INAUGURAL – DISSERTATION

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht – Karls – Universität

Heidelberg

vorgelegt von

Heinen, Tobias, M.Sc.

aus Duisburg

Tag der mündlichen Prüfung:

Statistical Models for Single-Cell Genetics

Betreuer: Prof. Dr. Ullrich Köthe

Prof. Dr. Oliver Stegle

Abstract

The impact of genetic variation on molecular traits such as gene expression is known to vary substantially across molecular contexts, such as cell types and states. Conventional approaches for molecular phenotyping rely on bulk sequencing data, representing aggregate measurements across thousands or millions of cells from sorted cell populations or whole tissue sections. As a result, these methods often lack the resolution to capture fine-grained cellular heterogeneity, impacting discovery and explanatory power for mapping genetic associations. More recently, single-cell sequencing technologies have fundamentally changed our ability to study cellular identity. By isolating and analyzing individual cells, these methods have revealed the remarkable diversity that exists even within seemingly homogeneous cell populations and have enabled researchers to create detailed profiles of the underlying transcriptomic and epigenetic landscape. Integrating single-cell measurements with genotyping data opens up new opportunities for linking genetic variation to molecular processes in a context-specific manner, to improve disease diagnosis, risk prediction and the design of therapeutic interventions.

Existing analysis strategies typically require aggregation of measurements in discrete cell clusters, in order to apply statistical methods originally developed for bulk-sequencing data. As a consequence, these methods may fail to capture more subtle allelic regulation and interaction effects with continuous biological processes such cell differentiation or development. This thesis develops three new computational methods designed to model genetic effects at the level of individual cells, to harness the full potential of single-cell measurements. The first method, scDALI, enables the assessment of allelic imbalance in single-cell count data, particularly for effects associated with specific cell types and states. The model is applied to open chromatin data from *Drosophila Melanogaster* embryos to examine allele-specific changes in chromatin accessibility across development. In addition, a variant of the variational autoencoder model is presented, aiding in the inference of cell types and states from

temporally resolved, high-dimensional chromatin profiles. The second contribution, CellRegMap, is the first principled method and statistical test to identify genetic variants with context-specific effects on gene expression, requiring no prior discretization of single-cell profiles. The model is validated using a semi-synthetic simulation framework, demonstrating the importance of accounting for both donor- and cell-level sources of confounding variation to obtain calibrated test statistics. Finally, the third method, LIVI, uses a variational autoencoder model with adversarial regularization to disentangle donor-specific and shared gene expression variation in population-scale transcriptome data. Once trained, the LIVI latent space can be used to define efficient tests for identifing persistent or context-specific genetic effects on expression factors.

Zusammenfassung

Der Einfluss genetischer Variation auf molekulare Merkmale wie die Genexpression kann je nach molekularem Kontext, wie beispielsweise dem Zelltypen und -zustand, erheblich variieren. Herkömmliche Ansätze zur molekularen Phänotypisierung basieren auf Bulk-Sequenzierung, die aggregierte Messungen über Tausende oder Millionen von Zellen aus sortierten Zellpopulationen oder ganzen Gewebeschnitten darstellen. Diesen Methoden bieten meist nicht die benötigte Auflösung, um zelluläre Heterogenität detailliert zu erfassen, was sich auf die statistische Power bei der Kartierung genetischer Zusammenhänge auswirkt. In jüngerer Zeit haben Technologien zur Einzelzellsequenzierung unsere Fähigkeit, die zelluläre Identität zu untersuchen, grundlegend verändert. Durch die Isolierung und Analyse einzelner Zellen haben diese Methoden die bemerkenswerte Vielfalt aufgedeckt, die selbst innerhalb scheinbar homogener Zellpopulationen existiert, und es der Wissenschaft ermöglicht, detaillierte Profile der zugrunde liegenden transkriptomischen und epigenetischen Landschaft zu erstellen. Die Integration von Einzelzellmessungen mit Genotypisierungsdaten eröffnet neue Möglichkeiten, genetische Variation kontextspezifisch mit molekularen Prozessen zu verknüpfen, um die Krankheitsdiagnose, Risikovorhersage und die Entwicklung therapeutischer Interventionen zu verbessern.

Bestehende Analysestrategien erfordern typischerweise die Aggregation von Messungen in diskreten Zellclustern, um statistische Methoden anzuwenden, die ursprünglich für Daten aus der Bulk-Sequenzierung entwickelt wurden. Infolgedessen gelingt es diesen Methoden teilweise nicht, subtilere genetische Regulations- und Interaktionseffekte mit kontinuierlichen biologischen Prozessen wie der Zelldifferenzierung oder -entwicklung zu erfassen. Diese Arbeit entwickelt drei neue Methoden zur Modellierung genetischer Effekte auf der Ebene einzelner Zellen, um das volle Potenzial von Einzelzellmessungen auszuschöpfen. Die erste Methode, scDALI, ermöglicht die Analyse der allelischen Verteilung in Einzelzelldaten, insbesondere für Effekte, die mit bestimmten Zelltypen und -zuständen verbunden

sind. Das Modell wird auf Chromatindaten von *Drosophila Melanogaster*-Embryonen angewendet, um allelspezifische Veränderungen in der Zugänglichkeit des Chromatin im Laufe der embryonalen Entwicklung zu untersuchen. Darüber hinaus wird eine Variante des Variational-Autoencoder-Modells vorgestellt, die die Inferenz von Zelltypen und -zuständen aus zeitlich aufgelösten, hochdimensionalen Chromatinprofilen unterstützt. Der zweite Beitrag, CellRegMap, ist die erste Methode zur Identifizierung genetischer Varianten mit kontextspezifischen Auswirkungen auf die Genexpression, die keine vorherige Diskretisierung von Einzelzellprofilen erfordern. Das Modell wird mithilfe semi-synthetischer Daten validiert. Die Analyse zeigt, wie wichtig es ist, Störvariationen sowohl auf Spender- als auch auf Zellebene zu berücksichtigen, um kalibrierte Teststatistiken zu erhalten. Die dritte Methode, LIVI, ist ein Variational-Autoencoder-Modell mit spezieller Regularisierung, um spenderspezifische und geteilte Genexpressionsvariationen in Transkriptomdaten im Populationsmaßstab zu entflechten. Nach dem Training kann der LIVI-Latentraum zur Definition effizienter Tests zur Identifizierung persistenter oder kontextspezifischer genetischer Effekte auf Expressionsfaktoren verwendet werden.

Acknowledgements

First and foremost, I would like to thank my supervisor at the German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ), Prof. Oliver Stegle, for his outstanding support and guidance throughout my PhD. I would also like to express my gratitude to Prof. Ullrich Köthe for accepting me as a doctoral candidate, as well as to Heidelberg University, the DKFZ and the European Molecular Biology Laboratory (EMBL) for providing a fantastic environment for scientific research.

As is the nature of interdisciplinary research, this work would not have been possible without my collaborators. In particular, I would like to thank Dr. Stefano Secchia, Dr. Anna Cuomo and Danai Vagiaki for their valuable insights and joint efforts. I would also like to extend my thanks to all other past and present members of the Stegle group for stimulating discussions and advice.

Last but certainly not least, I thank my family, for cheering me on all the way. Thank you Mom, Dad & Juju. Thank you Hana, thank you Jakob. Daddy has finished writing his book.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The study of how genetic information encodes observable traits lies at the heart of modern genetics and biology. This process encompasses the transformation of genetic sequences into functional molecules that ultimately give rise to the characteristics and behaviours of living organisms. Understanding the mechanisms underlying this complex process is a fundamental challenge in biological research, as it holds the key to deciphering the intricacies of development, evolution, and disease.

1.1 From genotype to phenotype

The *genotype* of an organism refers to the genetic information stored in its DNA sequence. Genes, functional units within the DNA sequence, are transcribed into RNA in a process known as gene expression. The resulting RNA may be directly functional or used as a template for synthesizing proteins, which shape cellular structure or fulfil essential roles in the catalysis of metabolic reactions, DNA replication, cell communication, or development [1]. Other non-protein-coding regions of the genome contain regulatory sequences that may be bound by proteins to initiate, enhance, or repress gene expression and control the amount of RNA produced by specific genes withing a cell. This intricate process of gene expression is used by all known life on earth, from single-celled prokaryotes to complex eukaryotic organisms such as ourselves. It is the most fundamental level at which genetic variation shapes observable traits, the *phenotype*.

The exploration of the genetic basis underlying complex traits and diseases has been a long-standing pursuit in human genetics. Enabled by advances in chip-based microarray sequencing technologies in the early 2000s, Genome-wide association studies (GWAS) [2, 3] have identified links between genetic variation and phenotypic traits of interest, such as disease susceptibility, drug response, or various quantitative traits. These studies involve scanning the genome of a large cohort of individuals to detect genetic variants, typically single nucleotide polymorphisms (SNPs), that are statistically correlated with the trait under investigation. The first GWA studies focused on binary case-control designs to chart the genetic basis of a variety of diseases, such as age-related macular degeneration [3], Crohn's disease [4], inflammatory bowel disease [5] and type 2 diabetes [6]. Subsequently, GWA studies have started to increasingly incorporate quantitative traits such as BMI or blood pressure [7]. This development is driven in part by the recognition that selecting a control group can be challenging, as healthy individuals may develop a disease later in life, thus reducing power to detect genetic links [2]. Through large-scale collaborations and meta-analyses [8, 9], and in particular the generation of extensive population-based cohorts and biobanks [10, 11], GWAS continues to uncover thousands of genetic loci associated with diverse traits, providing valuable insights into the genetic architecture of complex phenotypes [12, 13]. As of June 2023, the NHGRI-EBI Catalog of human genome-wide association studies [14] contains 6,422 publications and lists more than half a million SNP-trait associations.

While GWAS have been successful in identifying genetic variants associated with complex traits, understanding the functional implications of these variants is often challenging. In some cases, a genetic variant may directly affect the structure of a protein, e.g., by changing a section of DNA coding for one of the amino acids within its sequence. A prominent example is sickle cell anaemia, a blood disorder caused by genetic mutations in both copies of the β-globin (HBB) gene on chromosome 11 [15][1]. Diseases such as sickle cell anaemia are also known as monogenic disorders, caused by genetic mutations in a single gene. For many common health disorders, however, there are multiple directions of added complexity. First, monogenicity is the exception rather than the rule, with many common disorders being driven by multiple genetic factors (polygenic disorders), as well as environmental exposures and interactions thereof [16, 17]. Second, most of the associated signals derived from GWAS

[1] HBB produces haemoglobin, the oxygen-transport protein found within red blood cells. The abnormal haemoglobin forms long strands within red blood cells, leading to reduced cellular elasticity that prevents these cells from passing through narrow capillaries.

map to noncoding regions of DNA [18], which make up close to 99% of the human genome [19]. Third, because whole sections of DNA (haplotypes) are inherited together from a single parent, there is substantial correlation between variants. This nonrandom association of genomic loci, termed linkage disequilibrium (LD), has been used in early microarray-based GWAS to reduce costs, by only sequencing a set of selected tag SNPs which can serve as surrogates for groups of tightly linked variants [20]. As a consequence, however, pinpointing association signals to specific causal variants is often difficult [20].

1.2 Exploring the genetic basis of gene regulation

Expression quantitative trait loci (eQTL) are genetic variants that regulate gene expression levels, either by directly affecting transcription (in which case they are termed *cis*-acting) or through other regulatory mechanisms (*trans*-acting) [21, 22] (**Fig.**1.1). *Cis*-regulatory elements are small sections of DNA (usually 5-12 nucleotides in length) dispersed throughout the genome, which can be bound by regulatory proteins known as transcription factors (TFs) to control the expression of nearby[2] genes [1]. A *cis*-acting variant may for instance hamper TF binding, thus repressing gene expression of associated target genes. *Trans*-acting variants influence gene expression of more distant genes by acting through another molecule, e.g., by repressing the expression of an intermediary transcription factor in *cis*.

eQTL studies examine the correlation between the prevalence of genetic mutations and RNA transcript abundances of either local or distal genes, in order to link genetic variants to molecular processes and cellular functions. The widespread availability of high-throughput genotyping and transcriptomics technologies, such as large-scale genotyping arrays, whole-genome sequencing, and bulk RNA sequencing (RNA-seq) [24], has facilitated the application of this approach to a wide range of biological systems, from cell line models [25, 26] to post-mortem collected human tissues [27].

The concept that underlies eQTL mapping has also been applied to other molecular traits, such as chromatin accessibility [28–30] and other epigenetic features [31, 32], to improve the identification of causal eQTL variants and better understand their functional implications.

[2] Many *cis*-regulatory elements are located close a gene's transcription start site (TSS). However, by definition, a *cis*-regulatory element is only required to be on the same chromosome as its target gene. To regulate transcription, the element may be brought into close proximity of its target through DNA looping [1]. In practice, many studies consider only variants within a fixed (e.g. 1 megabase) window around a gene of interest.

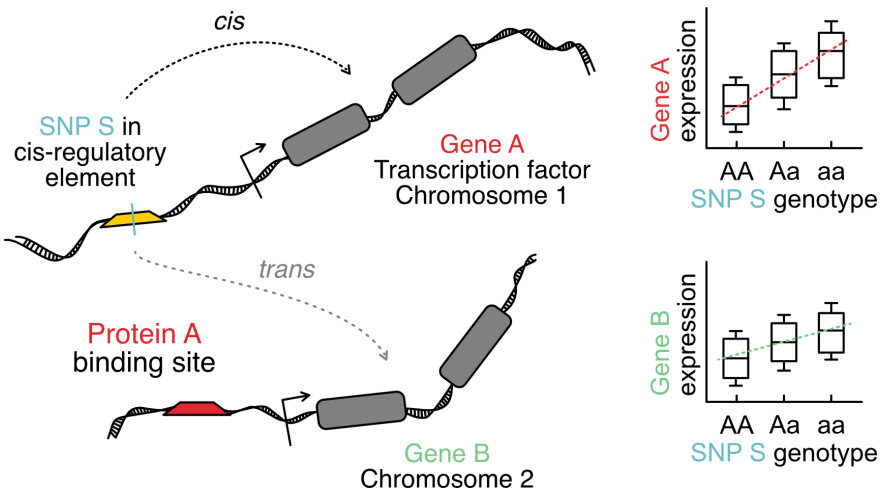

Figure 1.1. *Cis* and *trans* eQTL. A SNP S in a regulatory region proximal to gene A transcription start site (TSS, angled arrow) affects its expression in *cis*. Gene A encodes a transcription factor, leading to downstream *trans*-regulatory genetic effects on gene B. Grey boxes denote exonic regions. Adapted from [23].

The spatial organization and compaction of chromatin, the mixture of DNA and proteins found in the cell nucleus, is an essential regulatory mechanism governing gene expression (**Fig.** 1.2) [1]. DNA is wrapped around histone proteins to form nucleosomes, basic structural units of the chromatin architecture. The chromatin structure of different cell types and states is highly specific and determines which protein factors and other elements of the transcriptional machinery, such as RNA polymerase, have access to the genetic code. Genetic variants are known to interact with chromatin accessibility, for instance, by impacting the dynamic modification of nucleosome arrangements through DNA-binding chromatin remodelers [33]. Systematic association analyses of genetic variation and chromatin accessibility, quantified using high throughput sequencing-based assays such as DNase-seq [34] or ATAC-seq [35], have revealed changes in transcription factor accessibility as a major mechanism mediating the effects of non-coding genetic variation on gene expression [28].

1.2.1 Allelic imbalance

In a diploid organism like humans, each gene typically has two alleles (variants of a gene), one inherited from each parent. Standard QTL mapping efforts consider the aggregate expression of both alleles and map genetic effects on molecular traits based on inter-individual

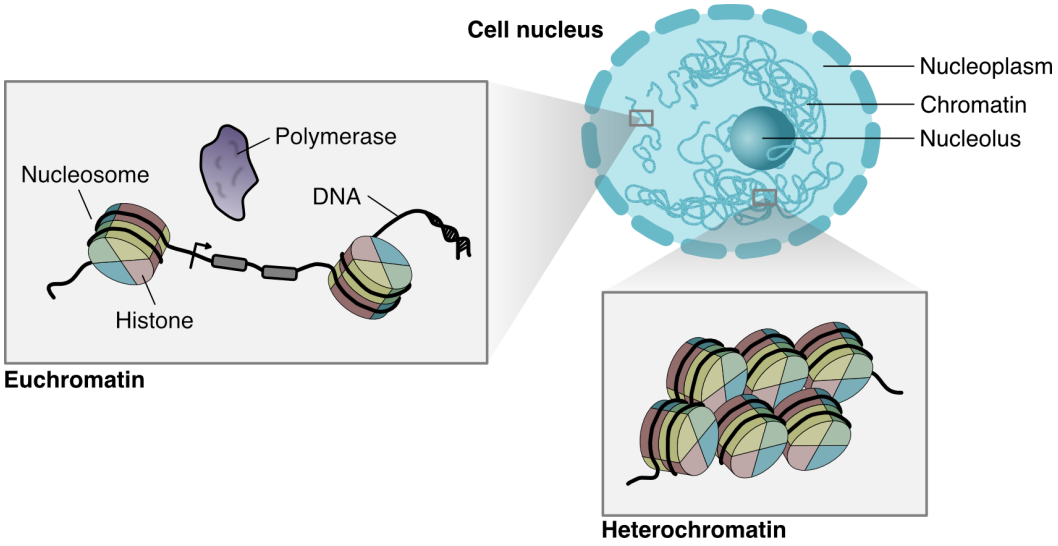

Figure 1.2. Chromatin accessibility as a regulatory mechanism. Chromatin refers to the mixture of DNA and proteins found in the nucleoplasm within the cell nucleus. DNA winds around histone octamers, protein complexes formed by eight histone proteins, to form structural units known as nucleosomes. Euchromatin refers to lightly packed chromatin, where the DNA can easily be accessed by protein factors and RNA polymerase to initiate gene expression. Tightly packed heterochromatin is less accessible to the transcriptional machinery.

variation across a population. Alternatively, sequencing-based phenotyping may also be used to measure allele-specific signals, which allows to map genetic effects even in a single individual (**Fig.** 1.3). Using heterozygous variants to distinguish parental haplotypes, sequenced transcripts can be assigned to either allele and allelic differences can be quantified [36]. Several studies showed that allelic imbalances are common in various molecular traits, such as transcription factor binding, chromatin states and gene expression [28,37–39]. These differences in allelic activitiy can arise as the result of *cis*-regulatory effects on a trait, if an individual is heterozygous for the causal variant, providing additional information on genetic associations that is largely complementary to inter-individual variation [40].

1.2.2 Challenges in genetic association mapping

The design of statistical methods for GWA or eQTL studies faces similar challenges. Avoiding the identification of spurious associations necessitates careful consideration of possible confounding factors, such as population structure [41,42]. Furthermore, because both GWA

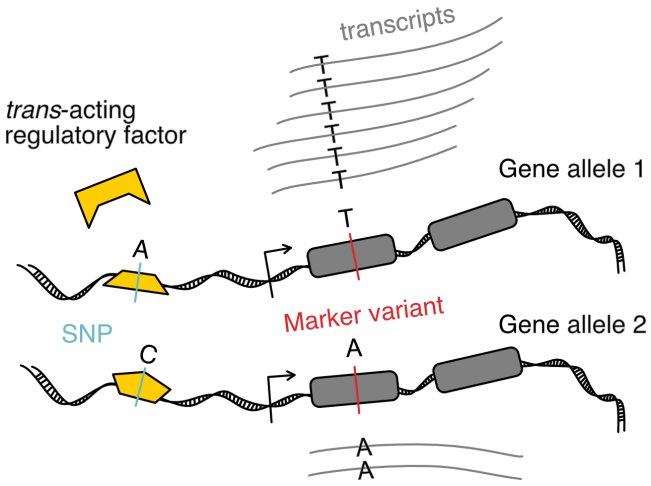

Figure 1.3. The genetic basis of allelic imbalance. Using heterozygous variants as natural markers (red), sequenced transcripts can be assigned to either allele in diploid organisms. A SNP in a *cis*-regulatory element impacts binding of the associated *trans*-acting factor, thereby suppressing transcription of allele 2 and causing allelic imbalance.

and eQTL studies frequently perform hundreds of thousands to millions of tests, multiple testing correction is vital to reducing the probability of obtaining false positives associations by chance alone [43, 44]. Linear mixed models (LMMs) are a commonly used class of models in genetic association analyses, as they offer a flexible and effective way of decomposing different sources of outcome variation in a regression framework [45–48]. LMMs are also an essential component of the modeling approaches developed in this thesis, and will be discussed in greater detail in chapter 2.

