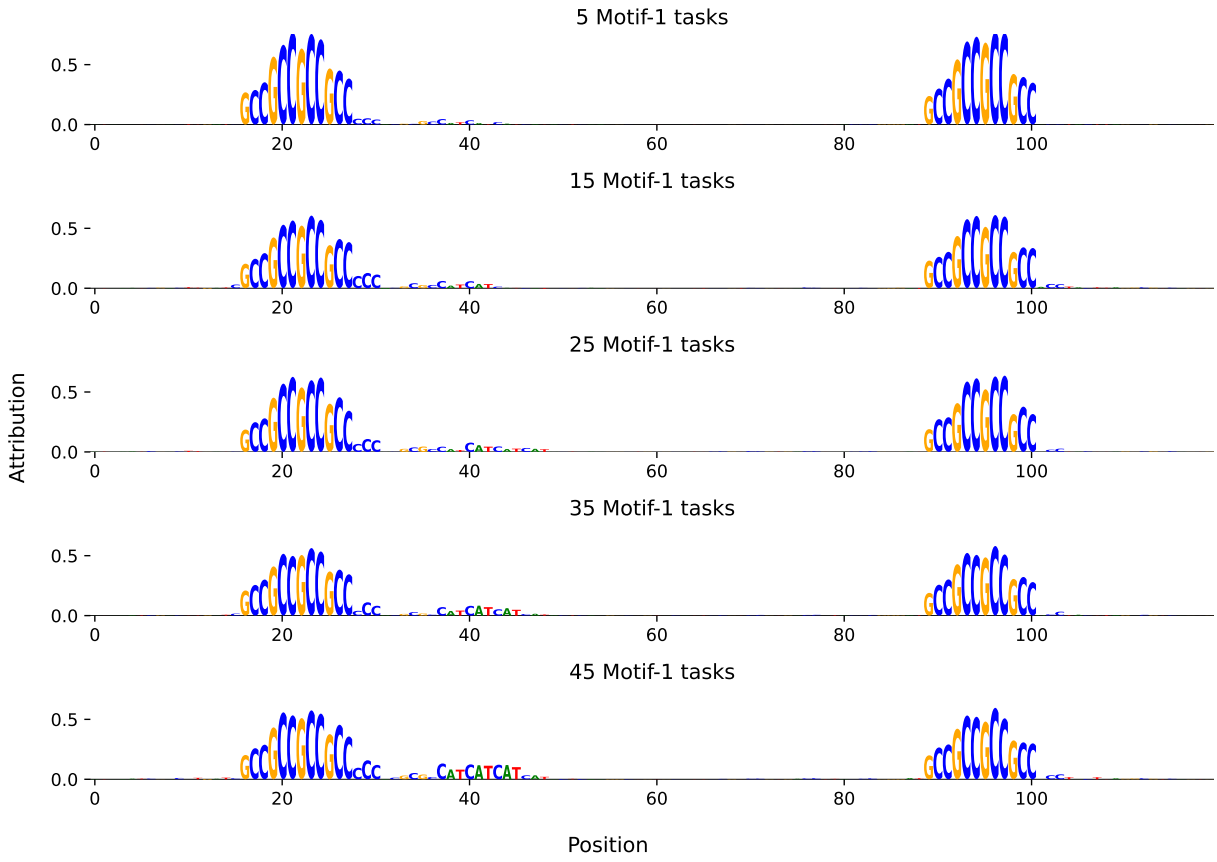

Figure 4.7: As the number of Motif-1 tasks increases, the feature attributions to Motif-1 increase even when attributing the Motif-2 head, whose goal was to predict n_2. The attributions for a representative subsequence is shown. DeepSHAP values (Y-axis) are plotted against the position of the sequence that was sampled (X-axis) Motif-2 (GCCGCCGCCGCC) appears to have consistent attributions, but Motif-1 (CATCATCATCAT) appears to have increasing attributions as the number of Motif-1 tasks increases.