1.2.3 From variant to gene to function

Combining GWAS and eQTL analyses has significantly advanced our understanding of the functional consequences of genetic variation, enabling the identification of putative mechanistic links between genetic variation and phenotypic outcomes [49, 50]. A range of statistical methods have been developed to determine if two association signals from GWAS and eQTL studies likely share the same underlying causal variant [51] and whether the GWAS phenotype is mediated by an intermediate expression trait (Mendelian randomization [52]). By mapping the regulatory effects of disease-associated genetic variants, an integrative approach can offer insights into the biological pathways perturbed in diseases and highlights

key genes and molecular processes involved in disease pathogenesis [21]. In fact, it has been estimated that drugs focusing on therapeutic targets supported by genetic information are twice as likely to succeed as those lacking such evidence, from initial phase I studies to the final approval stage [53]. Nevertheless, across studies only a limited number (25-50%) of disease-associated loci from GWAS share effects with known tissue-level eQTL [27, 54, 55]. Given that GWAS variants are enriched in proximity to putative causal genes and often locate in known regulatory regions [54], it appears that most eQTL studies lack the resolution to discover these 'missing links' of genetic regulation [54].

1.3 Resolving molecular contexts

Cells are the fundamental building blocks of all life on earth and they exhibit remarkable diversity and complexity, both within and across tissues [1]. This cellular heterogeneity arises from the interplay of specific regulatory programs, governing the transcriptional landscape. As a result, gene expression levels can vary substantially across cell types and states - the molecular context. While historically, eQTL studies were primarily focused on blood cells [37, 56], as samples were readily accessible, more recent efforts have targeted a range of human tissues [27, 57–62]. These results highlighted the importance of stratifying by the molecular context, showing that between 29% to 80% of eQTL are cell type-specific. The regulatory impact of disease-associated variants on gene expression therefore needs to be evaluated in tissues and cell types relevant to the disease.

1.3.1 Single-cell sequencing

All of the above mentioned eQTL studies rely on *bulk* RNA sequencing assays, which measure average molecular profiles of gene expression across millions of cells from pooled cell populations [24]. As a consequence, these methods often fail to accurately capture the nuances of cellular heterogeneity, such as rare or previously unknown cell states, as well as continuous transitions. Single-cell sequencing [63–66] has transformed our ability to dissect the complex molecular contexts within individual cells, offering unprecedented resolution and granularity. By capturing genomic, transcriptomic, or epigenomic information from individual cells, these techniques enable the unbiased exploration of cellular heterogeneity, lineage dynamics, and the impact of cellular microenvironments [67–72]. The most widely used single-cell sequencing technology is single-cell RNA sequencing (scRNA-seq) [64–66, 73],

which profiles the transcriptome of thousands or even millions of individual cells in a single experiment. Using this data, researchers can identify cell types or continuous trajectories (e.g., across time or pseudo-time [74–76]), link cell states to individual genes (e.g., to define marker genes for diverse cell populations [71, 77]) and study subtle variation in gene expression such as transcriptional bursting [78].

Single-cell technologies have been in use for more than a decade. However, applying these assays to generate large multi-sample data necessary for genetic association testing required both technological improvements as well as considerable reductions in cost [79]. In 2013, a study first demonstrated the benefits of mapping genetic effects on gene expression distribution in single cells [80]. The authors showed that many heritable expression phenotypes such as bursting patterns and other dynamic expression fluctuations were masked when considering average gene expression measurements across many cells. Despite the modest size of the data (92 genes from the WNT pathway; 1,440 single cells from 15 individuals), using single-cell data improved statistical power to discover these effects. In the following years, advances in assay technologies and experimental design (e.g., multi-individual pooling and demultiplexing [81,82]) have made it feasible to sequence cells from hundreds of individuals to map eQTL at a genome-wide scale [81,83–85]. For example, recent work from Cuomo et al. [84], revealed dynamic genetic effects on cell fate decision in human induced pluripotent stem cells (iPSCs), illustrating how single-cell RNA-sequencing can be used to pinpoint genetic associations to previously inaccessible cell states. The number of published studies has more than doubled in between January and December 2022 [79], with the largest dataset now encompassing approximately one million cells from almost 1,000 individuals [86].

Single-cell sequencing data poses a number of challenges for the development of analyses strategies and computational methods [73, 87, 88]. Measurements typically exhibit high levels of noise and technical variability due to a variety of factors, such as amplification biases, contamination with ambient RNA, variation in sequencing depth per cell and experimental batch effects [87, 89, 90]. Owing to the small amount of biological material, the total number of sequencing reads for an individual cell is usually low, leading very sparse measurements when evaluating molecular traits at a genome-wide scale [91,92]. Designing statistical methods that successfully account for biological and technical sources of noise in this data, remains an area of ongoing research [87, 92, 93]. It is of particular importance in the context of single-cell data integration, to form comprehensive atlases of cellular

heterogeneity across multiple datasets from different samples, experiments and sequencing technologies [69,71,87,94]. Moreover, modern single-cell technologies are now frequently used to profile millions of cells [95,96], making computational scalability a key priority in method development.

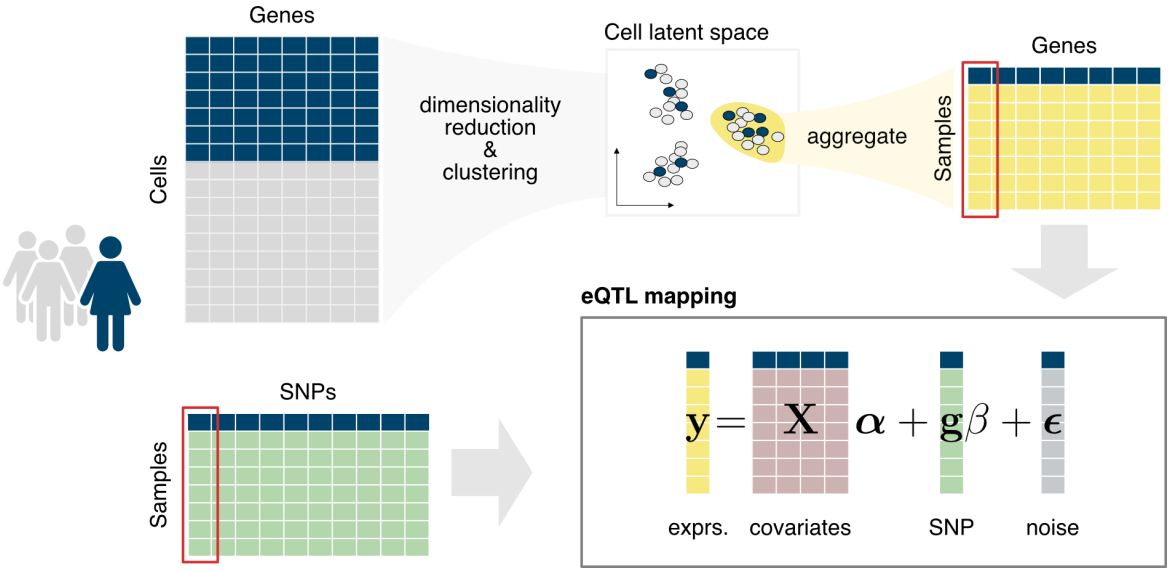

Figure 1.4. eQTL mapping based on aggregate single-cell gene expression counts. Cell clusters are identified using the observed expression profiles. Counts are then aggregated within each cluster, to obtain a sample-by-gene matrix of 'pseudo-bulk' measurements. eQTL are mapped independently for every SNP-gene pair (example highlighted in red) and cluster. Data from a single individual is highlighted in dark blue.

1.3.2 Single-cell eQTL mapping

To map genetic effects from single-cell sequencing data, cells are typically first divided into discrete groups based on their gene expression profiles [84,85,97,98]. Most commonly, the goal is to identify cell types, e.g., using clustering algorithms or by mapping to a reference dataset [86]. Alternatively, binning may be used, for example, to stratify cells along a continuous differentiation trajectory [84]. By considering aggregate expression ('pseudo-bulk') counts within each group, genetic effects can then be identified using established association tests originally designed for bulk sequencing data [97]. An overview of the workflow is shown in **Fig.**1.4. This approach immediately benefits from the plethora of data normalization and analysis strategies developed for processing scRNA-seq data and implemented

in popular software suites such as Seurat [99] or Scanpy [100]. However, there are also drawbacks to pseudo-bulk aggregation. While the data-driven stratification of single cells improves on bulk sequencing approaches, there is still a possibility that biologically relevant heterogeneity within discrete populations is obscured. For instance, important regulatory changes along a differentiation trajectory might not align with a chosen sequence of time intervals. Additionally, downstream association tests also neglect covariation in expression and genetic regulation between cell groups. These simplifications are likely detrimental to the accurate estimation of nuisance parameters and genetic effect sizes, leading to reduced discovery power.

1.4 Contributions

This thesis focusses on the development of statistical and computational methods designed to capture the relationship between genetic variation and quantitative traits such as gene expression from single-cell sequencing data. Importantly, these methods do not require prior discretization of cell states, and instead aim to harness the full potential of single-cell measurements. To this end, I combine classical statistical methodology, such as linear mixed models and statistical hypothesis testing, with recent advances in latent variable modeling from the field of machine learning.

First, I present scDALI, a statistical test and analysis framework for allelic imbalance in single-cell count data. scDALI is the first principled method for assessing allele-specific variation in a single-cell context, allowing to test for effects that align with particular cell types and states. The model is applied to scATAC-seq data of *Drosophila Melanogaster* embryos, to characterize allele-specific changes in chromatin accessibility across development. As part of this analysis, I also describe a novel variant of the Variational autoencoder model, targeted at temporally resolved scATAC-seq data, to infer cell types and states from high dimensional open chromatin profiles. I show how some of the core ideas underlying scDALI can be translated to the multi-sample setting and introduce CellRegMap, a statistical test designed specifically for identifying eQTL with single-cell resolution. CellRegMap uses the linear mixed model framework to solve a variance decomposition problem, allowing to assess the fraction of gene expression variation driven by genetics, cell state, as well as confounding variables. Importantly, CellRegMap models interactions between cell state and genetics, allowing to capture genetic effects that are specific to certain cell types or states.

I validate the model using a semi-synthetic simulation framework, designed to capture confounding effects present in real data, to assess statistical calibration and power.

Lastly, I present LIVI, an extension of the Variational Autoencoder model designed to disentangle donor-specific and shared, canonical gene expression variation from population-scale scRNA-seq data. LIVI uses an adversarial approach to remove donor-specific effects from the observed expression profiles, and then re-introduces these effects in an interpretable linear interaction model. Notably, cell-state-specific donor effects are summarized in trainable embedding vectors at level of donors (rather than individual cells), that can be efficiently screened for associations with genetic information or clinical covariates.

1.5 Thesis outline

The remainder of this thesis is organized as follows:

- **Chapter** 2 covers the mathematical foundations for the modeling approaches presented in this thesis. First, I present a brief overview of linear mixed models and statistical hypothesis tests for genetic association analyses. Second, this chapter provides an introduction to the variational autoencoder, a flexible latent variable framework, facilitating scalable probabilistic inference when using complex, non-linear observation models. Relevant extensions, such as conditional variational autoencoders and observation models for single-cell data, are also discussed.

- In **Chapter** 3 presents the scDALI model and applications to scATAC-seq and scRNA-seq data, as well as simulation studies. Subsequently, the CellRegMap model and semi-synthetic simulation framework are introduced. The chapter finishes with an example application of CellRegMap to map genetic effects on gene expression in scRNA-seq data from human induced pluripotent stem cells (iPSCs), differentiating towards definitive endoderm.

- **Chapter** 4 introduces the LIVI model. The method is validated on simulated data and applied to a large dataset of one million human peripheral blood mononuclear cells (PBMCs).

- **Chapter** 5 summarizes the thesis and discusses future research directions.

Chapter 2

Mathematical foundations

This chapter reviews key mathematical concepts used in this thesis and is structured into two parts. In section 2.1, I discuss linear mixed models (LMMs), a family of statistical models commonly used to perform genetic association analyses such as GWAS or eQTL studies. The second major part of this chapter, section 2.2, provides an introduction to the variational autoencoder model (VAE). VAEs constitute a class of probabilistic latent variable models, as well as an inference scheme, both of which are typically implemented using artificial neural networks.

Terminology

I will denote matrices using bold capitalized letters, for example, $\mathbf{X} \in \mathbb{R}^{N \times D}$. Similarly, vectors will be denoted using bold lower-case letters, e.g., $\mathbf{x} \in \mathbb{R}^{D}$. By default, all vectors are column-oriented. For example, the n-th row of $\mathbf{X} \in \mathbb{R}^{N \times D}$ is $\mathbf{x}_n^T$, the vector transpose of $\mathbf{x}_n$. The notation for matrix calculus follows [101]. In particular, for any multivariate function f of $\mathbf{x}$, $f : \mathbb{R}^N \mapsto \mathbb{R}$, $\frac{\partial f}{\partial \mathbf{x}}$ and $\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T}$ denote the gradient and Hessian of f, respectively.

2.1 Linear mixed models

Linear mixed models are an extension of the linear regression framework. These models incorporate random effect terms to represent covariance structure between observations, providing a flexible way to control for confounding sources of phenotypic variation such as relatedness between individuals. This section contains a brief overview of the model

and relevant inference techniques. For a comprehensive introduction to linear mixed models see for example [102]. Sections 2.1.1-2.1.2 introduce ordinary linear regression models and the general mixed model description. Parameter estimation using (restricted) maximum likelihood is discussed in sections 2.1.3 and 2.1.4. Section 2.1.5 covers basic likelihood-based methods for statistical hypothesis testing. In section 2.1.6, a specific score-based test for variance components in linear mixed models is derived. Section 2.1.7 briefly discusses variance component analysis and normalization. Non-Gaussian likelihood models and inference methods are introduced in section 2.1.8. Finally, section 2.1.9 addresses specific challenges in genetic association screens, in particular, confounding effects and multiple testing adjustments.

2.1.1 Linear models for independent observations

We consider a dataset $\mathbf{X} \in \mathbb{R}^{N \times D}, \mathbf{y} \in \mathbb{R}^N$ of N input-output pairs $(\mathbf{x}_n, y_n)$. Linear regression models the mean of the output variable as a linear function of the input vector,

$$\mathbb{E}[\mathbf{y}] = \mathbf{X}^T \boldsymbol{\beta}, \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^D$ is the coefficient or fixed effect vector[1]. In linear regression we assume that the data is *independent and identically distributed* (i.i.d) under a Gaussian observation model. That is, the joint distribution of output variables $\mathbf{y} \in \mathbb{R}^N$ follows a multivariate Gaussian distribution

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{\mathcal{I}}_N), \tag{2.2}$$

where $\boldsymbol{\mathcal{I}}_N \in \mathbb{R}^{N \times N}$ is the identity matrix. Under this model, any correlation in the response variable $\mathbf{y}$ emerges as a result of the linear dependence on $\mathbf{X}$; the measurement errors are independent. This assumption markedly simplifies parameter estimation, but may be inappropriate in many practical applications. A classical example is an analysis discussed in Laird and Ware [103], where data from 200 school children was collected across multiple time points to assess the effect of air pollution on pulmonary function. In such longitudinal studies, serial correlations among measurements from the same individual are to be expected. A similar challenge is often encountered in genomic studies, where multiple biological samples are collected from the same individual or groups of genetically related individuals.

[1] $\boldsymbol{\beta}$ may include an intercept term, by adding a constant column of 1s to $\mathbf{X}$

2.1.2 Mixed-effect models

Linear mixed models, sometimes also called multilevel or hierarchical models, extend the linear model introduced in the previous section to include both fixed and random effects. The result is a flexible framework for capturing correlations between observations. In its general form, the univariate model is typically given as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\mathbf{y} \in \mathbb{R}^N$ is the response vector, $\mathbf{X} \in \mathbb{R}^{N \times D}$ is the covariate matrix for fixed effects $\boldsymbol{\beta} \in \mathbb{R}^D$, $\mathbf{Z} \in \mathbb{R}^{N \times F}$ is the covariate matrix for random effects $\mathbf{u} \in \mathbb{R}^F$ and $\boldsymbol{\epsilon} \in \mathbb{R}^N$ captures residual noise. Unlike the the fixed effect vector $\boldsymbol{\beta}$, for which the model assumes a single true parameter in the population, $\mathbf{u}$ is itself a random variable,

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \tag{2.4}$$

where $\mathbf{K}$ typically involves unknown dispersion parameters or variance components that need to be estimated. As in 2.2, the residual errors are assumed to be independent and normally distributed,

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\mathcal{I}}_N). \tag{2.5}$$

Using basic properties of the multivariate normal distribution to marginalize over $\mathbf{u}$, one can obtain the marginal likelihood

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{K}\mathbf{Z}^T + \sigma^2 \boldsymbol{\mathcal{I}}_N). \tag{2.6}$$

In contrast to eq. 2.2, the distribution of $\mathbf{y}$ is no longer required to factorize across observations. In fact, this model can represent arbitrary covariance structure: Any real symmetric matrix $\boldsymbol{\Sigma}$ is positive semi-definite (that is, a valid covariance matrix) if and only if it may be decomposed as $\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{Z}^T$ for a real matrix $\mathbf{Z}$[2].

2.1.3 Maximum likelihood estimation

Let $\mathbf{V} = \mathbf{Z}\mathbf{K}\mathbf{Z}^T + \sigma^2 \boldsymbol{\mathcal{I}}_{\boldsymbol{N}}$ be the covariance matrix of the marginal model and let $\boldsymbol{\theta}$ denote the vector of unknown variance components of $\mathbf{V}$. The most common method of estimation

[2]For all real symmetric positive semi-definite matrices $\boldsymbol{\Sigma}$, there exists a decomposition $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a is a real lower triangular matrix with non-negative diagonal entries (Cholesky decomposition). Conversely, $\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} = \mathbf{v}^T\mathbf{Z}\mathbf{Z}^T\mathbf{v} = ||\mathbf{Z}^T\mathbf{v}||_2^2 \geq 0$ for all $\mathbf{v}$.

for the mixed model parameters is maximum likelihood (ML). The joint probability density function under the marginal model is

$$f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2}|\mathbf{V}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}, \tag{2.7}$$

where $|\mathbf{V}|$ denotes the determinant of $\mathbf{V}$. Therefore, the log-likelihood function is

$$\ell(\boldsymbol{\beta},\boldsymbol{\theta};\mathbf{y}) = \log f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \text{const.} \tag{2.8}$$

Differentiating with respect to the unknown parameters gives the necessary conditions for a global maximum of the likelihood function

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} - \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \tag{2.9}$$

$$\frac{\partial \ell}{\partial \theta_r} = \frac{1}{2}\left\{ (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_r}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_r}\right)\right\} = 0, \tag{2.10}$$

where tr(.) is the matrix trace. For simplicity, assume that $\mathbf{X}$ is of full (column) rank. Let $(\hat{\boldsymbol{\beta}}_{ML}, \hat{\boldsymbol{\theta}}_{ML})$ be a maximizer. From 2.9 we obtain the following closed-form expression for $\hat{\boldsymbol{\beta}}_{ML}$

$$\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{2.11}$$

where $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}}_{ML})$ is the marginal covariance matrix based on the estimated variance components $\hat{\boldsymbol{\theta}}_{ML}$. Now, let

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}. \tag{2.12}$$

Note that using 2.11, we have

$$\mathbf{P}\mathbf{y} = \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \tag{2.13}$$

that is, $\mathbf{P}$ removes the fixed effects from $\mathbf{y}$. Combining 2.9 and 2.10 one finds that the ML estimator for $\boldsymbol{\theta}$ satisfies

$$\mathbf{y}^T\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta_r}\mathbf{P}\mathbf{y} = \text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_r}\right), \tag{2.14}$$

which can be solved independently of $\hat{\boldsymbol{\beta}}_{ML}$.