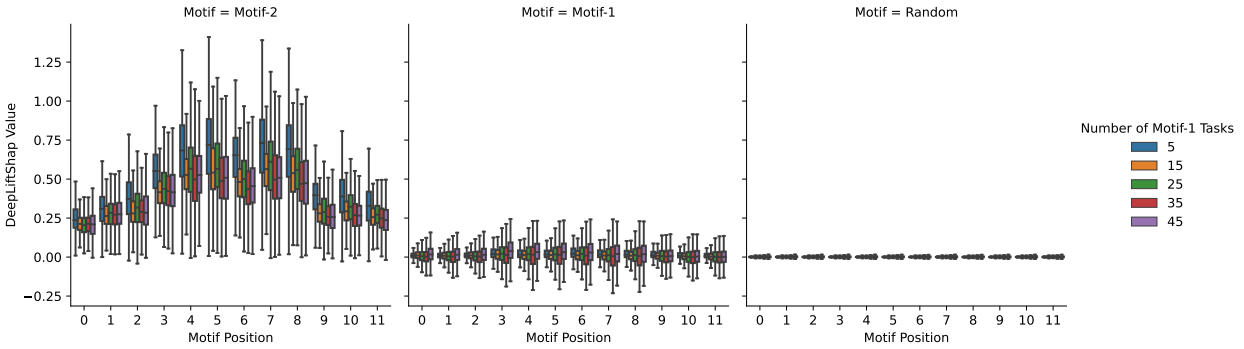

Figure 4.8: We systematically investigated the attributions from DeepLiftSHAP on the Motif-2 task for the Motif-2 sequence (left), the Motif-1 sequence (center), and random positions (right). We plotted boxplots of the DeepLiftSHAP values (Y-axis) against the position in the motif (X-axis) for different sequences. The color represents the number of Motif-1 tasks that were used in the training.

of the number of Motif-1 tasks, and also across the different initializations of our models. These steps were done to avoid artifacts of the attributions due to model initialization or randomness. We found that Motif-2 had the greatest DeepLiftSHAP values, with a peak towards the middle of the motif, consistent with our simple example (Figure 4.7). We also noted that there appeared to be a slight downward trend in the attributions for Motif-2 as the number of Motif-1 tasks increased. This may be because the models are attributing more towards Motif-1 as its weight increases. We further found that the Motif-1 attributions were more variable than those of random background noise, indicating that the model spuriously increases the Motif-1 attributions. This variance increased as the number of Motif-1 tasks increased, again evidencing the bias resulting from having many Motif-1 tasks. Finally, the low attribution values of the random sequence compared to both Motif-2 and Motif-1 serve as a negative control.