For the general case of arbitrary covariance structure, no closed-form maximum likelihood solutions exists. Instead, numerical optimization methods such as Newton–Raphson, Fisher scoring or expectation-maximization (EM) need to be used. In the case of independent observations, however, estimation is straightforward:

Example: Linear regression

If we assume no random effects, such that $\mathbf{V} = \sigma^2 \boldsymbol{\mathcal{I}}_N$, one obtains the well-known linear least squares estimator

$$\hat{\boldsymbol{\beta}}_{ML} = (\hat{\sigma}^2_{ML}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\sigma}^{-2}_{ML}\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.15}$$

which can be computed independently of the variance estimate. From 2.10 it then follows that the maximum likelihood estimate for σ^2 is

$$\hat{\sigma}^2_{ML} = \frac{1}{N}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}). \tag{2.16}$$

2.1.4 Restricted maximum likelihood estimation

Under suitable conditions, the ML estimator is both consistent and asymptotically normal. However, the estimator for variance components is biased and systematically underestimates the population value. As an illustrative example, consider the sample variance estimator

$$\tilde{s}^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n - \bar{y})^2,\ \bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n, \tag{2.17}$$

which can be derived as the ML estimate of the variance for a sample $(y_1, \ldots, y_N)$ from a univariate Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with unknown μ and σ^2. It is easy to show that

$$\mathbb{E}[\tilde{s}^2] = \frac{N-1}{N}\sigma^2, \tag{2.18}$$

and thus $\text{Bias}(\tilde{s}^2, \sigma^2) = \mathbb{E}[\tilde{s}^2 - \sigma^2] < 0$. This suggests a simple modification, known as Bessel's correction, to construct an unbiased estimator:

$$s^2 = \frac{N}{N-1}\tilde{s}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \bar{y})^2. \tag{2.19}$$

Intuitively, Bessel's correction adjusts for the degrees of freedom of the residual vector $(y_1 - \bar{y}, \ldots, y_N - \bar{y})$: While there are N independent observations, there are only $N-1$ independent residuals, as

$$\sum_{n=1}^{N}(y_n - \bar{y}) = 0. \tag{2.20}$$

Patterson and Thompson [104] derived a general procedure, later termed restricted maximum likelihood estimation (ReML), to adjust for bias in the estimation of variance components in

linear models with Gaussian observation noise. Their proposed approach involves finding linear transformations of $\mathbf{y}$, referred to as error contrasts, which preserve all the information on variance components while eliminating fixed effect parameters.

Returning to the model in eq. 2.3, the ReML method replaces $\mathbf{y}$ with $\mathbf{z} = \mathbf{A}^T\mathbf{y}$, where $\mathbf{A} \in \mathbb{R}^{N\times(N-\text{rank}(\mathbf{X}))}$ is chosen such that

$$\text{rank}(A) = N - \text{rank}(X), \;\; \mathbf{A}^T\mathbf{X} = \mathbf{0}. \tag{2.21}$$

Therefore, $\mathbb{E}[\mathbf{z}] = \mathbf{A}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$. As $\mathbf{y}$ is normally distributed, it follows that

$$\mathbf{A}^T\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^T\mathbf{V}\mathbf{A}), \tag{2.22}$$

where again $\mathbf{V} = \mathbf{Z}\mathbf{K}\mathbf{Z}^T + \sigma^2\boldsymbol{\mathcal{I}}_{\boldsymbol{N}}$ is the covariance matrix of the marginal model 2.6 with unknown parameters $\boldsymbol{\theta}$. The ReML estimator is defined as the maximum likelihood solution for the transformed variable $\mathbf{z}$. The log-likelihood, here also known as restricted log-likelihood, is

$$\ell_{ReML}(\boldsymbol{\theta}; \mathbf{y}) = -\frac{1}{2}\log|\mathbf{A}^T\mathbf{V}\mathbf{A}| - \frac{1}{2}\mathbf{z}^T(\mathbf{A}^T\mathbf{V}\mathbf{A})^{-1}\mathbf{z} + \text{const.} \tag{2.23}$$

Computing the derivative of the restricted log-likelihood and setting it to zero, one can show that the maximizer needs to satisfy

$$\mathbf{y}^T\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_r}\mathbf{P}\mathbf{y} = \text{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_r}\right), \tag{2.24}$$

where $\mathbf{P}$ is defined in 2.13 [102]. Note that the above equation is independent of $\mathbf{A}$ and so the ReML estimator does not depend on the choice of $\mathbf{A}$ (which is not unique).

Example: Linear regression

Consider again the linear model for independent observations in 2.2. The ReML estimator is

$$\hat{\sigma}^2_{ReML} = \frac{1}{N - \text{rank}(\mathbf{X})}(\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})^T(\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \tag{2.25}$$

$$= \frac{N}{N - \text{rank}(\mathbf{X})}\hat{\sigma}^2_{ML} \tag{2.26}$$

Similar to Bessel's correction, the ReML estimator replaces N by $N - \text{rank}(\mathbf{X})$ in eq. 2.16 to account for a loss in the degrees of freedom after estimating the fixed effect vector $\boldsymbol{\beta}$.

2.1.5 Likelihood-based testing

The previous sections have introduced point estimation in linear mixed models using maximum likelihood and restricted maximum likelihood estimation. Another important type of statistical inference for linear mixed models is hypothesis testing. Following the expository note by A. Buse [105], we will review three different tests commonly employed in the context of linear mixed models and compare these within the framework of maximum likelihood methods: The likelihood-ratio test, the score test and the Wald test. All of these tests aim to assess the statistical evidence for a particular parameter restriction. For instance, an association test for fixed effect d compares the null hypothesis

$$H_0 : \beta_d = 0, \tag{2.27}$$

to the alternative

$$H_1 : \beta_d \neq 0. \tag{2.28}$$

Similarly, we might be interested in testing for restrictions of the variance component vector $\boldsymbol{\theta}$. All three tests in this section leverage the likelihood function to derive a test statistic, T, that summarizes the data. Under some regularity conditions, all statistics follow χ^2 (chi-squared) distributions under the null asymptotically. Given the observed value t of the test statistic T, a p-value $P = p(T \geq t \mid H_0)$ may then be computed to assess whether or not H_0 can be rejected at a chosen significance level, e.g. $P < 0.05$.

Likelihood-ratio test

Let $\boldsymbol{\phi}$ denote all of the unknown model parameters and let $\mathbf{r}(\boldsymbol{\phi}) = \mathbf{0}$ be a set of g functional restrictions on the parameter vector. Furthermore, let $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\phi}}_0$ be the maximum likelihood estimates for the unrestricted and restricted models, respectively. The likelihood-ratio test statistic directly compares the likelihood of the data for the two models

$$\text{LR} = 2(\ell(\hat{\boldsymbol{\phi}}) - \ell(\hat{\boldsymbol{\phi}}_0)). \tag{2.29}$$

For the case of closed-form likelihood functions, as is the case for linear mixed models, LR is straightforward to implement. Furthermore, under some regularity conditions and assuming the null hypothesis is true, the test statistic follows a chi-squared distribution with g degrees of freedom asymptotically (Wilks' theorem)

$$\text{LR} \sim \chi^2(g). \tag{2.30}$$

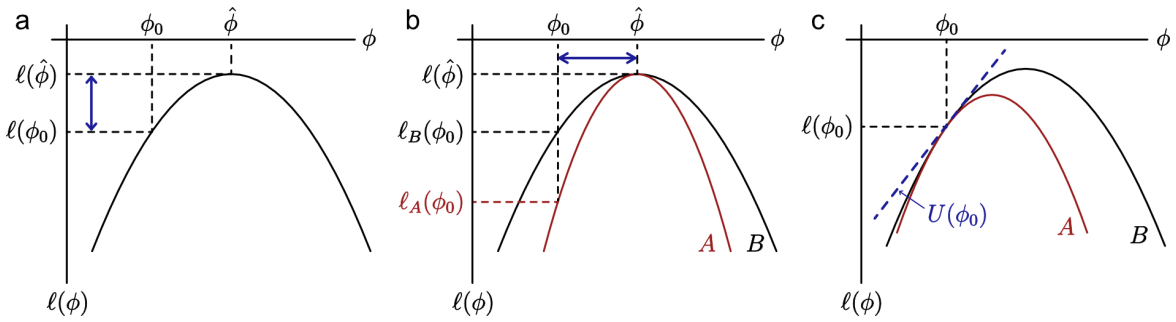

Figure 2.1. Geometric intuition for likelihood-ratio test, Wald test and score test for the hypothesis $H_0 : \phi = \phi_0$ vs. $H_1 : \phi \neq \phi_0$. Key elements of the test statistics are highlighted in blue. (a) The likelihood-ratio test directly compares the log-likelihoods $\ell(\phi_0)$ and $\ell(\hat{\phi})$ for null parameter and maximum likelihood estimate $\hat{\phi}$, respectively. (b) The Wald test assesses the difference between the maximum likelihood estimate $\hat{\phi}$ and null parameter ϕ_0, scaled by the curvature of the log-likelihood at $\hat{\phi}$. Shown are the log-likelihoods for two datasets A and B with identical global optimum but varying curvature. (c) The score test statistic evaluates the derivative of the log-likelihood at ϕ_0, weighted by the inverse curvature. Shown are the log-likelihoods for two datasets A and B with identical derivative at ϕ_0 but varying curvature. Adapted from [105].

Fig. 2.1a depicts a simple example, where we assume that ϕ is unidimensional and the log-likelihood is parabolic.

Wald test

Note that in **Fig.** 2.1a, the difference in likelihood may be expressed as a function of both the distance between $\hat{\phi}$ and ϕ_0, as well as the curvature of the log-likelihood function. For the same distance between the restricted and unrestricted parameters, a greater curvature will amplify the difference in their associated log-likelihoods (**Fig.** 2.1b). The Wald test statistic [106] is based on exactly this intuition. In the unidimensional case with linear restriction, $r(\boldsymbol{\phi}) = \phi - \phi_0$, the statistic is defined as

$$\mathrm{W} = (\hat{\phi} - \phi_0)^2 C(\hat{\phi}), \tag{2.31}$$

where $C(\hat{\phi})$ is the absolute value of the second derivative $\frac{\partial^2 \ell}{\partial \phi^2}\big|_{\phi=\hat{\phi}}$. Typically, $C(\hat{\phi})$ is replaced by its expectation $I(\hat{\phi}) = \mathbb{E}[C(\hat{\phi})]$, known as the Fisher information. However, as $C(\hat{\phi})$ is a consistent estimator of $I(\hat{\phi})$, both variants are asymptotically equivalent. More

generally, for arbitrary $\boldsymbol{\phi}$ and $\mathbf{r}$,

$$\mathrm{W} = \mathbf{r}(\hat{\boldsymbol{\phi}})^T(\mathbf{R}(\hat{\boldsymbol{\phi}})\mathbf{I}(\hat{\boldsymbol{\phi}})^{-1}\mathbf{R}(\hat{\boldsymbol{\phi}})^T)^{-1}\mathbf{r}(\hat{\boldsymbol{\phi}}), \tag{2.32}$$

where $\mathbf{R}(\hat{\boldsymbol{\phi}})$ is the matrix of partial derivatives (Jacobi matrix) of $\mathbf{r}$ evaluated at $\hat{\boldsymbol{\phi}}$. As the LR statistic, W follows a $\chi^2(g)$ distribution asymptotically under suitable assumptions if the null hypothesis holds true. Note that in contrast to the likelihood-ratio test, which requires fitting both the restricted and unrestricted models, the Wald test is solely based on the unrestricted estimate $\hat{\boldsymbol{\phi}}$.

Score test

In many cases the null model is of considerably lower complexity than the alternative. While the Wald test eliminates the computational burden of fitting the null model, an even more appealing approach would therefore be to only estimate the restricted parameter $\boldsymbol{\phi}_0$. Consider again the unidimensional case with linear restriction. Since the unrestricted estimate $\hat{\phi}$ maximizes the likelihood, we have

$$\mathcal{U}(\hat{\phi}) = \left.\frac{\partial \ell}{\partial \phi}\right|_{\phi=\hat{\phi}} = 0, \tag{2.33}$$

where $\mathcal{U}$ is known as the score function. If the null hypothesis is true, the parameter estimates $\hat{\phi}_0$ for the restricted model will be close to the unrestricted estimates $\hat{\phi}$. The score test [107] (**Fig.**2.1c), also known as Lagrange multiplier test, is based on the idea that a departure from the null hypothesis should result in a deviation of $\mathcal{U}(\hat{\phi})$ from zero (assuming that the null is sufficiently close to the global maximum). As in the case of the Wald test, the relationship between derivative and the likelihood depends on the curvature of the log-likelihood function. In particular, the greater the curvature, the closer $\hat{\phi}_0$ will be to the maximizer $\hat{\phi}$. The score test statistic is

$$\mathrm{S} = \mathcal{U}(\hat{\phi}_0)^2 C(\hat{\phi}_0)^{-1}, \tag{2.34}$$

where again, $C(\hat{\phi}_0)^{-1}$ is typically replaced by $I(\hat{\phi}_0)^{-1}$. The analog for the general case is

$$\mathrm{S} = \boldsymbol{\mathcal{U}}(\hat{\boldsymbol{\phi}}_0)^T\mathbf{I}(\hat{\boldsymbol{\phi}}_0)^{-1}\boldsymbol{\mathcal{U}}(\hat{\boldsymbol{\phi}}_0), \tag{2.35}$$

where, once again, the limiting distribution is $\chi^2(g)$ under the null.

Example: Testing for fixed effects in LMMs

As an example, we compare the likelihood-ratio, Wald and score statistics when testing for the fixed effects in the linear mixed model framework. Let again $\mathbf{V} = \mathbf{Z}\mathbf{K}\mathbf{Z}^T + \sigma^2\boldsymbol{\mathcal{I}}_N$ be the covariance matrix of the marginal model 2.6, such that

$$\mathbf{y} = \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}), \tag{2.36}$$

and for simplicity we assume $\mathbf{V}$ is known. Define the null and alternative hypotheses as

$$H_0 : \beta_d = 0 \text{ vs. } H_1 : \beta_d \neq 0. \tag{2.37}$$

Using the notation from the previous subsections, we can also express the null hypothesis using a function of restrictions $H_0 : r(\boldsymbol{\beta}) = 0$, where $r(\boldsymbol{\beta}) = \boldsymbol{\rho}^T\boldsymbol{\beta}$ and $\rho_j = 1$ if $j = d$ and $\rho_j = 0$ otherwise. That is, $\boldsymbol{\rho}$ is the vector of first derivatives $\rho_j = \frac{\partial r}{\partial \beta_j}$.

We previously derived the log-likelihood function and first order derivatives in section 2.1.3. Using these results, we find that

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -(\mathbf{X}^T\mathbf{V}\mathbf{X}), \tag{2.38}$$

that is, for known $\mathbf{V}$, the Hessian is constant and curvature and information matrix are equivalent. Recall that the unrestricted maximum likelihood estimate for the fixed effect vector is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}. \tag{2.39}$$

To obtain the restricted estimator, we solve the constrained optimization problem using the method of Lagrange multipliers

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} + \lambda\mathbf{r} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} - \mathbf{X}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\rho} = \mathbf{0} \tag{2.40}$$

After some rearrangements one obtains the solution

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}} + \lambda\mathbf{A}^{-1}\boldsymbol{\rho}, \tag{2.41}$$

with $\mathbf{A} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$ and

$$\lambda = \boldsymbol{\rho}^T\hat{\boldsymbol{\beta}}(\boldsymbol{\rho}^T\mathbf{A}^{-1}\boldsymbol{\rho})^{-1}. \tag{2.42}$$

Now, let the residual vectors for the null and alternative models be denoted by $\mathbf{u}_0 = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0$ and $\mathbf{u} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_1$. Note, that

$$\mathbf{u}_0 = \mathbf{u} - \lambda\mathbf{X}\mathbf{A}^{-1}\boldsymbol{\rho}, \tag{2.43}$$

and thus

$$\ell(\hat{\boldsymbol{\beta}}_0) = -\frac{1}{2}\mathbf{u}_0^T\mathbf{V}^{-1}\mathbf{u}_0 + \text{const.} \tag{2.44}$$

$$= -\frac{1}{2}\mathbf{u}^T\mathbf{V}^{-1}\mathbf{u}^T + (\lambda\mathbf{X}\mathbf{A}^{-1}\boldsymbol{\rho})^T\mathbf{V}^{-1}\mathbf{u} - \frac{1}{2}(\lambda\mathbf{X}\mathbf{A}^{-1}\boldsymbol{\rho})^T\mathbf{V}^{-1}(\lambda\mathbf{X}\mathbf{A}^{-1}\boldsymbol{\rho}) + \text{const.} \tag{2.45}$$

Using the definition of $\hat{\boldsymbol{\beta}}$, it is easy to show that $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{u} = \mathbf{0}$. Therefore,

$$\ell(\hat{\boldsymbol{\beta}}_0) = -\frac{1}{2}\mathbf{u}^T\mathbf{V}^{-1}\mathbf{u}^T - \frac{\lambda^2}{2}\boldsymbol{\rho}^T\mathbf{A}^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\mathbf{A}^{-1}\boldsymbol{\rho} + \text{const.} \tag{2.46}$$

$$= -\frac{1}{2}\mathbf{u}^T\mathbf{V}^{-1}\mathbf{u}^T - \frac{1}{2}(\boldsymbol{\rho}^T\hat{\boldsymbol{\beta}})^2(\boldsymbol{\rho}^T\mathbf{A}^{-1}\boldsymbol{\rho})^{-1} + \text{const.}, \tag{2.47}$$

where the last equality follows from the definition of $\mathbf{A}$ and λ. Using this identity, we obtain the difference of log-likelihoods as

$$\text{LR} = 2[\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_0)] = 2[-\frac{1}{2}\mathbf{u}^T\mathbf{V}^{-1}\mathbf{u} + \frac{1}{2}\mathbf{u}_0^T\mathbf{V}^{-1}\mathbf{u}_0] \tag{2.48}$$

$$= (\boldsymbol{\rho}^T\hat{\boldsymbol{\beta}})^2(\boldsymbol{\rho}^T\mathbf{A}^{-1}\boldsymbol{\rho})^{-1}. \tag{2.49}$$

Combining the definition of the Wald test statistic 2.32 and eq. 2.38 we also find that

$$\text{W} = (\boldsymbol{\rho}^T\hat{\boldsymbol{\beta}})^2(\boldsymbol{\rho}^T\mathbf{A}^{-1}\boldsymbol{\rho})^{-1} = \text{LR}. \tag{2.50}$$

For evaluating the score statistic, note that the score function is

$$\boldsymbol{\mathcal{U}}(\hat{\boldsymbol{\beta}}_0) = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} - \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}(\hat{\boldsymbol{\beta}} + \lambda\mathbf{A}^{-1}\boldsymbol{\rho}) \tag{2.51}$$

$$= -\lambda\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\mathbf{A}^{-1}\boldsymbol{\rho} = -\lambda\boldsymbol{\rho}, \tag{2.52}$$

as $\hat{\boldsymbol{\beta}}$ is a root of $\boldsymbol{\mathcal{U}}$. Using the definition of the score test 2.35 and eq. 2.38, we have

$$\text{S} = \lambda^2\boldsymbol{\rho}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\rho} = (\boldsymbol{\rho}^T\hat{\boldsymbol{\beta}})^2(\boldsymbol{\rho}^T\mathbf{A}^{-1}\boldsymbol{\rho})^{-1} = \text{LR}. \tag{2.53}$$

Choosing a test statistic

The previous sections introduced a geometric interpretation of the likelihood ratio, Wald and score test statistics. In this context, the Wald and score tests can be interpreted as approximations to the likelihood ratio test. We have seen that all three tests are numerically equivalent when testing for fixed effects in linear mixed models. In fact, this result holds for any tests of linear restrictions if the log-likelihood is quadratic [105]. More generally, it can

be shown that W $\geq$ LR $\geq$ S. So should one test be preferred over the other? Under some regularity conditions, all three test are asymptotically equivalent under the null and follow a $\chi^2(g)$ distribution, where g is the number of parameter restrictions. Additionally, all three tests are asymptotically locally most powerful tests, that is, they are most powerful for small deviations from the null hypothesis [108]. When considering computational scalability, the score and Wald tests are often more appealing than the likelihood-ratio test. The score test in particular only requires fitting the restricted null model, leading to considerable speedups when the score statistic itself can be computed efficiently.

An important special case occurs when the null hypothesis places the parameter on the boundary of the parameter space, for instance, when testing if non-negative variance components are different from zero. In this case the regularity conditions are violated and the test statistics no longer follow a χ^2 distribution asymptotically under the null [109]. Self and Liang showed that when the marginal covariance matrix has block-diagonal structure, a limiting distribution for the likelihood ratio statistic may still be derived [109]. Similar extensions are possible for score tests [110]. In the next section we will discuss a score-based test to assess arbitrary covariance matrices in LMMs.

2.1.6 Variance component score tests for linear mixed models

Consider the following linear mixed model in marginal form

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \tau^2\mathbf{K} + \sigma^2\boldsymbol{\mathcal{I}_N}), \tag{2.54}$$

where $\mathbf{K}$ is a known covariance matrix. We want to assess whether or not the y_n co-vary according to $\mathbf{K}$, that is,

$$H_0 : \tau^2 = 0 \text{ vs. } H_1 : \tau^2 > 0. \tag{2.55}$$

Such a test places τ^2 on the boundary of the parameter space under the null. As mentioned in the previous section, this violates the sufficient conditions under which the score or likelihood-ratio test statistics are known to asymptotically follow a $\chi^2(1)$ distribution. However, it is possible to derive a score-based statistic and associated distribution under the

null [111–113]. The score for τ^2 is is (see eq. 2.10),

$$\mathcal{U}_\tau(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \left.\frac{\partial \ell}{\partial \tau^2}\right|_{\tau^2=0,\, \boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \sigma^2=\hat{\sigma}^2} \tag{2.56}$$

$$= \frac{1}{2}\left\{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \hat{\mathbf{V}}_0^{-1}\mathbf{K}\hat{\mathbf{V}}_0^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) - \mathrm{tr}\left(\hat{\mathbf{V}}_0^{-1}\mathbf{K}\right)\right\} \tag{2.57}$$

$$= \frac{1}{2}\mathbf{y}^T\hat{\mathbf{P}}_0\mathbf{K}\hat{\mathbf{P}}_0\mathbf{y} - \frac{1}{2}\mathrm{tr}\left(\hat{\mathbf{V}}_0^{-1}\mathbf{K}\right), \tag{2.58}$$

where $\hat{\mathbf{V}}_0 = \hat{\sigma}^2\boldsymbol{\mathcal{I}}_{\boldsymbol{N}}$ is the estimated marginal covariance under the null $\tau^2 = 0$ and $\hat{\mathbf{P}}_0$ projects out the fixed effects,

$$\hat{\mathbf{P}}_0 = \hat{\mathbf{V}}_0^{-1} - \hat{\mathbf{V}}_0^{-1}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}_0^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}_0^{-1}. \tag{2.59}$$

We choose

$$U_\tau(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \frac{1}{2}\mathbf{y}^T\hat{\mathbf{P}}_0\mathbf{K}\hat{\mathbf{P}}_0\mathbf{y}, \tag{2.60}$$

the first term in 2.58, as a test statistic. Let $(\hat{\boldsymbol{\beta}}_0, \hat{\sigma}_0^2)$ denote the maximum-likelihood estimators under the null model. We will see that if the null hypothesis is true, $\tau^2 = 0$, $U_\tau(\hat{\boldsymbol{\beta}}_0, \hat{\sigma}_0^2)$ is asymptotically distributed according to a linear combination of chi-squared random variables [111]. While such distributions are difficult to evaluate directly, they can be approximated closely, e.g., with Davies method based on numerical inversion of the characteristic function [114] or using moment matching [115].

To derive the asymptotic distribution of $U_\tau(\hat{\boldsymbol{\beta}}_0, \hat{\sigma}_0^2)$, it is sufficient to consider the asymptotic behaviour of $U_\tau(\boldsymbol{\beta}_0, \sigma_0^2)$ under $\tau^2 = 0$, where $\boldsymbol{\beta}_0, \sigma_0^2$ are the true values of the parameters. This is because the maximum-likelihood estimators $\hat{\boldsymbol{\beta}}_0, \hat{\sigma}_0^2$ are consistent estimators of the true values $\boldsymbol{\beta}_0, \sigma_0^2$ under standard regularity conditions. Let $\mathbf{P}_0$ be as in 2.59, but with $\hat{\mathbf{V}}_0$ replaced by $\mathbf{V}_0 = \sigma_0^2\boldsymbol{\mathcal{I}}_N$ for the true parameter σ_0^2. We rewrite the test statistic as

$$U_\tau(\boldsymbol{\beta}_0, \sigma_0^2) = \frac{1}{2}\mathbf{y}^T\mathbf{P}_0\mathbf{V}_0^{1/2}\mathbf{V}_0^{-1/2}\mathbf{K}\mathbf{V}_0^{-1/2}\mathbf{V}_0^{1/2}\mathbf{P}_0\mathbf{y}, \tag{2.61}$$

where $\mathbf{V}_0^{1/2}$ is the matrix square root of $\mathbf{V}_0$. Note that using basic properties of the multivariate normal distribution,

$$\mathbf{V}_0^{1/2}\mathbf{P}_0\mathbf{y} = \mathbf{V}_0^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_N). \tag{2.62}$$

Now, let $\psi_1 \geq \cdots \geq \psi_R$ be the ordered, non-zero eigenvalues of the matrix $\frac{1}{2}\mathbf{V}_0^{-1/2}\mathbf{K}\mathbf{V}_0^{-1/2}$ and $\boldsymbol{\Psi} = \mathrm{diag}(\psi_i)$. Furthermore, let $\mathbf{H}$ be the $R \times N$ matrix of associated eigenvectors, such

that $\mathbf{V}_0^{-1/2}\mathbf{K}\mathbf{V}_0^{-1/2} = \mathbf{H}^T\mathbf{\Psi}\mathbf{H}$ and $\mathbf{H}\mathbf{H}^T = \boldsymbol{\mathcal{I}}_R$. Then,

$$U_\tau(\boldsymbol{\beta}_0, \sigma_0^2) = \frac{1}{2}\mathbf{y}^T\mathbf{P}_0\mathbf{K}\mathbf{P}_0\mathbf{y} = \mathbf{z}^T\mathbf{\Psi}\mathbf{z} = \sum_{i=1}^{R}\psi_i z_i^2 \tag{2.63}$$

where $\mathbf{z} = \mathbf{H}\mathbf{V}_0^{1/2}\mathbf{P}_0\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_R)$. That is, the z_i are independent standard normal variables and thus each z_i^2 follows a chi-squared distribution with one degree of freedom, $z_i^2 \sim \chi^2(1)$.

Note that the above derivation also generalizes to the case of heteroscedastic observation noise and models with additional variance components. That is, $\hat{\mathbf{V}}_0$ may be an arbitrary positive semi-definite covariance matrix.

2.1.7 Variance decomposition

In many cases one is interested in evaluating the individual contributions of different variance components in a Gaussian model to determine the most important drivers of variation of the output variable $\mathbf{y}$. Suppose, $\mathbf{y}$ is distributed as

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \sum_i \sigma_i^2\mathbf{K}_i) \tag{2.64}$$

with known covariance matrices $\mathbf{K}_i \in \mathbb{R}^{N\times N}$ and parameters σ_i^2. Equivalently, one may write $\mathbf{y} = \sum_i \sigma_i\mathbf{z}_i$, where $\mathbf{z_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_i)$ are independent Gaussian random variables. We define the variance explained by each term $\sigma_i^2\mathbf{K_i}$ as the expected sample variance of the vector $\sigma_i\mathbf{z}_i$. [116].

Let $\bar{z}_i = \frac{1}{N}\mathbf{z}_i^T\mathbf{1}$ be the sample mean. The (unbiased) sample variance is [117]

$$\begin{aligned}
\mathrm{Var}_S(\mathbf{z}_i) &= \frac{1}{N-1}(\mathbf{z}_i - \bar{z}_i\mathbf{1})^T(\mathbf{z}_i - \bar{z}_i\mathbf{1}) && (2.65)\\
&= \frac{1}{N-1}(\mathbf{z}_i^T\mathbf{z}_i - N\bar{z}_i^2) && (2.66)\\
&= \frac{1}{N-1}\mathrm{tr}\left(\mathbf{z}_i^T\mathbf{z}_i - \frac{1}{N}\mathbf{z}_i^T\mathbf{z}_i\mathbf{1}\mathbf{1}^T\right). && (2.67)
\end{aligned}$$

Taking the expectation w.r.t. $\mathbf{z}_i$ gives

$$\mathbb{E}[\mathrm{Var}_S(\mathbf{z}_i)] = \frac{1}{N-1}\mathrm{tr}(\mathbf{K}\mathbf{P}), \tag{2.68}$$

where

$$\mathbf{P} = \boldsymbol{\mathcal{I}}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^T. \tag{2.69}$$

projects out the sample mean. In particular, if $\mathbf{K}_i$ has been scaled such that $\mathbb{E}[\text{Var}_S(\mathbf{z}_i)] = 1$, the variance explained by the term $\sigma_i^2\mathbf{K}_i$ is simply

$$\mathbb{E}[\text{Var}_S(\sigma_i\mathbf{z}_i)] = \sigma_i^2. \tag{2.70}$$

2.1.8 Handling non-Gaussian responses

The linear mixed model framework as introduced in section 2.1.2 models both random effects and residual noise as Gaussian random variables. Under this assumption, integrating out the random effects is straightforward and the marginal model 2.6 is again Gaussian. As a result, likelihood-based statistical inference techniques can be readily applied. In some cases even closed-form solutions to the maximum-likelihood problem may be available; otherwise estimators can be computed efficiently using second-order numerical optimization methods. In practice, however, the assumption of normality is frequently violated, in particular when working with discrete response variables such as categorical or count data. If the distribution of the response variables is significantly skewed, heavy-tailed, or has multiple modes, a Gaussian model may fail to accurately represent the underlying distribution, leading to poor predictive performance and reduced statistical power. One solution to this problem is to use suitable preprocessing techniques to bring the response variables closer to a Gaussian distribution. Commonly, variance-stabilizing transformations such as the Anscombe [118], Box-Cox [119], or simple logarithmic transforms are applied to decouple the mean and variance of the data. Subsequently, Gaussian models may be applied. While this analysis strategy benefits from all the above-mentioned advantages of Gaussian likelihoods, it greatly depends on the potency of the chosen preprocessing methods to successfully 'gaussianize' the data. Alternatively, one may forgo the assumption of Gaussian residuals and look towards alternative distributions. In this section we will introduce generalized error models for classical linear regression and linear mixed models as well as relevant inference techniques.

Generalized likelihood models

Consider again the linear mixed model in its general form as introduced in section 2.1.2,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{2.71}$$

where $\boldsymbol{\beta}, \mathbf{u}$ are fixed and random effect vectors for the associated covariate matrices $\mathbf{X}, \mathbf{Z}$ and $\boldsymbol{\epsilon}$ captures zero-mean Gaussian observation noise. One can also view this model in a hierarchical manner, separating random and fixed effects from residual errors,

1. Latent Gaussian linear model: $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{K}\mathbf{Z}^T)$.

2. Gaussian likelihood model: $\mathbf{y} \mid \mathbf{u} \sim \mathcal{N}(\boldsymbol{\eta}, \sigma^2 \boldsymbol{\mathcal{I}}_N)$.

That is, the latent Gaussian variable $\boldsymbol{\eta}$ is mapped to the mean of a Gaussian likelihood model, $\mathbf{y} \mid \mathbf{u}$, and the y_n are conditionally independent given $\mathbf{u}$. We will denote the mean by $\boldsymbol{\mu} = \mathbb{E}[\mathbf{y} \mid \mathbf{u}]$ and make this mapping explicit, writing

$$\eta_n = g(\mu_n) \tag{2.72}$$

where g is known as a link function and in this case simply corresponds to the identity, $\eta_n = \mu_n$.

Generalized linear mixed models (GLMMs) extend the linear mixed model framework above by allowing for non-Gaussian likelihood models $\mathbf{y} \mid \mathbf{u}$, using suitable monotonic differentiable link functions g. We will assume the y_n are conditionally independent with means $\mathbb{E}[y_n \mid \mathbf{u}]$ and the conditional variance can be written as

$$\text{Var}(y_n \mid \mathbf{u}) = a_n(\phi) v(\mu_n), \tag{2.73}$$

where ϕ is an additional dispersion parameter and $a_i(\cdot)$ a known function and $v(\cdot)$ a known variance function. In particular, $\text{Var}(y_n \mid \mathbf{u})$ depends only on the n-th component of $\boldsymbol{\mu}$.

Example: Binomial response variable

Consider discrete response variables under a Binomial model, where y_n denotes the proportion of successes in d_i trials,

$$y_n \mid \mathbf{u} \sim \text{Bin}(d_n, \mu_n)/d_n. \tag{2.74}$$

Under this model we have

$$\text{Var}(y_n \mid \mathbf{u}) = \frac{1}{d_n} \mu_n (1 - \mu_n), \tag{2.75}$$

that is, $a_n(\phi) = a_n = 1/d_n$ and $v(\mu_n) = \mu_n(1 - \mu_n)$. Suppose $\mu_n = g^{-1}(\eta_n) = g^{-1}(\mathbf{x}_n^T \boldsymbol{\beta} + \mathbf{z}_n^T \mathbf{u})$, where g^{-1} denotes the inverse function of g. Then g^{-1} should satisfy $0 < g^{-1}(\eta_n) < 1$. Here, commonly used link functions for the binomial distribution are [120]

$$\eta_n = \log \frac{\mu_n}{1 - \mu_n} \qquad \text{(logit)} \tag{2.76}$$

$$\eta_n = \Phi^{-1}(\mu_n) \qquad \text{(probit)} \tag{2.77}$$

$$\eta_n = \log(-\log(1 - \mu_n)) \qquad \text{(complementary log-log)} \tag{2.78}$$

where $\Phi(\cdot)$ is the cumulative distribution function (c.d.f) of the standard normal distribution. The logit function can be obtained as the inverse c.d.f. of a logistic distribution and evaluates the log-odds of observing a value of 1 vs. 0. It is widely used due to its simple form and and interpretability: an increase in a particular fixed effect covariate multiplicatively scales the odds of the given outcome at a constant rate. The complementary log-log corresponds to the inverse c.d.f of the Gumbel or type-I generalized extreme value distribution. Unlike the logit or probit links functions, it is asymmetric around zero. **Fig.** 2.2 (left) shows all three link functions. The logistic and Gumbel distributions associated with the logit and complementary log-log link functions are not standardized to mean zero and variance one. For a better comparison, the link functions can be adjusted as shown in **Fig.** 2.2 (right), such that all links correspond to the inverse c.d.f of standardized distributions. In practice, the difference in empirical fit between these link functions is often negligible.

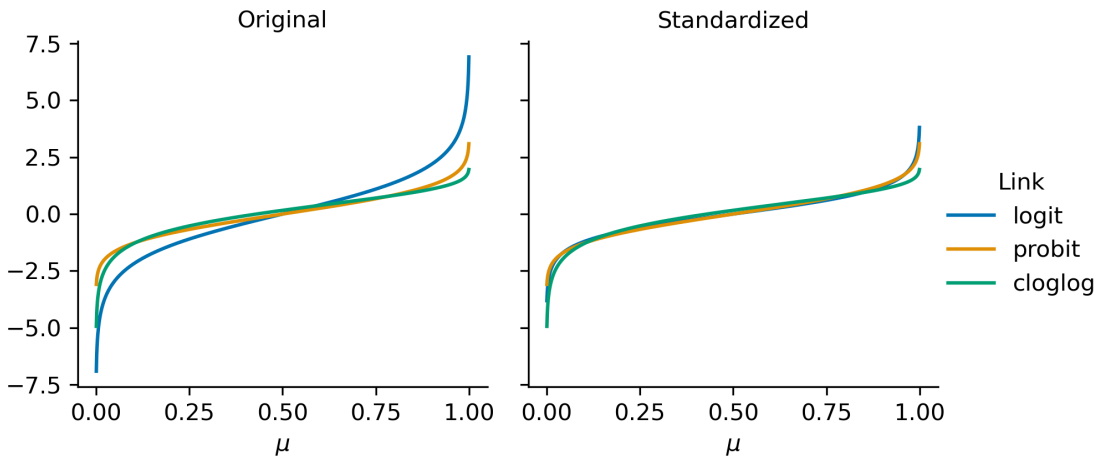

Figure 2.2. Left: Comparison of the logit, probit and complementary log-log link functions, corresponding to the inverse c.d.f of the logistic, standard normal and Gumbel distributions, respectively. Right: All link functions have been adjusted such that the associated distributions have mean zero and variance one.

The exponential family

Many theoretical results additionally require that the likelihood is an exponential family distribution, taking the form [120]

$$f(y_n) = \exp\left\{\frac{y_n\xi_n - b(\xi_n)}{a_n(\phi)} + c_i(y_n, \phi)\right\} \tag{2.79}$$

	Normal	Poisson	Binomial	Gamma
Notation	$\mathcal{N}(\mu, \sigma^2)$	Poisson(μ)	Bin$(d, \mu)/d$	Gamma(μ, ν)
Range of y	$(-\infty, \infty)$	$\mathbb{N}_0$	$\{k/d : k \in \mathbb{N}_0, k \leq d\}$	$(0, \infty)$
Dispersion ϕ	σ^2	1	$1/d$	ν^{-1}
$b(\xi)$	$\xi^2/2$	$\exp(\xi)$	$\log(1 + \exp(\xi))$	$-\log(-\xi)$
Canonical link	$\eta = \mu$	$\eta = \log(\mu)$	$\eta = \log(\frac{\mu}{1-\mu})$	$\eta = 1/\mu^2$

Table 2.1. Properties of common exponential family distributions.

for some functions $a_n(\cdot), b(\cdot)$ and $c_i(\cdot)$ and dispersion parameter ϕ. This family encompasses many common continuous and discrete distributions such as the normal, Poisson, binomial, and Gamma distributions. One can show that for members of the exponential family (see, e.g., [120]),

$$\mathbb{E}[y_n] = \mu_n = b'(\xi_n). \tag{2.80}$$

Therefore, if we choose the link function $g = h^{-1}$, where $h(\cdot) = b'(\cdot)$, we have

$$\xi_n = \eta_n. \tag{2.81}$$

In this case, g is known as the canonical link function. **Table** 2.1 shows the canonical links for some common univariate exponential family distributions. For instance, the logit is the canonical link function for the binomial and Bernoulli distributions. These link functions often turn out to be mathematically convenient, though there is no general reason to prefer canonical links from a statistical point of view. Note that while some authors restrict generalized linear mixed model likelihoods to the exponential family, we will not make this assumption here unless otherwise stated.

Fitting generalized linear models

In this section we will outline a simple iterative procedure for obtaining estimators for the fixed effect coefficients in non-Gaussian likelihood models without latent random effects, that is,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \quad \mathbb{E}[\mathbf{y}] = g^{-1}(\boldsymbol{\eta}), \tag{2.82}$$

where g^{-1} is applied element-wise. These models are a special case of the GLMM framework known as generalized linear models (GLMs). As before, we assume the y_n are independent with $\text{Var}(y_n \mid \mathbf{u}) = a_n(\phi)v(\mu_n)$.

Let $\hat{\boldsymbol{\eta}}^{(t)}$ be an estimate of $\boldsymbol{\eta}$ at iteration t and let $\hat{\boldsymbol{\mu}}^{(t)} = g^{-1}(\hat{\boldsymbol{\eta}}^{(t)})$. At each step, the algorithm calculates a working dependent variable based on a first-order Taylor approximation to the link function

$$\tilde{y}_n^{(t)} = \hat{\eta}_n^{(t)} + (y_n - \hat{\mu}_n^{(t)})g'(\hat{\mu}_n^{(t)}), \quad n = 1, \ldots, N, \tag{2.83}$$

where g' denotes the derivative of g. Under a linear transformation of $\mathbf{y}$, evaluating the mean and variance of the working variable is straightforward. We have

$$\text{Var}(\tilde{y}_n^{(t)}) = a_n(\phi)v(\mu_n)g'(\hat{\mu}_n^{(t)})^2, \quad n = 1, \ldots, N, \tag{2.84}$$

and, assuming that $\hat{\boldsymbol{\mu}}^{(t)}$ is an unbiased estimator for $\boldsymbol{\mu}$,

$$\mathbb{E}[\tilde{y}^{(t)}] = \hat{\boldsymbol{\eta}}^{(t)}. \tag{2.85}$$

To obtain an updated estimate of the fixed effect coefficients $\boldsymbol{\beta}$, we now assume that the working variable $\tilde{\mathbf{y}}^{(t)}$ can be modeled using a normal distribution,

$$\tilde{\mathbf{y}}^{(t)} \sim \mathcal{N}(\hat{\boldsymbol{\eta}}^{(t)}, \mathbf{V}^{(t)}), \tag{2.86}$$

where $\mathbf{V}^{(t)} = \text{diag}(a_n(\phi)v(\mu_n)g'(\hat{\mu}_n^{(t)})^2)$ and compute the maximum likelihood estimates by solving the associated weighted least squares problem

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\mathbf{X}^T\hat{\mathbf{W}}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}^{(t)}\mathbf{y}, \tag{2.87}$$

where $\mathbf{W}^{(t)}$ is the inverse of $\mathbf{V}^{(t)}$. The resulting algorithm is known as iterated weighted least squares (IWLS): Starting from the initial estimate $\boldsymbol{\mu}^{(0)} = \mathbf{y}$, this method alternates between updating the coefficient vector and working dependent variable until convergence.