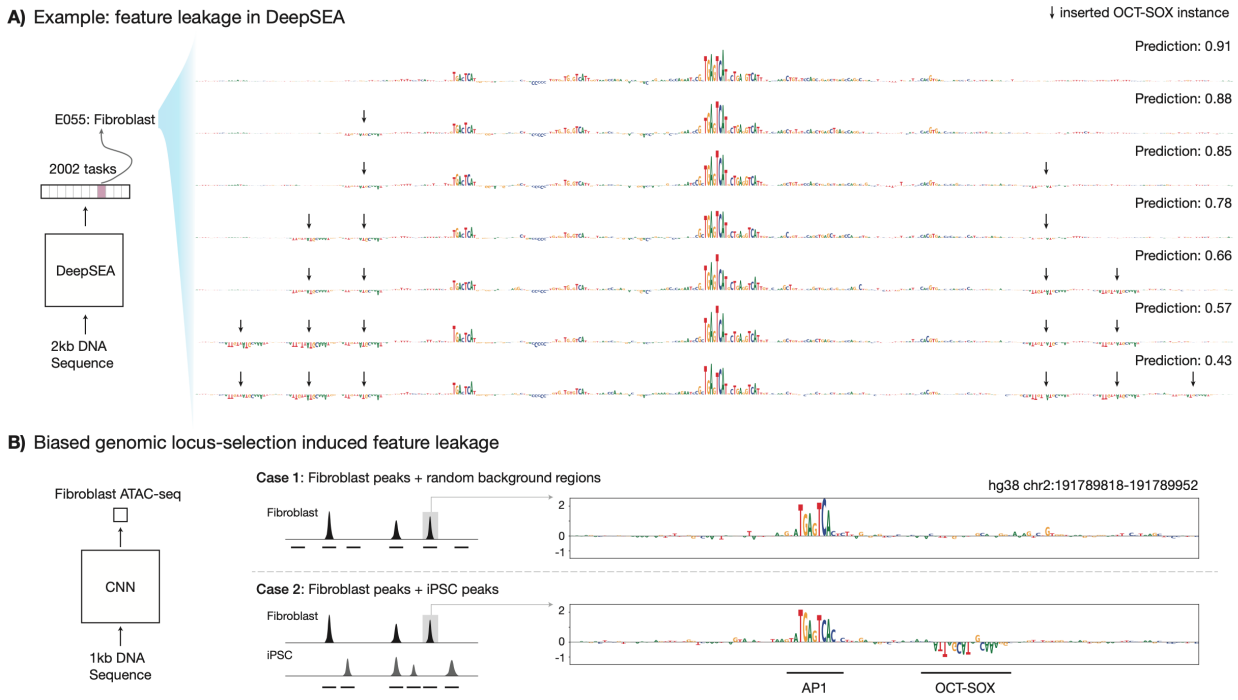

Figure 4.9: (A) Example of feature leakage in DeepSEA: gradually adding instances of Oct-Sox motifs at a genomic fibroblast enhancer, without disturbing existing predictive motifs, reduces the predicted probability of observing a peak for the Roadmap Epigenomics E055 fibroblast task. (B) Training a single-task model to predict fibroblast ATAC-seq signal on peaks from both fibroblasts and iPSCs results in learning the Oct-Sox motif, which is an iPSC-specific TF inactive in fibroblasts. In contrast, training the model with random background regions does not result in learning the Oct-Sox motif.

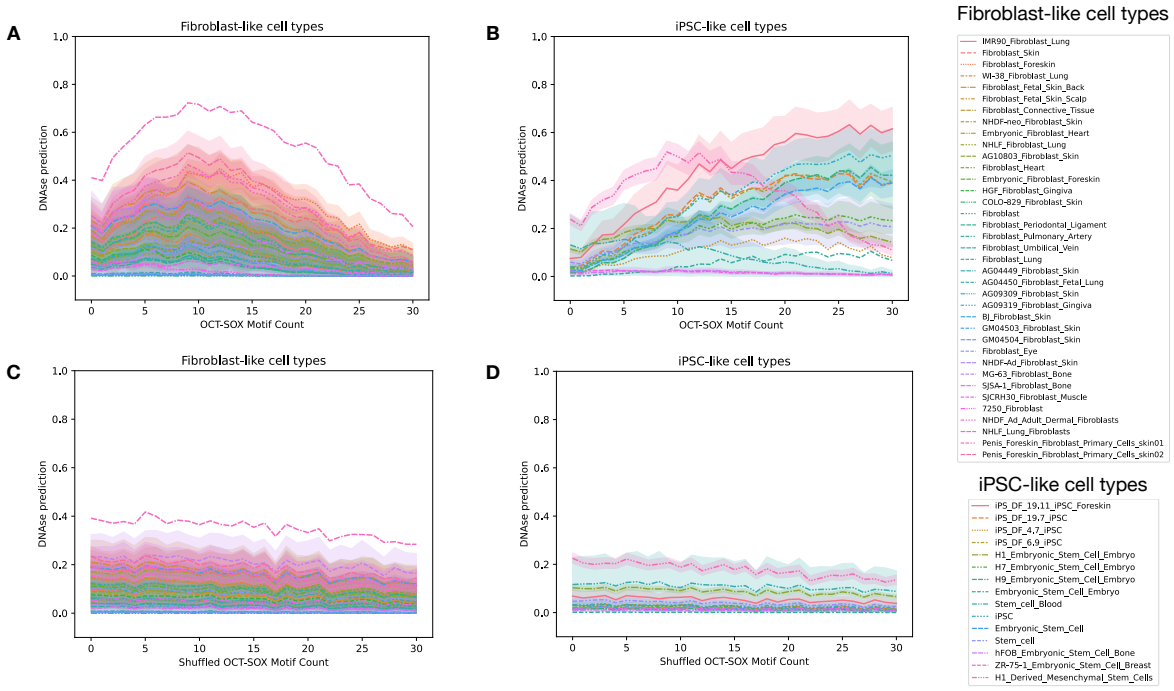

Figure 4.10: Oct-Sox motif affects fibroblast signal compared to shuffled motif baseline. (A) Fibroblast DNase predictions with Oct-Sox motif spiked in (B) iPSC DNase predictions with Oct-Sox motif spiked in (C) Fibroblast DNase predictions with shuffled Oct-Sox motif spiked in (D) iPSC DNase predictions with shuffled Oct-Sox motif spiked in

4.5.4 Examples in published models

Beyond simulated examples, we turn our attention to published models which may also exhibit biologically implausible feature attributions. For example, the DeepSEA model makes the biologically implausible prediction that insertion of instances of the Oct-Sox motif into genomic sequences significantly decreases DNase-seq signal in fibroblasts (Figure 4.9A). However, DNase-seq profiles in fibroblasts should not be sensitive to the Oct-Sox heterodimer, as the pluripotency TFs Oct4 and Sox2 are not expressed in fibroblasts. This problem is widespread across prediction tasks in the DeepSEA model, as well as the Enformer and scBasset models. We additionally found that in this model, it was possible to see feature leakage even when training single-task models on peaks from fibroblast along with peaks from other tasks (Fig 4.9B).

Additionally, we investigated SEI, a model that that takes as input any 4 kilobase window along the human genome, and predicts 21,907 chromatin profiles [Chen et al., 2022]. We simulated random order-0 sequence using the *simdna* package. We inserted the AP-1 motif (ATGAGTCAT), which is active in fibroblast, in the center of the sequence, where predictions are made in order to start with a peak region for fibroblast. We then inserted anywhere between 0 and 30 copies of the Oct-Sox motif (CCTTTGTTATGCAAAT) into the sequence as well in order to investigate how Oct-Sox modulated the peaks in fibroblast. We pulled out the DNase predictions from SEI of any cell type that was a fibroblast cell type or an iPSC cell type and plotted the DNase predictions against the number of Oct-Sox copies. We also repeated this process for a shuffled version of the Oct-Sox motif (TTATTGGACTCTATAC). We found that after 10 copies of the Oct-Sox motif were inserted into the fibroblast cells, Oct-Sox began to suppress the DNase signal (Figure 4.10A). On the other hand, the motif began to increase the DNase signal for the iPSC cell types (Figure 4.10B). While it is expected that the Oct-Sox motif increases the DNA accessibility of iPSC cell types, it is problematic that it also modulates the fibroblast accessibility, because fibroblasts do not express Oct-Sox and DNA accessibility should not be affected by the motif. Our negative controls (Figure

4.10C,D) demonstrate that changes are less prominent when using the shuffled Oct-Sox motif, as compared to the motif itself. These results demonstrate that SEI exhibits some of the behaviors that may be attributable to leakage through the shared latent layer.

4.6 Discussion

In this work we examined the occurrence of biologically implausible attributions due to either a multitask model including a shared embedding layer amongst many tasks or due to peak selection and class balancing. We analyzed this problem through theoretical analysis, simple simulations, and the use of published models. We found that peak selection in data can lead to biases in the estimates in linear models, and we expect this phenomenon to also occur in published models (Sections 4.4.1-4.5.1). We provided some theoretical basis for why this might occur to argue that training of these models should occur genome-wide in order to best avoid bias issues. We found that in addition to the problem with peak selection, there is a separate issue in multitask models that share a latent space (Section 4.5.3). In these models, there is a possibility when many tasks are related and there are a few distinct tasks that the distinct tasks train slower and traces of the other tasks may appear when feature attribution methods are used.

Due to the widespread nature of multitask models and use of interpretation methods such as SHAP, it is critical that we better understand how bias may occur in multitask settings. In these models, we showed that the commonplace practice of selecting peaks for training data may potentially lead to spurious results. Hence, if a model is not trained across the genome, care must be taken in interpreting downstream analyses as there may be inconsistencies due to the selection. In our examples, we showed that published models can lead to biologically implausible feature attributions. Hence, utmost care must be taken when interpreting attribution results in large sequence-based models. A common suggested use of sequence-based models is to better understand the relationship between sequence and genomic features. Spurious attributions used in this setting can lead to inaccurate and misleading results. Additionally, we showed that training can be slowed down for distinct

tasks when many unrelated tasks are included. Furthermore, even if the model has concluded training for most of the related tasks, it can still be improving on the distinct task. Looking at overall loss can hide the continuing improvement in the distinct task. Due to the increasing demands in training time, these phenomena can lead to models that are undertrained in some tasks. It can also be the case that in some examples, by the time the distinct task is trained, the remaining tasks are overfit. Hence, it is important to consider the loss for each of the tasks individually.

While at present we do not have a suggested solution for feature attributions, we hypothesize that a feature attribution method could be developed to only consider the head layers as opposed to through the entire network. For the problem with peak selection, we suggest that ideally models are trained genome-wide to avoid the possibility of inducing a bias through peak selection, but in cases where the training data is too large to train genome-wide, we suggest that care be taken in training and in peak selection.