IWLS uses a linear approximation to the link function to map the original response variables to the latent space of the linear predictor $\mathbf{X}\boldsymbol{\beta}$ and approximates the distribution of the transformed variable to be Gaussian. While this appears to be a rather crude approximation, it can be shown that for exponential family likelihood models, the resulting parameter updates are equivalent to Fisher scoring, a variant of Newton's method [120]. Let

$$\boldsymbol{\mathcal{U}}(\hat{\boldsymbol{\beta}}) = \left.\frac{\partial \ell}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \tag{2.88}$$

be the score function for $\boldsymbol{\beta}$ and let $\mathbf{C}(\hat{\boldsymbol{\beta}})$ be the negative Hessian of ℓ,

$$\mathbf{C}(\hat{\boldsymbol{\beta}})_{ds} = -\left.\frac{\partial \ell^2}{\partial \beta_d \partial \beta_s}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \tag{2.89}$$

also known as the observed information matrix. Starting from an initial guess $\hat{\boldsymbol{\beta}}^{(0)}$, Newton's method can be used to approximate a critical point of the log likelihood function. At each iteration, a refined estimate $\hat{\boldsymbol{\beta}}^{(t+1)}$ can be obtained by solving the system of linear equations

$$\mathbf{C}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}) = \boldsymbol{\mathcal{U}}(\hat{\boldsymbol{\beta}}). \tag{2.90}$$

Newton's method usually converges if the initial guess is sufficiently close to the unknown critical point and the Hessian is non-singular [121]. Under certain conditions it can be shown that the rate of convergence is at least quadratic in some neighborhood around the critical point [121]. The Fisher scoring algorithm replaces $\mathbf{C}(\hat{\boldsymbol{\beta}})$ with its expectation, $\mathbf{I}(\hat{\boldsymbol{\beta}}) = \mathbb{E}[\mathbf{H}(\hat{\boldsymbol{\beta}})]$, the Fisher information matrix. While the expected information may be harder to derive, it is often faster to compute. Furthermore, for exponential family models with canonical link function, IWLS, Newton's method and Fisher scoring are all equivalent [120].

Approximate inference for generalized LMMs

All of the inference techniques discussed in the context of linear mixed models make use of the fact that evaluation of the marginal distribution, that is, the (high-dimensional) integral

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \mathbf{u}) p(\mathbf{u}) d\mathbf{u}, \tag{2.91}$$

is straightforward: For normally distributed random effects and Gaussian likelihood, 2.91 can be solved in closed form (see 2.6). In the case of GLMMs, however, where the error model is no longer restricted to the normal distribution, likelihood-based inference is much more challenging. A number of different methods have been developed to address the computational difficulties of parameter estimation in GLMMs, ranging from Monte Carlo expectation-maximization [122] for maximum likelihood estimation to Bayesian methods such as Gibb sampling [102] and variational inference [123]. Here, we will discuss a widely-used method proposed by Breslow and Clayton known as quasi-likelihood estimation [124]. The idea is to construct an objective that shares many of the same properties of a log-likelihood function, while being simpler to evaluate or approximate.

Let **y** be generated from the generalized linear mixed model,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{K}\mathbf{Z}^T), \tag{2.92}$$

$$\mathbb{E}[\mathbf{y} \,|\, \mathbf{u}] = g^{-1}(\boldsymbol{\eta}), \tag{2.93}$$

$$\text{Var}(y_n \,|\, \mathbf{u}) = a_n(\phi)v(\mu_n), \tag{2.94}$$

where g is a known link function, ϕ a dispersion parameter, $a_n(\cdot), v(\cdot)$ are known functions and the y_i are conditionally independent given **u**. Consider the function

$$U_n = U(\mu_n; y_n) = \frac{y_n - \mu_i}{a_n(\phi)v(\mu_n)} \tag{2.95}$$

It is easy to verify that U_n fulfils the following conditions:

$$\mathbb{E}[U_n] = 0, \tag{2.96}$$

$$\text{Var}(U_n) = 1/(a_n(\phi)v(\mu_n)), \tag{2.97}$$

$$-\mathbb{E}[\frac{\partial U_n}{\partial \mu_n}] = 1/(a_n(\phi)v(\mu_n)). \tag{2.98}$$

The above identities are essential properties of a log-likelihood derivative and form the basis of much of the first-order asymptotic theory related to likelihood-based inference [120]. This motivates the use of

$$d_n = -2\int_{y_n}^{\mu_n} \frac{y_n - s}{a_n(\phi)v(s)} ds, \tag{2.99}$$

termed the (quasi-)deviance measure of fit, as a approximate log-likelihood function for μ_n. In fact, if conditioned on **u** the y_i are drawn from an exponential family, in which $a_n(\phi) = \phi/w_n$ for some known weight w_n, one can show that $d_n = 2\phi\{\ell(y_n; y_n) - \ell(\mu_n; y_n)\}$ [120]. Note that $\ell(y_n; y_n)$ is independent of μ_n, and therefore maximizing the log-likelihood $\sum_n \ell(\mu_n; y_n)$ is equivalent to minimizing $\sum_n d_n$ with respect to $\boldsymbol{\mu}$.

An approximate marginal likelihood, known as known as the integrated quasi-likelihood, may be constructed by integrating out the latent random effects **u** [124]:

$$L_Q \propto |\mathbf{K}|^{-1/2} \int \exp\left\{ -\frac{1}{2}\sum_{n=1}^{N} d_n - \frac{1}{2}\mathbf{u}^T\mathbf{K}^{-1}\mathbf{u} \right\} d\mathbf{u}. \tag{2.100}$$

While L_Q can still not be evaluated in closed form for arbitrary non-Gaussian likelihoods, it is amenable to approximation by Laplace's method. Specifically, let

$$q(\mathbf{u}) = \frac{1}{2}\left(\sum_{n=1}^{N} d_n + \mathbf{u}^T\mathbf{K}^{-1}\mathbf{u} \right), \tag{2.101}$$

and suppose $\tilde{\mathbf{u}}$ is a minimizer of $q(\mathbf{u})$ such that $\frac{\partial q}{\partial \mathbf{u}}|_{\mathbf{u}=\tilde{\mathbf{u}}} = \mathbf{0}$, that is,

$$\mathbf{K}^{-1}\mathbf{u} - \sum_{n=1}^{N} \frac{y_n - \mu_n}{a_n(\phi)v(\mu_n)g'(\mu_n)} z_n = \mathbf{0}, \tag{2.102}$$

where $\mu_n = \mathbf{x}_n^T\boldsymbol{\beta} + \mathbf{z}_n^T\mathbf{u}$. Laplace's method uses a second order Taylor expansion of $q(\cdot)$ at the minimizer $\tilde{\mathbf{u}}$,

$$q(\mathbf{u}) \approx q(\tilde{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \tilde{\mathbf{u}})^T \frac{\partial^2 q}{\partial \mathbf{u} \partial \mathbf{u}^T}\bigg|_{\mathbf{u}=\tilde{\mathbf{u}}} (\mathbf{u} - \tilde{\mathbf{u}}), \tag{2.103}$$

to reduce the 2.100 to a Gaussian integral which can easily be evaluated. Specifically, the logarithm of L_Q, termed ℓ_Q is given by

$$\ell_Q \approx -\frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\log\left|\frac{\partial^2 q}{\partial \mathbf{u} \partial \mathbf{u}^T}\bigg|_{\mathbf{u}=\tilde{\mathbf{u}}}\right| - q(\tilde{\mathbf{u}}) + \text{const.} \tag{2.104}$$

Now, it can be shown that [124]

$$\frac{\partial^2 q}{\partial \mathbf{u} \partial \mathbf{u}^T}\bigg|_{\mathbf{u}=\tilde{\mathbf{u}}} = \mathbf{K}^{-1} + \sum_{n=1}^{N} \frac{\mathbf{z}_n \mathbf{z}_n^T}{a_n(\phi)v(\mu_n)g'(\mu_n)^2} + \mathbf{R}, \tag{2.105}$$

where the term $\mathbf{R}$ is zero in expectation and is, in probability, of lower order order than the leading terms as a function of the number of observations N. In fact, for canonical link functions $\mathbf{R}$ equals zero [124]. Note that the denominator of 2.105 corresponds to the IWLS weights 2.84, $\mathbf{W} = \text{diag}((a_n(\phi)v(\mu_n)g'(\mu_n)^2))^{-1}$. Dropping the remainder $\mathbf{R}$, we have

$$\frac{\partial^2 q}{\partial \mathbf{u} \partial \mathbf{u}^T}\bigg|_{\mathbf{u}=\tilde{\mathbf{u}}} \approx \mathbf{Z}^T\mathbf{W}\mathbf{Z} + \mathbf{K}^{-1}. \tag{2.106}$$

Combining the approximations 2.106 and 2.104, we obtain

$$\ell_Q \approx -\frac{1}{2}\left(\log|\boldsymbol{\mathcal{I}}_N + \mathbf{Z}^T\mathbf{W}\mathbf{Z}\mathbf{K}| + \sum_{n=1}^{N} \tilde{d}_n + \tilde{\mathbf{u}}^T\mathbf{K}^{-1}\tilde{\mathbf{u}} \right) + \text{const.}, \tag{2.107}$$

where $\tilde{d}_n$ is d_n with $\mathbf{u} = \tilde{\mathbf{u}}$. Under the additional assumption that the IWLS weights vary slowly as a function of the mean $\boldsymbol{\mu}$, Breslow and Clayton [124] propose to ignore the first term in 2.107 and optimize $\boldsymbol{\beta}$ and $\mathbf{u} = \mathbf{u}(\boldsymbol{\beta})$ to jointly maximize the following objective,

$$\ell_{PQ}(\boldsymbol{\beta}, \mathbf{u}) = -\frac{1}{2}\left(\sum_{n=1}^{N} d_n + \mathbf{u}^T\mathbf{K}^{-1}\mathbf{u} \right) = -q(\mathbf{u}). \tag{2.108}$$

where the subscript PQ stands for penalized quasi-log-likelihood. Computing the roots of the derivatives w.r.t. $\boldsymbol{\beta}$ and $\mathbf{u}$ gives the system of nonlinear equations

$$\sum_{n=1}^{N} \frac{y_n - \mu_n}{a_n(\phi)v(\mu_n)g'(\mu_n)} x_n = \mathbf{0}, \tag{2.109}$$

$$\sum_{n=1}^{N} \frac{y_n - \mu_n}{a_n(\phi)v(\mu_n)g'(\mu_n)} z_n - \mathbf{K}^{-1}\mathbf{u} = \mathbf{0}. \tag{2.110}$$

Green [125] proposed 2.108 as a penalized likelihood for semi-parametric regression models and derived a Fisher scoring algorithm to iteratively optimize $\boldsymbol{\beta}$ and $\mathbf{u}$. His algorithm was modified by Breslow and Clayton [124], who showed the solutions to 2.109 and 2.110 via Fisher scoring can be expressed as the best linear unbiased estimators (BLUE) in the normal-theory linear mixed model,

$$\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \tilde{\boldsymbol{\epsilon}} \tag{2.111}$$

$$\tilde{\boldsymbol{\epsilon}} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}^{-1}) \tag{2.112}$$

for the working response variable $\tilde{\mathbf{y}}$,

$$\tilde{y}_n = \eta_n + (y_n - \mu_n)g'(\mu_n), \quad n = 1, \ldots, N. \tag{2.113}$$

The resulting algorithm is similar to the IWLS procedure for the GLM model discussed in the previous subsection. First, compute $\tilde{\mathbf{y}}$ at the current estimate of $\boldsymbol{\beta}$ and $\mathbf{u}$. Then, updated estimates for $\boldsymbol{\beta}$ and $\mathbf{u}$ can be obtained as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\tilde{\mathbf{y}} \tag{2.114}$$

and

$$\hat{\mathbf{u}} = \mathbf{K}\mathbf{Z}^T\mathbf{V}^{-1}(\tilde{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{2.115}$$

where $\mathbf{V} = \mathbf{W}^{-1} + \mathbf{Z}\mathbf{K}\mathbf{Z}^T$ denotes the marginal covariance matrix for 2.111.

In practice, the covariance matrix $\mathbf{K}$ of $\mathbf{u}$ often contains unknown variance components $\boldsymbol{\theta}$ that need to be estimated. In the derivations above, we have ignored the dependence on $\boldsymbol{\theta}$. Breslow and Clayton propose to substitute the estimates $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}) = (\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \hat{\mathbf{u}}(\boldsymbol{\theta}))$ at convergence into 2.107, to construct a profile quasi-log-likelihood for optimizing $\boldsymbol{\theta}$. They show that, following further approximations, the objective corresponds the profile likelihood (or its REML version) for the associated normal theory model 2.111.

It should be noted that, due to the various approximations made throughout the derivation of the penalized quasi-log-likelihood objective, the resulting estimators are known to be inconsistent [102]. Therefore, large samples sizes will not alleviate the bias introduced by the approximations. While bias-corrected variants based on higher-order Laplace approximations have been proposed [126], they also cannot eliminate the bias asymptotically.

A relevant special case occurs when the variance components of $\mathbf{u}$ approach zero. In this situation, the distribution of the random effects $p(\mathbf{u})$ is concentrated near its mode. As a consequence, the Laplace approximation, which is based on an expansion at the mode of $p(\mathbf{u})$, becomes accurate [102]. This fact makes the quasi-likelihood approximation particularly useful when deriving hypothesis tests for zero variance components.

2.1.9 Challenges in genetic association analyses

Genetic association analyses aim to assess the statistical correlation between a genetic variant and a phenotypic trait. In genome-wide association studies (GWAS) one often considers organismal traits such as height, body-mass-index (BMI) or particular physical or mental disease states. Quantitative trait loci (QTL) studies focus on the genetic basis of molecular traits, such as gene expression, protein abundance or chromatin conformation, the spatial organization of DNA in the cell nucleus. One typically considers single nucleotide polymorphisms (SNPs), that is, genetic mutations of single bases at a specific position in the genome, which are present in at least 1% of the population. These SNPs are further assumed to be biallellic, that is, they generally take on only two different values corresponding to the prevalent (reference) genotype, say A, and most common alternative allele, say a. In order to model the effect of genetic variation on the phenotype, different encodings for heterozygous (Aa) and homozygous (AA or aa) genotypes may be considered

- Dominant model. $AA = 0, Aa = 1, aa = 1$
- Recessive model: $AA = 0, Aa = 0, aa = 1$
- Additive model: $AA = 0, Aa = 1, aa = 2$

The additive model, also known as allele dosage model, is a common choice in both GWAS and eQTL studies and will be used throughout this thesis. Let $\mathbf{g} \in \{0, 1, 2\}^N$ denote the genotype vector for a particular SNP of interest in N individuals and let $\mathbf{y} \in \mathbb{R}^N$ be a

(quantitative) trait. Using the basic regression framework, the genetic effect of **g** on **y** may be modeled as

$$\mathbf{y} = \mathbf{g}\beta + \boldsymbol{\epsilon} \tag{2.116}$$

where $\epsilon_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. As outlined in section 2.1.5, a test statistic T may then be constructed to assess $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$ and obtain a p-value.

Population structure and other confounding factors

In practice, it is often necessary to control for additional covariates that may correlate with the genotype **g** and dependent variable **y**, to improve statistical power and avoid an excess of false positive associations. Let $P = p(T \geq t \mid H_0)$ be the p-value for an observed value t of the test statistic T. If the model is appropriately specified, P should approximately follow a uniform distribution $P \sim U(0, 1)$ under the null hypothesis. A Q-Q plot comparing the distribution of the observed test statistic to a uniform distribution is a useful visual diagnostic that can indicate whether a test has produced more significant associations than is to be expected by chance (**Fig.** 2.3). Poor model calibration due to unmodeled confounding factors typically leads to a departure from the null across the entire distribution, whereas the presence of genuine genetic signals will generate deviations at the tail end of the range.

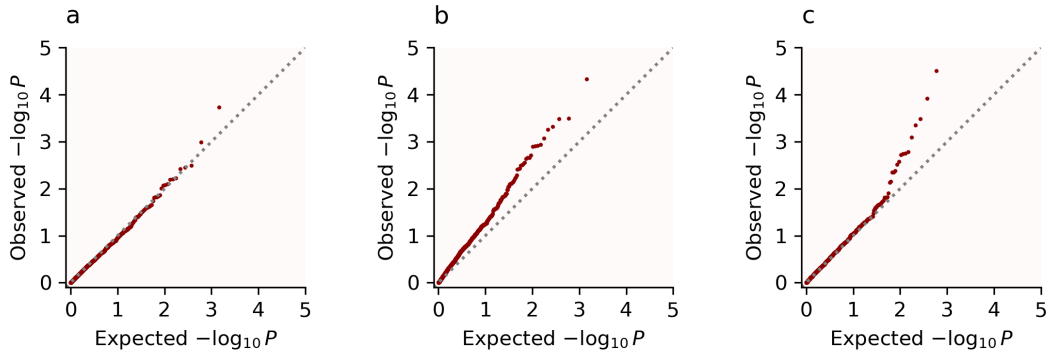

Figure 2.3. Example Q-Q plots showing the quantiles of the observed p-value distribution (y axis) and expected distribution under the null (x axis). (a) If no associations are present and the test statistic is calibrated, points are expected to closely follow the diagonal. (b) Unmodeled confounding variables produce a systematic deviation from the null distribution. (c) Genuine genetic effects lead to a deviation at the tail of the highly significant range.

Additional covariates $\mathbf{X}$ are commonly modeled as fixed effects,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{g}\beta + \boldsymbol{\epsilon}. \tag{2.117}$$

For instance, these may include batch identifiers that can indicate potential technical effects due to differences in sample preparation and processing. In gene expression QTL (eQTL) studies, one commonly controls for broad transcriptional trends across multiple genes, corresponding to potential known and hidden sources of confounding variation. These global patterns can be identified using principal component analysis (PCA) or other linear and non-linear methods for factor analysis or dimensionality reduction [127–131].

Population substructure, such as ethnic background as well as familial and higher-order relatedness, presents another important type of confounding variation. [132,133]. The prevalence of a phenotype as well as the distribution of allele frequencies are known to be population-specific [41], which can lead to the identification of spurious association signals. Using genome-wide genotyping data, it is possible to infer empirical patterns that accurately distinguish subtle ancestries and population substructures [134, 135]. Leading principal components (PCs) of the genotype data representing major axes of genetic variation can be included as covariates in the regression model 2.117 to account for global population effects [132,136]. Alternatively, the data may be used to construct a kinship matrix $\mathbf{R} \in N \times N$ of genetic relatedness [47,48] (see [137] for common used similarity measures) which may serve as the covariance matrix for a random effect term in a linear mixed model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{g}\beta + \mathbf{u} + \boldsymbol{\epsilon}, \tag{2.118}$$

where $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{R})$. For example, Let $\mathbf{G}$ denote the $N \times D$ matrix of genome-wide genotype measurements for N SNPs, where typically $N < D$, and let $\mathbf{R} = \mathbf{G}\mathbf{G}^T$ be proportional to the empirical covariance. Using the singular value decomposition of $\mathbf{G}$, it is easy to show that $\mathbf{R} = \mathbf{Z}\mathbf{Z}^T$, where $\mathbf{Z} \in \mathbb{R}^{N\times N}$ are the principal components of $\mathbf{G}$ weighed by their singular values. Therefore the model may equivalently be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{g}\beta + \mathbf{Z}\tilde{\mathbf{u}} + \boldsymbol{\epsilon}, \tag{2.119}$$

where $\tilde{\mathbf{u}} \sim \mathcal{N}(\mathbf{0}, \tau^2\boldsymbol{\mathcal{I}}_N)$. That is, the random effect model is regressing $\mathbf{y}$ on the full set of principal components. In the fixed effect model 2.117, the number of PCs cannot be of the same order of N while maintaining reasonable statistical power [138]. Therefore, LMMs are often better suited to controlling for subtle population effects that may not be

captured by the leading principal components of the genotype matrix. This improvement in modeling capacity comes at the cost of computational complexity. Fitting the LMM 2.118 requires storing and inverting the marginal covariance matrix, which scale as the square and cube of the sample size N, respectively. However, in the context of genetic association analyses, where the same (known) covariance matrix $\mathbf{R}$ is used to test thousands of SNP-trait pairs, optimizations are possible [42, 47, 139, 140]. These methods run in $O(N^2)$ once initial computations of order $O(N^3)$ have been performed. If the kinship matrix is of low rank, further improvements are possible, reducing the cost to an initial $O(N^2)$ and $O(N)$ per SNP-trait pair. For example, FaST-LMM [139] uses the spectral decomposition of the kinship matrix to rotate the responses $\mathbf{y}$ and covariates such that they become uncorrelated. The transformed data can then be fitted efficiently using linear regression without random effects.