Chapter 5

CONCLUSION

In our first project, **Comparing Heritability Estimators under Alternative Structures of Linkage Disequilibrium**, we critically examined published methods of estimating heritability estimates. Heritability quantifies whether a trait can be explained genetically or not. For example, if a trait is highly heritable, it may suggest that attempting to change environmental factors to manipulate that trait is not possible. On the other hand, if a trait, for example a disease, is highly heritable, it may mean that individuals who have a family history of that disease should exercise additional precautions for screenings and preventative measures. In our work, we demonstrated that LD may impact estimates of heritability. In particular, we found that for some forms of LD, estimates of heritability increased as LD increased, yet for others, the opposite trend occurred. It is important to understand these differences as we continue to explore the heritability of traits, as LD may lead to inaccurate estimates. Furthermore, we established the equivalence between some historical estimators and modern estimates and provided some theoretical rationale for this equivalence. With this equivalence, we found that even though some modern methods were designed to account for LD, the overall effect may still be limited. This work demonstrates that there is still effort to be made in accounting for LD in the context of heritability.

As we look forward in the heritability estimation literature, we are excited by the prospects of new data. Whereas many heritability estimates are based off of SNPs, this was largely because whole-genome data was not available, and even when it was, it was computationally intractable to use it. However, it has been hypothesized that part of the missing heritability issue is due to rare variants that are missed in a typical SNP chip panel, which focuses on SNPs with minor allele frequence greater than 1%. Exciting new work has been coming out

which assesses heritability on the basis of whole-genome sequencing [Wainschtein et al., 2022]. This effort has become possible in part due to our increasing ability to compute and also due to the efforts of large consortiums that collect large amounts of sequencing information. As whole-genome estimates of heritability become more prevalent, our work in considering LD becomes more relevant as well, because LD is strongest as the positions get closer. As the data starts to become denser, the positions that are considered get closer, and LD becomes a greater issue. A possible future direction looking into LD at the scale of whole-genome data would be interesting. Furthermore, work has been conducted examining the use of nonlinear methods in heritability estimation [Kerin and Marchini, 2020]. This work looks at the gene-environment interactions and may be a good step beyond classical narrow-sense heritability. We believe that these two directions are interesting and look forward to seeing where the field goes.

In our second project, **Matrix prior for data transfer between single cell data types in latent Dirichlet allocation**, we developed a tool that could be used for the improvement of single-cell analyses in LDA. Our method takes information from the output of LDA run on a large-scale reference dataset, and transfers the semantics of that dataset onto a smaller dataset. We do so in a computationally light-weight way so that subsequent analyses on the smaller dataset are not slowed down compared to analyzing them separately. Furthermore, our method is not limited to one data type, and it can generally be used for data modalities apart from scATAC-seq, which used as a case study in our project. This tool is a step in better using the wealth of data that is already available in the literature. Many atlas datasets exploring the single-cell landscape in particular tissues of model organisms exist and may potentially be leveraged to improve our understanding of fine-grained cell types. As an example, Durham et al. [2021] create an extensive scATAC-seq atlas of *C. elegans*. With our method, this could be a valuable resource for other scientists hoping to conduct smaller-scale experiments of *C. elegans*. This kind of atlas exists for many other model organisms as well. We hope that our tool serves as one step towards enabling us to discover and understand more fine-grained cell types.

One of our goals in this project was to be able to transfer information between data modalities. Because scATAC-seq is a newer technology, we expected that data scarcity may be an issue. We had originally aimed to incorporate scRNA-seq data in the analysis of scATAC-seq data, since scRNA-seq is a more mature technology with more experiments that have already been conducted. In our experiments, however, we were not able to see improvements in cell-type resolution in scATAC-seq data through the use of scRNA-seq data. One possible reason for this is that we used a simple translation between scATAC-seq and scRNA-seq data, where we simply summed the cut sites of scATAC-seq over gene-bodies, in order to move both scATAC-seq and scRNA-seq data into the same axis. Recent work [Wu et al., 2021, Zhang et al., 2022] has worked on translation between the two data modalities, and it is possible that incorporating an approach like this would improve our results in combining scATAC and scRNA-seq. Another interesting direction would be ways to incorporate large-scale datasets in analyses apart of LDA. While LDA is a powerful tool to analyze these single-cell data, it is not the only method that practitioners use to analyze their data. We believe that the problem of leveraging established data to improve analyses in new experiments is an important problem, and may hold promise in settings outside of LDA.

In our third project, **Data leakage in sequence-based multitask genomics models**, we examined a type of model that has been gaining in popularity due to its promise towards *in silico* mutagenesis. These models have become possible to train and create due to the increase in publicly available data, for example through the ENCODE project [Feingold et al., 2004], among other publicly available data sources. These data sources provide genomic tracks for a wide array of different assays and cell types, enabling machine learning methods to dissect the syntax of the genome. We analyzed the ability of these models to predict biologically meaningful predictions, and we demonstrated that although the models may perform well in some cases, in other cases, the models produce counterintuitive results that may mislead the user. We showcased through simulation that the use of the shared latent representation in a multitask setting and the selection of peaks in a joint manner are both plausible reasons for

these spurious results. We believe that these results demonstrate that great caution must be exercised when interpreting results from these machine-learning models, as it is hard to tell when the results are trustworthy, and we hope that our work leads to more improvements to the models so that they can become useful and accurate tools to advance our knowledge of genomics.

Beyond genomic information, sequence based models have also begun to appear in predicting 3D chromatin architecture data [Zhou, 2022, Fudenberg et al., 2020]. These models have the same goal of predicting genomic features using sequence, but now aim to predict contact maps of genomes. In other words, they aim to predict how regions of the genome interact with each other from sequence alone. They have shown great promise in producing predictions that rival those of competing methods that include additional sequencing information. We are interested in how biologically plausible these 3D chromatin prediction methods are, and are excited for their possibilities. Finally, we are excited by the prospect of validating *in silico* mutagenesis through real whole-genome sequencing data [Sasse et al., 2023], demonstrating that when sequence-to-function models are put in real world scenarios, they may fail to perform as expected. Sasse et al. [2023] also propose that training on diverse genomes with their associated gene expression measurements may be a promising avenue to achieve accurate *in silico* mutagenesis.

We have presented three projects at the intersection of the rapidly evolving fields of genomics and machine learning. We believe that we have made contributions in machine learning methodology to better investigate these data.

BIBLIOGRAPHY

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015.

Davide-Carlo Ambrosetti, Claudio Basilico, and Lisa Dailey. Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Molecular and Cellular Biology*, 17(11):6321–6329, 1997.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021a.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, 2021b.

Shankar Balasubramanian, David Klenerman, and Colin Barnes. Arrayed polynucleotides and their use in genome analysis, January 30 2003. US Patent App. 10/153,240.

David M Blei, A Ng, and M Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003.

Thomas J Bouchard Jr, David T Lykken, Matthew McGue, Nancy L Segal, and Auke Tellegen. Sources of human psychological differences: The Minnesota study of twins reared apart. *Science*, 250(4978):223–228, 1990.

Sharon R Browning and Brian L Browning. Population structure can inflate SNP-based heritability estimates. *The American Journal of Human Genetics*, 89(1):191–193, 2011.

Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109(1):21–29, 2015.

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, 2015.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Mario PL Calus and Jérémie Vandenplas. SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. *Genetics Selection Evolution*, 50(1):1–11, 2018.

Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, et al. Comprehensive

single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017.

Kathleen M Chen, Aaron K Wong, Olga G Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature Genetics*, 54(7): 940–949, 2022.

Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

Laura Clarke, Susan Fairley, Xiangqun Zheng-Bradley, Ian Streeter, Emily Perry, Ernesto Lowy, Anne-Marie Tassé, and Paul Flicek. The International Genome Sample Resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Research*, 45(D1):D854–D859, 2017.

Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

Robert R Corbeil and Shayle R Searle. Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38, 1976.

James F Crow and Motoo Kimura. *An introduction to population genetics theory*. Harper & Row, 1970.

Darren A Cusanovich, Andrew J Hill, Delasa Aghamirzaie, Riza M Daza, Hannah A Pliner, Joel B Berletch, Galina N Filippova, Xingfan Huang, Lena Christiansen, William S DeWitt, et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5): 1309–1324, 2018.

William M Darling. A theoretical and practical implementation tutorial on topic modeling and Gibbs sampling. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 642–647, 2011.

Kushal K Dey, Chiaowen Joyce Hsiao, and Matthew Stephens. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genetics*, 13(3): e1006599, 2017.

Lee H Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2): 269–284, 2014.

Timothy J Durham, Riza M Daza, Louis Gevirtzman, Darren Cusanovich, Olubusayo Bolonduro, William Stafford Noble, Jay Shendure, and Robert H Waterston. Comprehensive characterization of tissue-specific chromatin accessibility in L2 C. elegans nematodes. *Genome Research*, pages gr–271791, 2021.

Luke M Evans, Rasool Tahmasbi, Matt Jones, Scott I Vrieze, Gonçalo R Abecasis, Sayantan Das, Douglas W Bjelland, Teresa R de Candia, Jian Yang, Michael E Goddard, et al. Narrow-sense heritability estimation of complex traits using identity-by-descent information. *Heredity*, 121(6):616–630, 2018a.

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, 2018b.

Rongxin Fang, Sebastian Preissl, Yang Li, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K Shiau, Xinzhu Zhou, Fangming Xie, et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nature communications*, 12(1): 1337, 2021.