Multiple testing correction

Typical GWAS or eQTL studies test thousands to millions of variant-trait pairs, making appropriate multiple testing adjustments essential to limit the number of false positive findings (type I error) [2, 12].

A simple strategy is to bound the family-wise error rate (FWER), that is the probability of observing at least one false discovery across all tested variants for a particular trait. A classical solution is the Bonferroni procedure, which rejects the null hypothesis $H_0^{(i)}$ for test i if

$$P_i < \frac{\alpha}{m}, \tag{2.120}$$

where P_i is the estimated p-value for $H_0^{(i)}$, α is a chosen FWER threshold (e.g., $\alpha = 0.05$) and m is the total number of tested hypotheses [43]. Here, $P_i \cdot m$ is sometimes referred to as the Bonferroni-adjusted p-value. The proof simply follows from Boole's inequality, as

$$\text{FWER} = p\big(\bigcup_{i=1}^{m_0}\{P_i < \frac{\alpha}{m}\}\big) \leq p\big(\sum_{i=1}^{m_0}\{P_i < \frac{\alpha}{m}\}\big) = m_0\frac{\alpha}{m}, \tag{2.121}$$

where m_0 is the number of null hypothesis that are true and we have used that p-values are uniform under the null. The Bonferroni method ensures that FWER $< \alpha$ in the strong sense, that is, FWER control is guaranteed for any configuration of true and false hypotheses tested (as opposed to weak control, where FWER $< \alpha$ is only guaranteed to hold when all null hypotheses are true). Notably, it does not assume independence of the test statistics and

conversely does not account for dependencies between tested variants (linkage disequilibrium). As a result, the bound may be loose in practice. When performing very large numbers of tests the Bonferroni correction (and FWER control more generally) can be limiting, as it exerts very stringent control on the number of false positive associations. In fact, procedures that control for the FWER in the strong sense, often have substantially less power than the individual tests at the same significance thresholds [141].

As an alternative approach, one may choose to bound the false discovery rate (FDR), that is, the expected proportion of positive associations that are false (incorrect rejection of the null hypothesis) [141]. The concept of FDR control for the multiple testing problem, as well as a controlling procedure were first introduced in a seminal paper by Benjamini and Hochberg in 1995. Assume the p-values P_i are sorted in ascending order. The following algorithm (Benjamini-Hochberg step-up procedure) controls the FDR at level α[3]:

- Choose the largest k such that $P_k \leq \frac{k}{m}\alpha$,
- Reject all null hypotheses $H_0^{(i)}$ for $i = 1, \ldots, k$.

Equivalently, an adjusted p-value may be computed as $Q_i = \frac{P_i \cdot m}{i}$ and $H_0^{(i)}$ is rejected if $Q_i < \alpha$. The Benjamini-Hochberg procedure is valid for independent test statistics, but can also be applied to correlated p-values under certain conditions [142]. Alternatively, the Benjamini–Yekutieli procedure controls the FDR under arbitrary dependence assumptions [143], but tends to be less powerful [142]. The development of methods for FDR control remains an active area of research [142].

2.2 The variational autoencoder

The second part of this chapter introduces the variational autoencoder (VAE) [144], a flexible latent variable model widely used in machine learning for tasks such as data generation, dimensionality reduction, and representation learning. Section 2.2.1 gives an overview of stochastic variational inference (SVI) and the VAE as introduced by Kingma and Welling in 2013. I also discuss the conditional VAE in section 2.2.2, a model that factors out latent effects driven by an observed auxiliary variable (usually a discrete label) and enables condi-

[3] Similar to the Bonferroni adjustment, the BH method actually controls its error rate at $m_0 \frac{\alpha}{m}$ where m_0 is the number of null hypothese that are true [142].

tional data generation. The chapter concludes with a discussion of scVI [131], a variational autoencoder model for single-cell sequencing data, in section 2.2.3.

2.2.1 Auto-encoding variational Bayes

The variational autoencoder models a distribution over high-dimensional datapoints, $\mathbf{x}_n \in \mathbb{R}^D$, $n = 1, \ldots, N$, as the result of a generative process. As such, the VAE is part of the broader class of generative models. These models are simulators, trained to generate new examples that resemble the training data, but are not exactly the same. Within the constraints of the specific modeling framework, a generative model forms a hypothesis of how the observations could have been generated in the real world, by identifying patterns and regularities in the data. For example, in order to reproduce the complex dependencies between pixels in image data, a model will likely need to learn some internal representation of the depicted shapes, poses, lighting and so forth. Similarly, generating high-dimensional transcriptomics data requires an understanding of the specific gene expression patterns in different tissues and cell types. In many scientific applications, these learned representations are of primary interest to the practitioner.

Latent variable models

Formally, a VAE is a type of probabilistic latent variable model. The high dimensional observations $\mathbf{x} \in \mathbb{R}^D$ are assumed to be generated from a distribution that depends on unobserved latent states $\mathbf{z} \in \mathbb{R}^K$, $K << D$. Here, $\mathbf{z}$ encodes systematic, low-dimensional structure among the observed variables that is independent of observation noise. A priori, $\mathbf{z}$ is assumed to be drawn from a distribution $p(\mathbf{z})$, where typically $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_K)$ (although the prior may also depend on additional unknown parameters). The latent variables are then mapped to the parameters of a conditional likelihood model for the observed data, $p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z})$, using a deterministic function $f_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$. For example, continuous observations may be modeled under a Gaussian noise model as $\mathbf{x} \,|\, \mathbf{z} \sim \mathcal{N}(f_{\boldsymbol{\theta}}(\mathbf{z}), \sigma^2 \boldsymbol{\mathcal{I}}_D)$ where σ^2 is a global hyperparameter. Taken together, the prior $p(\mathbf{z})$ and conditional likelihood $p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z})$ define the generative model and marginal distribution for $\mathbf{x}$

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{z}) d\mathbf{z} = \mathbb{E}_{p(\mathbf{z})}[p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z})]. \tag{2.122}$$

Once $\mathbf{x}$ has been observed, an updated posterior distribution over the latent states can be formulated using Bayes rule as

$$p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})} = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})}. \tag{2.123}$$

Suppose we are given a dataset of i.i.d samples $\mathbf{x}_n \in \mathbb{R}^D$, $n = 1, \ldots, N$. We are interested in obtaining estimates of the model parameters θ and evaluating the posterior distribution of latent states $p(\mathbf{z}_n \mid \mathbf{x}_n)$. As discussed in the context of linear mixed models, even when $f_{\boldsymbol{\theta}}$ is linear, the integral 2.122 is generally intractable for arbitrary likelihood models $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$.

The evidence lower bound (ELBO)

Variational Bayesian methods use a particular decomposition of the log-marginal likelihood 2.122 to learn an approximate posterior distribution and perform parameter inference. Using the definition of the posterior 2.123, note that we can write for any distribution $q(\mathbf{z})$,

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z})}[\log p_{\boldsymbol{\theta}}(\mathbf{x})] \tag{2.124}$$

$$= \mathbb{E}_{q(\mathbf{z})}[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})}] = \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})q(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})q(\mathbf{z})}\right] \tag{2.125}$$

$$= \underbrace{-\mathbb{E}_{q(\mathbf{z})}\left[\log \frac{q(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}\right]}_{\mathcal{L}} + \underbrace{\mathbb{E}_{q(\mathbf{z})}\left[\log \frac{q(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})}\right]}_{\mathrm{KL}(q(\mathbf{z}) \,||\, p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}))}, \tag{2.126}$$

where $\mathrm{KL}(q(\mathbf{z}) \,||\, p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}))$ is the Kullback-Leibler divergence, a statistical distance between $q(\mathbf{z})$ and the true posterior. The first term in 2.126, known as the evidence lower bound (ELBO) on the marginal likelihood of $\mathbf{x}$, is independent of the unknown true posterior and marginal distribution. Note that because the KL-divergence is always non-negative, we have

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathcal{L}. \tag{2.127}$$

Furthermore, equality holds if and only if $q(\mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$. The ELBO may equivalently be written as

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{z})}[\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})] - \mathrm{KL}(q(\mathbf{z}) \,||\, p(\mathbf{z})). \tag{2.128}$$

Here, the first term quantifies the ability of the model to accurately represent an observed data point $\mathbf{x}$ for given $\boldsymbol{\theta}$ and posterior approximation $q(\mathbf{z})$. The second term penalizes deviations from the specified prior distribution.

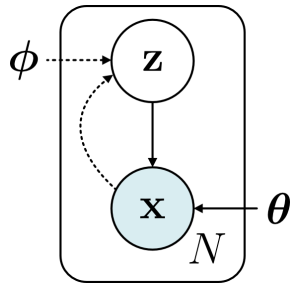

Figure 2.4. The VAE graphical model. Under the generative model, low-dimensional latent states $\mathbf{z}$ are mapped to the parameters of a conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ over high-dimensional observed variables $\mathbf{x}$ using a parametric function $f_{\boldsymbol{\theta}}$ (solid arrows). Dashed lines show the variational approximation $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ with parameters $\boldsymbol{\phi}$. Adapted from [144].

In variational inference, one first defines a family of, so-called variational distributions $q_{\boldsymbol{\phi}}(\mathbf{z})$ parameterized by $\boldsymbol{\phi}$ and chooses $\boldsymbol{\phi}$ to maximize the evidence lower bound $\mathcal{L}$. For any particular choice of $\boldsymbol{\theta}$, $\log p_{\boldsymbol{\theta}}$ is constant and maximizing the ELBO is equivalent to minimizing $\mathrm{KL}(q(\mathbf{z}) \,||\, p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}))$. The variational family is chosen such as to make inference tractable. At the same time, it needs to be expressive enough to provide a sufficiently close approximation to the true posterior distribution.

The VAE uses a parametric function $g_{\boldsymbol{\phi}}$ to map observed data points to the parameters of a distribution $q_{\boldsymbol{\phi}}(\mathbf{z}) = q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$, known as a recognition model. The complete graphical model representation is shown in **Fig.** 2.4. Both $f_{\boldsymbol{\theta}}$, which maps $\mathbf{z}$ to the parameters of the conditional likelihood, and $g_{\boldsymbol{\phi}}$ are implemented using neural networks, allowing for flexible, non-linear generative models and posterior approximations. Under the assumption that the conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ is differentiable w.r.t. to both $\boldsymbol{\theta}$ and $\mathbf{z}$, the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be optimized using gradient-based methods.

Stochastic gradient-based optimization of the ELBO

In order to enable gradient-based optimization of the ELBO $\mathcal{L}$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, suitable gradient estimators need to be derived. The main challenge is computing the gradient for $\boldsymbol{\phi}$. A standard Monte-Carlo estimator for the gradient of an expectation $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[f(\mathbf{z})]$ can be constructed as [145]

$$\frac{\partial}{\partial \boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[f(\mathbf{z})] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}\left[f(\mathbf{z}) \frac{\partial}{\partial \boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \right] \tag{2.129}$$

$$\approx \frac{1}{L} \sum_{l=1}^{L} f(\mathbf{z}^{(l)}) \log q_{\boldsymbol{\phi}}(\mathbf{z}^{(l)} \mid \mathbf{x}) \tag{2.130}$$

where $\mathbf{z}^{(l)}$ are samples from $q_{\boldsymbol{\phi}}(\cdot \,|\, \mathbf{x})$ and the first equality uses the properties of the log derivative,

$$\frac{\partial}{\partial \boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}) = \frac{\frac{\partial}{\partial \boldsymbol{\phi}} q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x})}{q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x})}. \tag{2.131}$$

However, this estimator has high variance in practice [145]. Kingma and Welling [144] propose an alternative estimator based on the fact that for many common distributions, the random variable $\tilde{\mathbf{z}} \sim q_{\boldsymbol{\phi}}(\cdot \,|\, \mathbf{x})$ can be reparameterized using an auxiliary transform of a noise variable $\boldsymbol{\epsilon}$,

$$\tilde{\mathbf{z}} = g_{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\epsilon}), \tag{2.132}$$

where the distribution of $\boldsymbol{\epsilon}$ does not depend on any parameters. For example, if $\tilde{\mathbf{z}}$ a multivariate Gaussian, $\tilde{\mathbf{z}} \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}^{\mathbf{2}}(\mathbf{x}))$, we can write

$$\tilde{\mathbf{z}} = \boldsymbol{\mu}(\mathbf{x}) + \text{diag}(\boldsymbol{\sigma}(\mathbf{x}))\boldsymbol{\epsilon} \tag{2.133}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_K)$ (similarly for other location-scale family distributions). As another example, if the distribution of $\tilde{\mathbf{z}}$ has a tractable inverse c.d.f., $F^{-1}_{\tilde{\mathbf{z}}}(\cdot, \mathbf{x})$,

$$\tilde{\mathbf{z}} = F^{-1}_{\tilde{\mathbf{z}}}(\boldsymbol{\epsilon}, \mathbf{x}) \tag{2.134}$$

where $\boldsymbol{\epsilon}_k \overset{\text{i.i.d}}{\sim} U(0, 1)$. If the posterior distribution admits a representation of the form 2.132, a Monte-Carlo estimator for the expectation $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}[f(\mathbf{z})]$ may be formed as [144]

$$\mathbf{E}_{q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x})}[f(\mathbf{z})] = \mathbb{E}_{\boldsymbol{\epsilon}}[f(g_{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\epsilon}))] \tag{2.135}$$

$$\approx \frac{1}{L} \sum_{l=1}^{L} f(g_{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\epsilon}^{(l)})) \tag{2.136}$$

where $\boldsymbol{\epsilon}^{(l)}$ are samples of $\boldsymbol{\epsilon}$. The expression 2.136 can now easily be differentiated with respect to $\boldsymbol{\phi}$. Applying this estimator to the first term in 2.128 and using the VAE variational posterior, one obtains the approximate ELBO

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) \approx \frac{1}{L} \sum_{l=1}^{L} \log p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \tilde{\mathbf{z}}^{(l)}) + \text{KL}(q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}) \,||\, p(\mathbf{z})) \tag{2.137}$$

where $\tilde{\mathbf{z}}^{(l)} = g_{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\epsilon}^{(l)})$. The KL term can often be evaluated in closed form (e.g., for Gaussian prior and posterior distribution) or estimated using 2.136.

Given a dataset $\mathbf{X} \in \mathbb{R}^{N \times D}$ of i.i.d observations $\mathbf{x}_n$, an approximate lower bound of the marginal likelihood is

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}) = \sum_{n=1}^{N} \tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_n). \tag{2.138}$$

In practice, one typically performs gradient-based updates for random subsets $\mathbf{X}^{(M)} \in \mathbb{R}^{M \times D}$, $M < N$ of the full data, termed mini-batches. Mini-batch stochastic gradient descent [146, 147] has become the state of the art for neural network training, due to increased computational parallelism, reduced memory footprint and more importantly, improved generalization performance and optimization convergence [148–150]. The mini-batch objective is

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^{(M)}) = \sum_{n=1}^{M} \frac{N}{M} \tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_n^{(M)}). \tag{2.139}$$

where $\mathbf{x}_n^{(M)}$ are the observations in minibatch $\mathbf{X}^{(M)}$.

The idea of gradient-based stochastic variational inference (SVI) is general enough to be applicable to a wide variety of probabilistic models. For example, SVI forms the backbone of the probabilistic programming language Pyro [151], which builds on top of PyTorch [152], a GPU-accelerated framework for automatic differentiation and deep learning. Once a probabilistic generative model and approximate posterior have been specified, Pyro allows for automatic parameter inference and estimation of posterior probabilities.

Encoder-decoder interpretation

Consider the following simple VAE generative model for continuous data [144],

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_K) \tag{2.140}$$

$$\mathbf{x} \,|\, \mathbf{z} \sim \mathcal{N}(f_{\boldsymbol{\theta}}(\mathbf{z}), \tau^2 \boldsymbol{\mathcal{I}}_K) \tag{2.141}$$

where $f_{\boldsymbol{\theta}}$ is a multi-layer fully-connected neural network and τ^2 is a hyperparameter. The variational model is

$$q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}) = \mathcal{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \mathrm{diag}(\sigma^2_{\boldsymbol{\phi}}(\mathbf{x}))), \tag{2.142}$$

where both $\boldsymbol{\mu}_{\boldsymbol{\phi}} : \mathbb{R}^D \mapsto \mathbb{R}^K$ and $\sigma^2_{\boldsymbol{\phi}} : \mathbb{R}^D \mapsto \mathbb{R}^K$ are implemented using neural networks. To compute the approximate evidence lower bound for a given observation $\mathbf{x}$, one first computes the parameters $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})$ and $\sigma^2_{\boldsymbol{\phi}}(\mathbf{x})$ and draws $\mathbf{z}$ from the posterior distribution over latent states using the reparameterization trick (2.133). The sampled latent representation $\mathbf{z}$ is then mapped to the mean of the conditional distribution $p(\mathbf{x} \,|\, \mathbf{z})$. This 'forward pass' is visualized in **Fig.** 2.5. The variational distribution performs the role of an encoder, compressing the high-dimensional data $\mathbf{x}$ in order to pass it through the latent space bottleneck. The data is then reconstructed, using $f_{\boldsymbol{\theta}}$ as a decoder to map samples from $q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x})$ to the mean $\hat{\mathbf{x}}$ of

the conditional likelihood $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$. In this sense, the VAE can be viewed as a probabilistic variant of classical encoder-decoder neural network architectures, termed Autoencoder models [153]. By optimizing the evidence lower bound for a dataset $\mathbf{X} \in \mathbb{R}^{N \times D}$

$$\frac{1}{2} \sum_{n=1}^{N} ||\mathbf{x} - \hat{\mathbf{x}}||_2^2 / \tau^2 - \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{x} \mid \mathbf{x}) \,||\, p(\mathbf{z})) + \text{const.}, \tag{2.143}$$

the VAE is trained to balance reconstruction performance and the distance between approximate posterior and latent prior. Note that the hyperparameter τ^2 only affects the first term and therefore plays the role of a regularization parameter controlling the weighting of the two terms.

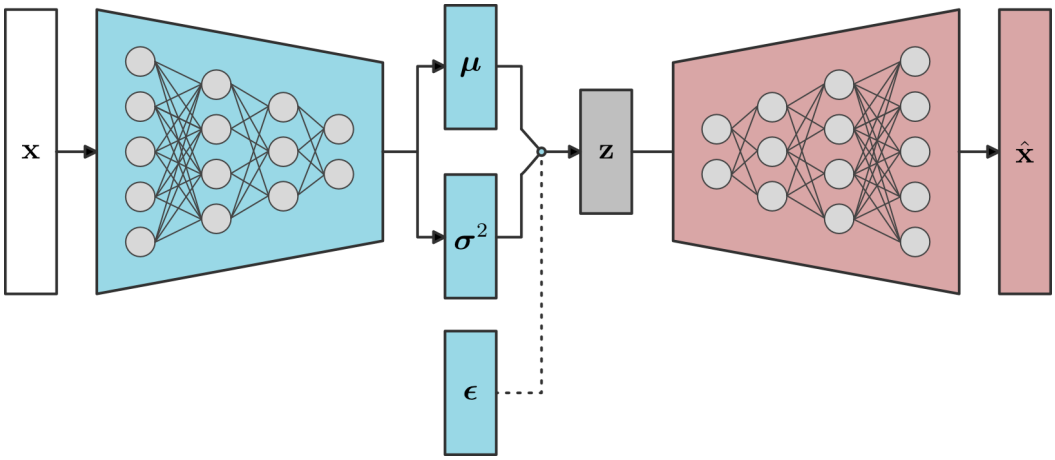

Figure 2.5. The VAE as an encoder-decoder architecture. Solid lines represent deterministic transformations. The dashed line corresponds to sampling from $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_K)$. During training, observed data points $\mathbf{x}$ are mapped to latent states using a probabilistic encoder $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ (blue). Using sampled latent representations, the data is approximately reconstructed as $\hat{\mathbf{x}}$, the mean of the conditional likelihood (red).