EA Feingold, PJ Good, MS Guyer, S Kamholz, L Liefer, K Wetterstrand, FS Collins, TR Gingeras, D Kampa, EA Sekinger, et al. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

Anthony P Fejes, Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven JM Jones. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, 2008.

R. A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

Geoff Fudenberg, David R Kelley, and Katherine S Pollard. Predicting 3D genome folding from DNA sequence with Akita. *Nature methods*, 17(11):1111–1117, 2020.

Stacey Gabriel and Liuda Ziaugra. SNP genotyping using Sequenom MassARRAY 7K platform. *Current Protocols in Human Genetics*, 42(1):2–12, 2004.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Katherine Wu, Michael Jayasuriya, Edouard Melhman, Maxime Langevin, Yining Liu, Jules Samaran, et al. scvi-tools: a library for deep probabilistic analysis of single-cell omics data. *bioRxiv*, 2021.

John Michael Gaziano, John Concato, Mary Brophy, Louis Fiore, Saiju Pyarajan, James Breeling, Stacey Whitbourne, Jennifer Deen, Colleen Shannon, Donald Humphries, et al. Million veteran program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, 70:214–223, 2016.

Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods*, 16(5):397–400, 2019.

JK Haseman and RC Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavioural Genetics*, 2(1):3–19, 1972.

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*, page 1, 2019.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

Shan Jiang, Zheng Huang, Yun Li, Chengwei Yu, Hao Yu, Yuwen Ke, Lan Jiang, and Jiang Liu. Single-cell chromatin accessibility and transcriptome atlas of mouse embryos. *Cell Reports*, 42(3), 2023.

Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24(1):1–29, 2023.

David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7): 990–999, 2016.

David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, 2018.

Matthew Kerin and Jonathan Marchini. A non-linear regression method for estimation of gene–environment heritability. *Bioinformatics*, 36(24):5632–5639, 2020.

Sobin Kim and Ashish Misra. SNP genotyping: technologies and biomedical applications. *Annual Review of Biomedical Engineering*, 9:289–320, 2007.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Marcin Kruczyk, Husen M Umer, Stefan Enroth, and Jan Komorowski. Peak Finder Metaserver-a novel application for finding peaks in ChIP-seq data. *BMC Bioinformatics*, 14:1–7, 2013.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

SH Lee, NR Wray, ME Goddard, and PM Visscher. Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics*, 88(3):294–305, 2011.

SH Lee, J Yang, ME Goddard, PM Visscher, and NR Wray. Estimation of pleiotropy between complex diseases using SNP-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542, 2012.

Zhaotong Lin, Souvik Seal, and Saonli Basu. Estimating SNP heritability in presence of population substructure in biobank-scale datasets. *Genetics*, 220(4):iyac015, 2022.

Jialin Liu, Chao Gao, Joshua Sodicoff, Velina Kozareva, Evan Z Macosko, and Joshua D Welch. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nature Protocols*, 15(11):3632–3662, 2020.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022.

Michael Lynch, Bruce Walsh, et al. *Genetics and Analysis of Quantitative Traits*, volume 1. Sinauer Sunderland, MA, 1998.

Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, 183(4):1103–1116, 2020.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7): 621–628, 2008.

National Human Genome Research Institute. The cost of sequencing a human genome, 2021. URL `https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost`.

Nao Nishida, Asako Koike, Atsushi Tajima, Yuko Ogasawara, Yoshimi Ishibashi, Yasuka Uehara, Ituro Inoue, and Katsushi Tokunaga. Evaluating the performance of Affymetrix SNP array 6.0 platform with 400 japanese individuals. *BMC Genomics*, 9:1–10, 2008.

Jonathan S Packer, Qin Zhu, Chau Huynh, Priya Sivaramakrishnan, Elicia Preston, Hannah Dueck, Derek Stefanik, Kai Tan, Cole Trapnell, Junhyong Kim, et al. A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science*, 365(6459): eaax1971, 2019.

H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

Ali Pazokitoroudi, Yue Wu, Kathryn S Burch, Kangcheng Hou, Aaron Zhou, Bogdan Pasaniuc, and Sriram Sankararaman. Efficient variance components analysis across millions of genomes. *Nature Communications*, 11(1):1–10, 2020.

Tinca JC Polderman, Beben Benyamin, Christiaan A De Leeuw, Patrick F Sullivan, Arjen Van Bochoven, Peter M Visscher, and Danielle Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7):702–709, 2015.