2.2.2 Conditional variational autoencoders

In many practical applications additional label information is available. For example, biological samples might be associated with different donors, disease statuses or technical replicates. In such cases, the label information can be incorporated into a variational autoencoder model in order to separate shared and label-specific variation. A conditional variational autoencoders (CVAE) extends the original VAE model by conditioning the generative process

- and commonly also the posterior approximation - on an observed label y [154, 155]. In the most simple case, the label information (one-hot encoded) is simply concatenated with the sampled latent states prior to the decoding step, such that

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z} \,|\, y) = p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z}, y)p(\mathbf{z}). \tag{2.144}$$

Alternatively, an explicit conditional prior $p_{\boldsymbol{\theta}}(\mathbf{z} \,|\, y)$ may be used. Since y is observed, the inference process using SVI is analogous to the standard VAE. However, the framework has also be extended to the semi-supervised case where y is partially unobserved, by introducing an additional recognition network $q_{\boldsymbol{\phi}}(y \,|\, \mathbf{x})$ such that $q_{\boldsymbol{\phi}}(\mathbf{x}, y \,|\, \mathbf{z}) = q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x})q_{\boldsymbol{\phi}}(y \,|\, \mathbf{x})$ [156].

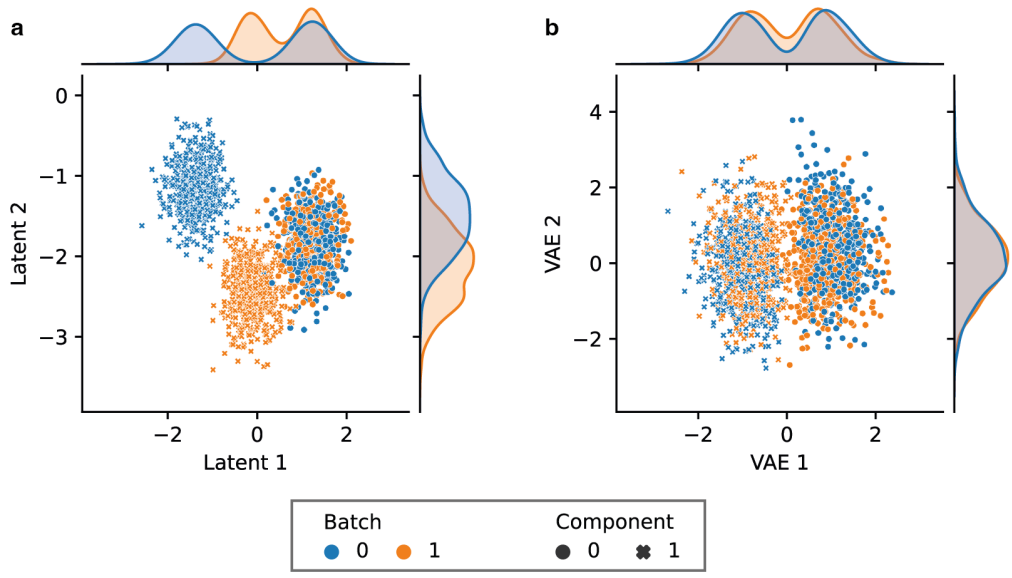

Figure 2.6. CVAE toy example. (a) Simulated latent space, sampled from two Gaussian mixture models (batch) with two components (component). (b) Integrated latent space inferred by a CVAE model, using the batch as an auxiliary observed variable during training.

A toy example of the CVAE for fully observed label information is given in **Fig.** 2.6. **Fig.** 2.6a shows the simulated two-dimensional latent space where points were sampled from two Gaussian mixture models (batch) with two components each. Suppose the differences between batches corresponds to nuisance variation, e.g., technical effects caused by the measurement process. 50-dimensional observed points were generated using a random projection matrix. The difference between batches is component-specific and therefore non-linear

in the two-dimensional latent space. By providing the batch ID as a conditional variable during training, a simple CVAE model with isotropic latent prior and Gaussian variational posterior[4] successfully integrates batches while maintaining the separation between mixture components (**Fig.** 2.6b).

In addition to removing label information from the latent representation, the CVAE may also be used to perform style transfer [156]: For a given point $\mathbf{x}$ with label y, the CVAE allows to generate new examples $\tilde{\mathbf{x}} \sim p_{\boldsymbol{\theta}}(\cdot\,|\mathbf{z}, \tilde{y})$ sharing the same latent state (style) $\mathbf{z} \sim q_{\boldsymbol{\phi}}(\cdot \mid \mathbf{x}, y)$ under alternative labels $\tilde{y}$ (content types). That is, style transfer poses the counterfactual question 'what would a data point $\mathbf{x}$ look like, had it been generated using a different label'. For example, CVAEs have been trained to visualize handwritten digits in different styles (e.g., stroke width, slant/upright) [144], perform neural machine translation [157], as well as designing drug-like molecules with specific chemical properties [158].

Identifiability and disentanglement

In representation learning, the goal is to identify latent factors $\mathbf{z}$ which explain systematic variation in high-dimensional observations $\mathbf{x}$. The concept of disentanglement refers to the property of a learned representation where the underlying factors of variation are separated and represented as distinct and independent dimensions or features [159, 160]. Intuitively, each latent factor should correspond to semantically meaningful concepts. For example, in images of faces, a disentangled representation may dedicate specific dimensions to factors such as pose, lighting and identity of a person. By learning disentangled latent representations, one may hope for improved generalization performance [161, 162], interpretability [163, 164] and faster learning in downstream abstract reasoning tasks [165]. The concept of disentanglement is also related to the more general idea of independent causal mechanisms in the causal inference literature [166, 167], positing that the causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.

The classical VAE model introduced in section 2.2.1 postulates a particular notion of disentanglement by assuming that the data-generating factors z_k, $k = 1, \ldots, K$, are stochastically independent under the prior $p(\mathbf{z}) = \prod_k p(z_k)$. In this context, inferring disentangled repre-

[4] Both encoder and decoder use a fully-connected network with one hidden layer (128 nodes). Trained on 1000 points using SGD with learning rate 10^{-3}.

sentations is closely related to the problem of model identifiability. Suppose $\mathbf{x}$ was generated from a latent variable model with joint density $p_{\boldsymbol{\theta}^*}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}^*}(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$ for some unknown parameter $\boldsymbol{\theta}^*$. The VAE framework allows to find a set of parameters $\boldsymbol{\theta}$ maximizing the marginal likelihood $p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})d\mathbf{z}$, such that

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \approx p_{\boldsymbol{\theta}^*}(\mathbf{x}). \tag{2.145}$$

However, equality of the marginal distributions does not necessarily imply $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ (model identifiability), and thus the corresponding joint densities $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}), p_{\boldsymbol{\theta}^*}(\mathbf{x}, \mathbf{z})$ may not be equal. In fact, Khemakhem et al. [168] proved that the true parameters are not identifiable in the general (unconditional) VAE model. As a consequence, even when the data was generated from independent factors, the VAE latent space may be entangled. However, Khemakhem et al. showed that by conditioning the prior distribution over latent variables on an auxiliary variable, such as a class label or time stamp, identifiability becomes possible up to some trivial transformations [168] given sufficient data.

2.2.3 VAEs for single-cell sequencing data

Variational autoencoder models have emerged as useful tools for the analysis of single-cell sequencing data [131, 169–173], offering a single framework for a variety of common tasks such as normalization, dimensionality reduction, imputation and differential expression analysis. These approaches have gained increasing popularity due to the exponential growth of single cell datasets in recent years [95], making computational scalability a primary concern. This section will describe scVI [131], one of the first first VAE-based methods dedicated to modeling single-cell transcriptome (scRNA-seq) measurements.

A standard preprocessing pipeline for single-cell RNA sequencing (scRNA-seq) data typically involves several steps to clean, normalize, and prepare the data for downstream analysis [174]. Following an initial quality assessment, the raw sequencing data is aligned to a reference genome [175] such that the number of reads mapping to each gene or transcript in the reference can be quantified. The result is a count matrix containing the measured expression of every gene in every cell. Cells are filtered based on the number of counts / detected genes and the fraction of mitochondrial read counts [174]. An unusually large number of detected genes may be the result of barcode collisions, where the measurements in a single 'cell' actually correspond to a mixture of read counts from multiple different cells. Few detected genes or a high fraction of mitochondrial reads can indicate a ruptured cell membrane

and associated a loss of cytoplasmic mRNA. The result is a filtered count matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ for N cells and D genes.

A typical scRNA-seq analysis workflow proceeds by normalizing the read counts to account for technical nuisance factors such as differences in library size / count depth, sampling noise and technical batch effects [174]. Depending on the normalization method, the data is then log-transformed to stabilize the variance and facilitate the application of methods that assume Gaussian observation noise [174]. Subsequently, the dimensionality of the data is reduced, e.g. using principal component analysis, to remove noise and enable the identification and visualization of cell types and states.

Starting with the filtered count matrix $\mathbf{X}$, single-cell variational inference (scVI) [131] offers an alternative strategy that allows for joint normalization and dimensionality reduction of single-cell RNA-seq data. Let s_n be the batch identifier for cell n and let $l_\mu(s_n), l_{\sigma^2}(s_n)$ be the empirical mean and variance of the log-library size (log of total counts) per batch. The core scVI generative model for $n = 1, \ldots, N$ cells and $d = 1, \ldots, D$ genes is as follows

$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_K) \tag{2.146}$$

$$l_n \,|\, s_n \sim \text{LogNormal}(l_\mu(s_n), l_{\sigma^2}(s_n)) \tag{2.147}$$

$$\boldsymbol{\rho}_n = f_\rho(\mathbf{z}_n, s_n) \tag{2.148}$$

$$w_{nd} \,|\, \rho_{nd} \sim \text{Gamma}(\rho_{nd}, \theta_d) \tag{2.149}$$

$$x_{nd} \,|\, w_{nd} \sim \text{Poisson}(l_n w_{nd}) \tag{2.150}$$

where f_ρ is a multi-layer fully-connected neural network. Here, $\mathbf{z}_n$ is a lower-dimensional representation of the cell state and l_n models a latent scaling factor. The final activation function of f_ρ is a softmax, such that $\sum_d \rho_d = 1$ and ρ_{nd} corresponds to the relative activity of gene d in cell n. The conditional likelihood of x_{nd} given ρ_{nd} and l_n is a Gamma-Poisson mixture with gene-specific dispersion parameter θ_d [92,176]. Given the relative proportions of mRNA transcripts and total counts in a cell, the distribution of transcript counts should be multinomial. Due to the large number of genes and total sequencing reads, the number of counts per gene should therefore be approximately Poisson-distributed (Poisson limit theorem). However, the count distribution for highly expressed genes often exhibits larger variance than is to be expected under a Poisson model [176–178], where mean and variance are determined by the same parameter. To address the issue of overdispersion, the expected

number of counts of the Poisson distribution is itself modeled as a random variable, following a Gamma distribution with mean ρ_{nd} and dispersion parameter θ_d. The density of the Gamma-Poisson mixture, also known as the negative binomial distribution, is available in closed form,

$$x_{nd} \,|\, \rho_{nd} \sim \text{NB}(l_n \rho_{nd}, \theta_d) \tag{2.151}$$

$$p(x_{nd} = k \,|\, \rho_{nd}) = \frac{\Gamma(\theta_d^{-1} + k)}{k! \Gamma(\theta_d^{-1})} \Big(\frac{\theta_d^{-1}}{\theta_d^{-1} + l_n \rho_{nd}}\Big)^{\theta_d^{-1}} \Big(\frac{l_n \rho_{nd}}{\theta_d^{-1} + l_n \rho_{nd}}\Big)^{k}. \tag{2.152}$$

and the mean and variance are given by

$$\mathbb{E}[x_{nd} \,|\, \rho_{nd}] = l_n \rho_{nd} \tag{2.153}$$

$$\text{Var}(x_{nd} \,|\, \rho_{nd}) = l_n \rho_{nd} + (\theta_d l_n) \rho_{nd}. \tag{2.154}$$

Contrary to the Poisson distribution, where mean and variance are equal, the variance of the negative binomial grows with the square of the mean. For small θ_d, the negative binomial approaches the Poisson distribution. The original scVI model introduced by Lopez et al [131] also considered an extended conditional likelihood model, designed to account for an overabundace of zero values compared to a standard negative binomial distribution. The resulting distribution is known as a zero-inflated negative binomial (ZINB). However, more recent studies have found that data generated by modern droplet-based single-cell technologies can be adequately modeled using a negative binomial distribution without zero inflation [179].

scVI infers an approximate posterior distribution over latent states, allowing to identify major axes of biological variation in the data while accounting for batch effects, library size and sampling noise. The scVI generative model is a special case of the CVAE framework. The variational distribution assumes independence between latent states and size factors (mean field approximation) [131]

$$q(\mathbf{z}_n, l_n \,|\, \mathbf{x}_n, s_n) = q(l_n \,|\, \mathbf{x}_n, s_n) q(\mathbf{z}_n, \,|\, \mathbf{x}_n, s_n). \tag{2.155}$$

Using multi-layer encoder networks, the observed data and batch ID are mapped to the parameters of the variational distribution, which is chosen to be Gaussian in the case of $\mathbf{z}_n$ and LogNormal for l_n. Note that a random variable follows a LogNormal distribution if its logarithm is Gaussian, making the LogNormal distribution amenable to the reparameterization trick. Furthermore, it allows for straightforward evaluation of the KL divergence between

prior and approximate posterior, by computing the analytic distance in the Gaussian base space. Under the variational approximation 2.155, the marginal likelihood can be optimized using the standard SVI algorithm introduced in section 2.2.1.

Chapter 3

Mixed effect models for single-cell genetics

The functional consequences of genetic variation can depend on the cellular context, the cell type or cell state [27, 180–183]. Historically, most existing studies had to rely on bulk sequencing data from tissue sections or sorted cell populations, measuring average molecular traits across thousands to millions of cells. As a result, these studies largely ignored more fine-grained cellular heterogeneity and associated genetic interaction effects. Advancements in single-cell sequencing technologies have made it possible to study cells within different cellular contexts, allowing to identify rare cell types and map continuous transitions. However, existing strategies for the analysis of genetic effects from single-cell sequencing data are predominantly based on methods originally developed for bulk sequencing data [84, 85, 97, 98], requiring discretization of cell states and subsequent aggregation of sequencing reads to create 'pseudo-bulk' profiles, as discussed in section 1.3.2.

In this chapter I describe a new approach based on the linear mixed model framework, taking a first step towards a systematic and impartial exploration and quantification of context-specific genetic effects at the level of single cells. The central idea is to summarize cell-state variation in a random effect covariance matrix, $\mathbf{K} = \mathbf{C}\mathbf{C}^T$, where $\mathbf{C}$ is a lower-dimensional embedding of the observed single-cell count data. The approach is conceptually related to linear mixed models used in GWAS and eQTL studies to account for genetic relatedness (section 2.1.9) and genotype-environment interactions [184]. Importantly, this formulation

allows to capture discrete cell states as well as continuous transitions such as differential or developmental trajectories. The idea is implemented in two concerted methods. First, I introduce scDALI , a model for detecting context-specific allelic imbalance from single-cell data. Second, I describe CellRegMap , the first statistical method designed specifically for single-cell eQTL testing. The contents of this chapter were published in [185] and [186].

3.1 scDALI: modeling allelic heterogeneity in single cells

Existing methods for genetic association analyses depend on profiling a large and genetically diverse group of individuals, which can be especially challenging for in vivo studies and non-human model systems. A potential alternative is to assess allele-specific signals [40, 187–189], which provide a complementary view of genetic effects on molecular traits and can be measured even in a single individual (section 1.2.1). When combined with single-cell technologies, this approach could offer a potent strategy for dissecting the functional consequences of genetic variations within complex tissues, encompassing diverse cell types and states. Studies focussing on context-specific allelic regulation are only beginning to emerge [98, 190], requiring principled computational tools to detect and quantify allelic imbalance at the level of single cells.

Here, I introduce single-cell differential allelic imbalance (scDALI), a statistical model and analysis framework for allele-specific quantifications of single-cell sequencing data. The model is designed to identify and test for different types of allelic imbalances, and differentiates homogeneous effects shared by all cells, from heterogeneous effects that align with specific cell types and states. The underlying problem is similar to differential expression testing, but instead of assessing variation in total expression counts, scDALI identifies differences in the fraction of counts originating from one of the gene alleles. scDALI can be used to estimate allelic imbalances from sparse sequencing data within individual cells, which allows for downstream visualization and interpretation of loci with significant allelic regulation. scDALI is generally applicable to sequence-based count data and can be used to model single-cell datasets generated using different technologies and modalities.

The model is validated on simulated data and applied to study allelic regulation of chromatin accessibility in a developmental timecourse of F1 *Drosophila melanogaster* embryos sampled at three different time points. To this end, I propose a VAE-based probabilistic model

to infer latent cell states and developmental stages from highly sparse open chromatin data (sci-ATAC-seq), while accounting for varying sampling intervals. Using the learned cell-state representation, scDALI detects hundreds of regulatory regions exhibiting heterogeneous allelic imbalance. Among these discoveries are regions with opposing allelic effects in different cell lineages, which would not be detected by bulk sequencing approaches. I then demonstrate how scDALI can be used to map the effects of known expression quantitative trait loci to specific cell types and states, by evaluating allelic imbalances of single-cell gene expression measurements of differentiating human induced pluripotent stem cells (iPSCs). These examples highlight how the proposed model can be employed across a range of species and data formats, harnessing single-cell technologies to eliminate the need for cell sorting. In particular, the approach is applicable to data from diverse outbred individuals as well as F1 crosses of inbred wild isolates.

Acknowledgements and contributions

This work was supervised by Oliver Stegle and Eileen Furlong. I developed and implemented the scDALI method and VAE model for open chromatin data. The F1 *Drosophila melanogaster* embryos were collected by Bingqing Zhao. Stefano Secchia adapted and applied the sci-ATAC-seq protocol with advice from James P. Reddington. I implemented the pipeline for processing the raw sci-ATAC-seq data and performed the simulation study and analyses of both datasets (developing F1 Drosophila embryos & scRNA-seq from differentiating human iPSCs). Stefano Secchia compared allele-specific effects discovered by scDALI to the activity of known transcription factors. Oliver Stegle, Eileen Furlong, Stefano Secchia and myself contributed to the interpretation of the results. All figures are my own work, unless otherwise stated.

3.1.1 The beta-binomial model for allele-specific quantifications

scDALI builds on the generalized linear mixed model framework (GLMM, section 2.1.8) to model context-specific allelic imbalance of quantitative traits, such as gene expression, chromatin accessibility or other epigenetic features, from single-cell sequencing data. For cells $n = 1, \ldots, N$ let d_n be the total number of reads mapping to a particular genomic feature such as a gene or peak of accessibility, and let k_n be the number of reads mapping to one of the two alleles in a diploid organism (see **Fig.** 1.3). For simplicity, I will assume the region may be assigned to a single haplotype, and refer to k_n as the maternal count.

scDALI builds on the Beta-Binomial likelihood model frequently used for modeling allelic imbalance from bulk-sequencing data [40, 188, 191, 192]. Suppose, for now, that all cells share the same underlying allelic rate μ, i.e., the expected relative activity of the maternal allele and let $r_n = k_n/d_n$ denote the allelic rate in cell n. In the simplest case, the k_n could be considered as independent draws from a binomial distribution, such that

$$r_n \sim \text{Bin}(d_n, \mu)/d_n. \tag{3.1}$$

Under the binomial model, mean and variance are coupled as $\text{Var}(r_n) = \frac{1}{d_n}\mu(1-\mu)$. Empirical allelic counts, however, often show greater variability than is to be expected under the binomial model. This is because cellular populations seldom exhibit complete homogeneity (e.g., owing to cell cycle influences), and allele-specific counts are susceptible to additional sources of variation stemming from both technical and biological factors. For example, gene expression is often discontinuous and follows stochastic patterns known as transcriptional bursting [193], leading to increased variability of expression counts both between gene alleles as well as cells. As a solution, μ may be replaced by a random variable $\tilde{\mu}$ with $\mathbb{E}[\tilde{\mu}] = \mu$. Specifically, I assume $\tilde{\mu}$ follows a beta distribution [40, 188, 191, 192] with dispersion parameter θ, that is,

$$r_n \sim \text{Bin}(d_n, \tilde{\mu})/d_n \tag{3.2}$$

$$\tilde{\mu} \sim \text{Beta}(\theta^{-1}\mu, \theta^{-1}(1-\mu)). \tag{3.3}$$

The resulting compound distribution for r_n, known as the Beta-Binomial distribution, is analytically tractable with variance

$$\text{Var}(r_n) = \frac{1}{d_n}\mu_n(1-\mu_n)\frac{\theta^{-1}+d_n}{\theta^{-1}+1} = a_n(\theta)v(\mu_n) \tag{3.4}$$

with variance function $v(\mu_n) = \mu_n(1-\mu_n)$ and $a_n(\theta) = \frac{1}{d_n}\frac{\theta^{-1}+d_n}{\theta^{-1}+1}$. When θ is large, draws will be similar to a Bernoulli distribution with counts coming almost exclusively from either allele. As θ tends to zero, the Beta-Binomial approaches the Binomial distribution.