Zhaohui S Qin, Jianjun Yu, Jincheng Shen, Christopher A Maher, Ming Hu, Shanker Kalyana-Sundaram, Jindan Yu, and Arul M Chinnaiyan. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, 11: 1–13, 2010.

Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, 4(8):651–657, 2007.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Nehmé Saksouk, Elisabeth Simboeck, and Jérôme Déjardin. Constitutive heterochromatin formation and transcription in mammals. *Epigenetics & Chromatin*, 8:1–17, 2015.

Fred Sanger and Alan R Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–448, 1975.

Alexander Sasse, Bernard Ng, Anna Spiro, Shinya Tasaki, David A Bennett, Christopher Gaiteri, Philip L De Jager, Maria Chikina, and Sara Mostafavi. How far are we from personalized gene expression prediction using sequence-to-expression deep neural networks? *bioRxiv*, pages 2023–03, 2023.

Alicia N Schep, Beijing Wu, Jason D Buenrostro, and William J Greenleaf. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature methods*, 14(10):975–978, 2017.

Armin Schwartzman, Andrew J Schork, Rong Zablocki, Wesley K Thompson, et al. A simple, consistent estimator of SNP heritability from genome-wide association studies. *The Annals of Applied Statistics*, 13(4):2509–2538, 2019.

Jay Shendure, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741): 1728–1732, 2005.

Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

Lingyun Song and Gregory E Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb–prot5384, 2010.

Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics*, 91(6): 1011–1021, 2012.

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.

Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*, 590(7845):290–299, 2021.

Evan D Tarbell and Tao Liu. HMMRATAC: a hidden markov modeler for ATAC-seq. *Nucleic Acids Research*, 47(16):e91–e91, 2019.

Mizuho Tomioka, Masazumi Nishimoto, Satoru Miyagi, Tomoko Katayanagi, Nobutaka Fukui, Hitoshi Niwa, Masami Muramatsu, and Akihiko Okuda. Identification of Sox-2 regulatory region which is under the control of Oct-3/4–Sox-2 complex. *Nucleic Acids Research*, 30(14):3202–3213, 2002.

Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021.

Peter M Visscher, Sarah E Medland, Manuel A R Ferreira, Katherine I Morley, Gu Zhu, Belinda K Cornes, Grant W Montgomery, and Nicholas G Martin. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics*, 2(3):e41, 2006.

Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era: concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255, 2008.

Pierrick Wainschtein, Deepti Jain, Zhili Zheng, L Adrienne Cupples, Aladdin H Shadyab, Barbara McKnight, Benjamin M Shoemaker, Braxton D Mitchell, et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics*, 54(3):263–273, 2022.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009.

Bowen Wang, Serge Sverdlov, and Elizabeth A Thompson. Efficient estimation of realized kinship from single nucleotide polymorphism genotypes. *Genetics*, 205(3):1063–1078, 2017.

Huanjun Wang, Yan Mei, Cheng Luo, Qun Huang, Zifeng Wang, Guan-Ming Lu, Lili Qin, Zhun Sun, Chao-Wen Huang, Zhi-Wen Yang, et al. Single-cell analyses reveal mechanisms

of cancer stem cell maintenance and epithelial–mesenchymal transition in recurrent bladder cancer scRNA-and scATAC-seq analyses of human bladder cancer. *Clinical Cancer Research*, 27(22):6265–6278, 2021.

Justin Wood, Patrick Tan, Wei Wang, and Corey Arnold. Source-lda: Enhancing probabilistic topic models using prior knowledge sources. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 411–422. IEEE, 2017.

Kevin E Wu, Kathryn E Yost, Howard Y Chang, and James Zou. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences*, 118(15):e2023070118, 2021.

J Yang, B. Benyamin, BP McEvoy, S Gordon, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42:565–569, 2010.

J Yang, T Ferreira, AP Morris, Medland SE, et al. Conditional and joint multiple-SNP analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nat Genet*, 44(4):369–375, 2012.

Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1): 76–82, 2011.

Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna AE Vinkhuyzen, Sang Hong Lee, Matthew R Robinson, John RB Perry, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47(10):1114, 2015.

Kihoon Yoon and Stephen Kwek. An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics. In *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, pages 6–pp. IEEE, 2005.

Ran Zhang, Laetitia Meng-Papaxanthos, Jean-Philippe Vert, and William Stafford Noble. Semi-supervised single-cell cross-modality translation using Polarbear. In *International Conference on Research in Computational Molecular Biology*, pages 20–35. Springer, 2022.

Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):1–9, 2008.

Jian Zhou. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nature genetics*, 54(5):725–734, 2022.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015.

Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, 2018.