Parameter estimation

Maximum-likelihood estimators for the mean and dispersion parameters p and θ of the Beta-Binomial distribution are not available in closed form. Minka [194] derived both fixed-point iterations and Newton-Raphson updates for the parameters of the Polya distributions, of which the Beta-Binomial distribution is a special case. Most bulk sequencing studies

are comprised of only a limited number of replicates and further assumptions are required to lower the estimation uncertainty, such as a shared mean-variance relationship between genomic features [191]. Single-cell sequencing datasets, on the other hand, typically contain thousands to millions of cells, allowing for direct estimation of p and θ separately for each feature of interest.

3.1.2 The scDALI model

The previous section discussed the basic Beta-Binomial model, under the assumption that cells are sampled from a homogeneous population with fixed expected allelic rate μ for all cells. When profiling heterogeneous systems, containing diverse cell types and states, this assumption is unlikely to hold. This is because genetic regulation and associated allele-specific effects frequently depend on the molecular context (section 1.3). Key to the scDALI approach is the observation that total counts, quantified at individual genomic features, provide a largely independent signal to allele-specific measurements, allowing to define cellular states in a data-driven manner. Established methods for the analysis of single-cell sequencing data use low-dimensional representations $\mathbf{C} \in \mathbb{E}^{N \times K}$ of the high-dimensional total count matrix to summarize biologically meaninful patterns in the data. For example, $\mathbf{C}$ may encode discrete cell clusters, principal components of the total count matrix or positions along a continuous trajectory or pseudo-temporal ordering [174, 195, 196]. For a given representation $\mathbf{C}$, scDALI constructs a cell-state covariance matrix $\mathbf{K} = \mathbf{K}(\mathbf{C}) \in \mathbb{R}^{N \times N}$ and models the latent allelic rates using a random effect model with Beta-Binomial likelihood,

$$r_n \,|\, \boldsymbol{\eta} \sim \text{BetaBin}(d_n, \theta^{-1}\mu_n, \theta^{-1}(1-\mu_n))/d_n \tag{3.5}$$

$$\mu_n = g^{-1}(\eta_i) \tag{3.6}$$

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{1}\alpha + \mathbf{X}\boldsymbol{\beta}, \sigma^2_{het}\mathbf{K}), \tag{3.7}$$

where $g(x) = \log(\frac{x}{1-x})$ denotes the logit link function. Here, $\mathbf{X}$ are optional covariates with associated fixed effect vector $\boldsymbol{\beta}$ and σ^2_{het} is a scaling parameter, quantifying the strength of context-specific, heterogeneous effects. The offset α models homogeneous imbalance, where $\alpha = 0$ corresponds to an expected allelic rate of 1/2. An overview of the model is given in **Fig.** 3.1.

scDALI uses a linear covariance function, such that $\mathbf{K} = \mathbf{C}\mathbf{C}^T$, in which case 3.7 corresponds to Bayesian linear regression of all cell-state dimensions $\mathbf{c}_k$, $k = 1, \ldots, K$ (columns

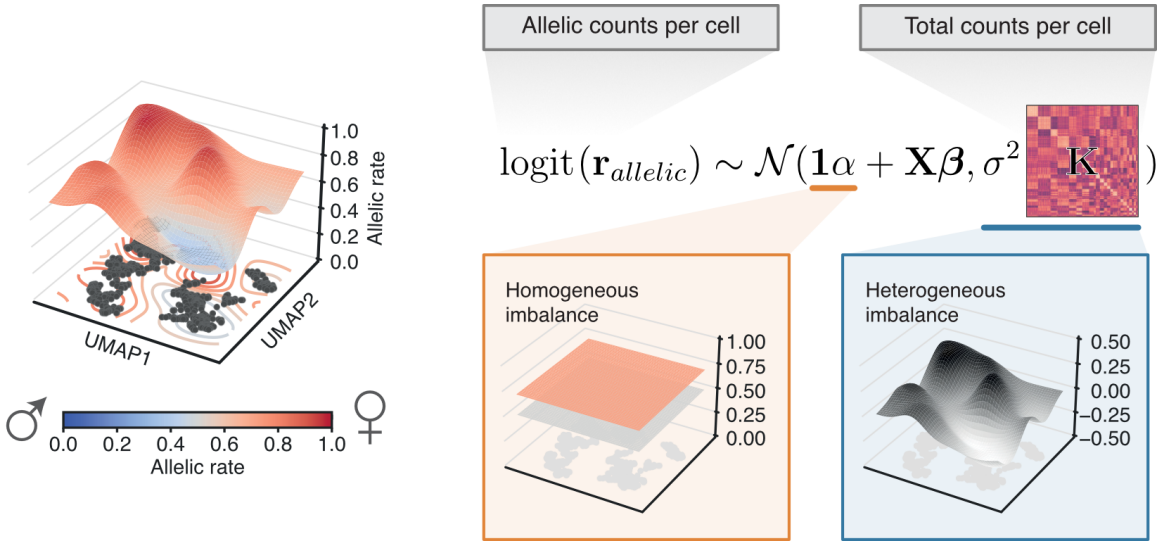

Figure 3.1. .scDALI integrates total and allele-specific read counts (compare with **Fig.** 1.3) to model allelic imbalance in a genomic region. The latent allelic rate, corresponding to the relative activity for a given haplotype, is decomposed into global (homogeneous) and cell-state-specific (heterogeneous) effects on a logit scale, using a generalized linear mixed model with Beta-Binomial likelihood. Heterogeneous imbalances are modeled using a cell-state covariance matrix derived from total read counts.

of **C**), as discussed in chapter 2. Nevertheless, non-linear effects can also be modeled by including additional transformed cell-state variables such as interactions $\mathbf{c}_k \odot \mathbf{c}_l$ or element-wise feature maps $\phi(\mathbf{c}_k)$. Alternatively, the model may be extended using common non-linear covariance functions from the Gaussian process (GP) literature [197].

3.1.3 Approximate inference

Due to the non-Gaussian conditional distribution $r_n \,|\, \boldsymbol{\eta}$, the marginal likelihood $p(\mathbf{r})$ is analytically intractable under the full scDALI model 3.5-3.7. I therefore resort to approximate inference techniques. Let the working response variable be defined as (see eq. 2.113)

$$\tilde{r}_n = \eta_n + (r_n - \mu_n)g'(\mu_n), \quad n = 1, \ldots, N. \tag{3.8}$$

Estimators for fixed effects $\alpha, \boldsymbol{\beta}$ may be obtained using the penalized quasi-likelihood approach (chapter 2, section 2.1.8), as iterative solutions to the working normal-theory approx-

imation (eq. 2.111) to the working variable $\tilde{\mathbf{r}}$,

$$\tilde{\mathbf{r}} = \mathbf{1}\alpha + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon} \tag{3.9}$$

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{het}\mathbf{K}) \tag{3.10}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}^{-1}) \tag{3.11}$$

where $\mathbf{W}^{-1}$ is diagonal and contains the inverse IWLS iterated weights (eq. 2.84),

$$\mathbf{W}^{-1} = \text{diag}(a_n(\theta)v(\mu_n)g'(\mu_n)^2). \tag{3.12}$$

with $v(\mu_n)$ and $a_n(\theta)$ as in eq. 3.4. Note that the derivative of the logit link is given by

$$g'(x) = \frac{1}{x(1-x)}, \tag{3.13}$$

and therefore

$$\mathbf{W} = \text{diag}(d_n\mu_n(1-\mu_n)\frac{\theta+1}{d_n\theta+1}). \tag{3.14}$$

The fixed effect estimates depend on the dispersion parameter θ and the variance component σ^2_{het}. Similar to the procedure proposed by Breslow and Clayton [124], estimators for the variance components may be obtained by optimizing the log-likelihood of the working normal-theory model at convergence. The complete algorithm alternates between solving the IWLS problem and optimizing θ and σ^2_{het}.

For large datasets, the penalized quasi-likelihood procedure tends to be slow. As an alternative, I consider sparse variational inference, a method originally developed for scalable Gaussian process regression, to optimize the model parameters and learn an approximate posterior distribution over latent allelic rates $p(\boldsymbol{\mu} \mid \mathbf{r})$ [198, 199]. Briefly, sparse variational inference learns a set of $M << N$ pseudo-inputs, referred to as inducing points, that are highly informative on the underlying function. A tractable bound on the marginal likelihood may then be derived. Computing the bound and its derivative scales linearly with N, and allows for a considerable reduction in computational complexity even for models with Gaussian observation noise, where evaluation of the marginal likelihood is of order $O(N^3)$ for arbitrary covariance matrices. An additional variational approximation is used to factorize the model such that the bound becomes amenable to stochastic optimization and non-Gaussian likelihoods can be handled using one-dimensional numerical integration [198,200] . In practice, estimating allelic rates under the Beta-Binomial model for thousands of candidate regions in large datasets can still be computationally challenging. For the analyses in this

chapter I therefore approximate the scDALI model, by replacing the Beta-Binomial with a homoscedasatic Gaussian likelihood for empirical alleic rates $r_n = k_n/d_n$. GPU-accelerated implementations of sparse variational inference are available as part of the Python libraries GPFlow [201] and GPytorch [202].

3.1.4 Statistical significance testing

The extent of allele-specific effects under the scDALI model is determined by the parameters α and σ^2_{het}. I consider the following three scenarios of null and alternative hypotheses, capturing different types of allelic imbalance:

scDALI-Het (Heterogeneous imbalance)

$$H_0^{het} : \sigma^2_{het} = 0 \text{ vs. } H_1^{het} : \sigma^2_{het} > 0 \tag{3.15}$$

scDALI-Hom (Homogeneous imbalance)

$$H_0^{hom} : \alpha = 0 \text{ vs. } H_1^{hom} : \alpha \neq 0 \tag{3.16}$$

scDALI-Joint (General imbalance)

$$H_0^{joint} : \sigma^2_{het} = 0, \alpha = 0 \text{ vs. } H_1^{joint} : \alpha \neq 0 \text{ or } \sigma^2_{het} > 0 \tag{3.17}$$

As discussed in the previous section, fitting the full scDALI model and evaluating the marginal likelihood $p(\mathbf{r})$ is computationally expensive for large datasets. I therefore derive score-based test statistics, which only require parameter estimates under the restricted null models.

scDALI-Het

scDALI-Het evaluates the evidence for heterogeneous allelic imbalance, that is $\sigma^2_{het} > 0$. The problem can be reduced to the variance component test for Gaussian LMMs introduced in chapter 2, section 2.1.6, by approximating the scDALI GLMM 3.5-3.7 using the normal-theory model 3.9-3.11. Zhang et al. [111] considered this approach in the context of semi-parametric mixed models, albeit motivated by a slightly different approximation to the quasi-likelihood function than the one proposed by Breslow and Clayton [124] (see section 2.1.8).

Under the null hypothesis $\sigma^2_{het} = 0$, scDALI reduces to a generalized linear model without random effects, such that the observations r_n are independent and parameters can be estimated efficiently. Let $\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0, \hat{\theta}_0$ be the iterated weighted least squares (IWLS) solutions to the working normal-theory model under the null GLM

$$\tilde{\mathbf{r}} = \mathbf{1}\alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.18}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}^{-1}), \tag{3.19}$$

with $\mathbf{W}$ defined in 3.14. Furthermore, let

$$\tilde{\mathbf{r}}_0 = \mathbf{1}\hat{\alpha}_0 + \mathbf{X}\hat{\boldsymbol{\beta}}_0 + g'(\hat{\boldsymbol{\mu}}_0)(\mathbf{r} - \hat{\boldsymbol{\mu}}_0) \tag{3.20}$$

be the working vector at convergence, where $g'(\cdot)$ is applied element-wise, and let

$$\hat{\boldsymbol{\mu}}_0 = g^{-1}(\mathbf{1}\hat{\alpha}_0 + \mathbf{X}\hat{\boldsymbol{\beta}}_0), \tag{3.21}$$

(again element-wise) with associated IWLS weights $\hat{\mathbf{W}}_0$. Following section 2.1.6, a score-based statistic for the variance component σ^2_{het} under the full normal-theory model 3.9-3.11 may be defined as

$$Q = \frac{1}{2}\tilde{\mathbf{r}}_0^T \hat{\mathbf{P}}_0^T \mathbf{K} \hat{\mathbf{P}}_0 \tilde{\mathbf{r}}_0, \tag{3.22}$$

where

$$\hat{\mathbf{P}}_0 = \hat{\mathbf{W}}_0 - \hat{\mathbf{W}}_0 \tilde{\mathbf{X}} (\tilde{\mathbf{X}}^T \hat{\mathbf{W}}_0 \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \hat{\mathbf{W}}_0. \tag{3.23}$$

is the projection matrix for fixed effects with $\tilde{\mathbf{X}} = \begin{bmatrix}\mathbf{X} & \mathbf{1}\end{bmatrix}$. Under the null hypothesis, Q is approximately distributed as a weighted sum of independent chi-squared random variables $z_i^2 \sim \chi^2(1)$ with one degree of freedom,

$$\sum_i \psi_i z_i^2. \tag{3.24}$$

where ψ_i are the ordered non-zero eigenvalues of $\frac{1}{2}\hat{\mathbf{V}}_0^{-1/2}\mathbf{K}\hat{\mathbf{V}}_0^{-1/2}$ with $\hat{\mathbf{V}}_0 = \hat{\mathbf{W}}_0^{-1}$. To evaluate the distribution of 3.24 and compute p-values, Davies method can be used [114, 203]. A Python implementation is provided by `limix` [1].

Computing the eigenvalues of an arbitrary $N \times N$ matrix requires $O(N^3)$ operations. However, for linear covariance function $\mathbf{K} = \mathbf{C}\mathbf{C}^T, \mathbf{C} \in \mathbb{R}^{N \times K}$ where $K << N$, a more efficient implementation is possible [184]. Note that for all matrices $\mathbf{A}$, one can show that

[1] `https://github.com/limix/chiscore`

eigenvalues($\mathbf{A}^T\mathbf{A}$) = eigenvalues($\mathbf{A}\mathbf{A}^T$) [101]. From this it follows that

$$\text{eigenvalues}(\frac{1}{2}\hat{\mathbf{V}}_0^{-1/2}\mathbf{K}\hat{\mathbf{V}}_0^{-1/2}) = \text{eigenvalues}(\frac{1}{2}(\mathbf{C}^T\hat{\mathbf{V}}_0^{-1/2})^T(\mathbf{C}^T\hat{\mathbf{V}}_0^{-1/2})) \tag{3.25}$$

$$= \text{eigenvalues}(\frac{1}{2}(\mathbf{C}^T\hat{\mathbf{V}}_0^{-1/2})(\mathbf{C}^T\hat{\mathbf{V}}_0^{-1/2})^T) \tag{3.26}$$

$$= \text{eigenvalues}(\frac{1}{2}\mathbf{C}^T\hat{\mathbf{V}}_0^{-1}\mathbf{C}), \tag{3.27}$$

which can be computed in $O(K^3)$ once the matrix product has been evaluated.

scDALI-Joint

The scDALI-Joint test builds on the approach originally developed for SKAT-O [113] and recently extended in [184] to jointly test for heterogeneous and homogeneous effects using a sequence of single-parameter variance component tests. To this end, I modify the original scDALI model 3.5-3.7 by making the homogeneous effect α a Gaussian random variable $\alpha \sim \mathcal{N}(0, \sigma^2_{hom})$ such that 3.7 becomes

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2_{hom}\mathbf{1}\mathbf{1}^T + \sigma^2_{het}\mathbf{K}). \tag{3.28}$$

The above may equivalently be written as

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2_{tot}[(1-\rho)\mathbf{1}\mathbf{1}^T + \rho\mathbf{K}]). \tag{3.29}$$

where $\sigma^2_{tot} = \sigma^2_{het} + \sigma^2_{hom}$ and $\rho = \sigma^2_{het}/\sigma^2_{hom} \in [0, 1]$. Intuitively, if $\mathbf{K}$ has been normalized appropriately (see section 2.1.7), σ^2_{tot} can be interpreted as the total variance explained by allele-specific effects, whereas ρ quantifies the relative extent of heterogeneous imbalance. Under the modified scDALI model 3.29, the scDALI-Joint null and alternative hypotheses may now be written as

$$H_0^{joint} : \sigma^2_{tot} = 0 \text{ vs. } H_1^{joint} : \sigma^2_{tot} > 0. \tag{3.30}$$

Note that for fixed ρ, a suitable test statistic can be constructed equivalently to scDALI-Het, assuming a modified kernel matrix

$$\mathbf{K}_\rho = (1-\rho)\mathbf{1}\mathbf{1}^T + \rho\mathbf{K}. \tag{3.31}$$

In practice, however, ρ is unknown. Following [113, 184], I perform a grid search over $\rho \in [0, 1]$ and aggregate the associated p-values:

1. For a pre-defined sequence of values $\rho_r \in [0, 1]$, compute p-values P_r using the scDALI-Het test with modified covariance matrix 3.31. As in [113,184], I replace the exact Davies method with Liu's modified moment matching approximation [115,203] when evaluating the null distribution 3.24, in order to improve computational efficiency.

2. Compute the combined statistic $T = \min_r P_r$. For the case of a LMM with linear covariance function $\mathbf{K} = \mathbf{C}\mathbf{C}^T$, Moore et al. [184] derived the distribution of T under the null and provided a numerical method for approximating the p-value. Assuming a normal-theory approximation to the scDALI model, I use this approach for all subsequent analyses. A more general but less powerful alternative is to use T directly, following a Bonferroni adjustment for the number of tested values ρ_r.

As a byproduct of the scDALI-Joint testing procedure, the value of ρ associated with the smallest p-value provides an estimate of the extent of allelic imbalance explained by heterogeneous effects.

scDALI-Hom

As a special case, the scDALI-Joint test can be used to identify homogeneous allelic imbalance by fixing $\rho = 0$ in 3.29. This test additionally assumes the absence of heterogeneous effects ($\sigma^2_{het} = 0$) and therefore evaluates a slightly different hypothesis from 3.16. However, by avoiding to fit a random effect model with unconstrained cell-state covariance $\sigma^2_{het}\mathbf{K}$, this test significantly reduces the computational burden and allows for fast screening of homogeneous effects even in large datasets.

3.1.5 Cell-state inference from open chromatin data

The application of scDALI to empirical data requires a suitable cell state definition $\mathbf{C} \in \mathbb{R}^{N\times K}$. Common analysis strategies for single-cell sequencing data rely on clustering algorithms to identify discrete cell types from the observed expression profiles [174]. In principle, scDALI can be used to identify differences in allelic rates between clusters, using one-hot-encoded cluster labels to define a block diagonal covariance matrix. However, such a representation will disregard covariation between clusters and likely fail to accurately represent continuous biological processes such as cellular differentiation or development. Instead, a lower-dimensional embedding of the total count matrix can be used as a general-purpose

approach for capturing both discrete and continuous effects. Variational autoencoders have emerged as a versatile, non-linear framework for latent space inference from scRNA-seq count data. In this subsection I describe a VAE variant adapted to open chromatin data (chapter 1; **Fig.** 1.2). Notably, the model integrates sampling times for developmental datasets to estimate a continuous pseudo-temporal ordering from few observed time points.

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is an experimental technique for assaying chromatin accessibility on a genome-wide scale [35]. This method uses a genetically modified, hyperactive Tn5 transposase enzyme, to cleave DNA in accessible regions of the genome and tag the insertion site by ligating adapters for high-thoughput sequencing. The tagged insertion ends are then purified, PCR-amplified and sequenced. Peaks of accessibility, genomic regions enriched with aligned sequencing reads, can then be identified using computational methods [204]. Building on advances in split-pool combinatorial indexing technologies and droplet microfluidics, allowing for efficient molecular barcoding and identification of nucleic acids from a large pool of fixed cells or nuclei, the ATAC-seq protocol has been adapted to single cells [68, 96, 205].

The analysis of single-cell ATAC-seq (scATAC-seq) data presents several unique challenges compared to scRNA-seq data. While single cells typically carry several transcripts of any expressed gene, there are only few copies of DNA (two in diploid organisms). Therefore, the probability of observing insertions in any particular genomic region is very low, leading to inherent data sparsity at the per-cell level. In scATAC-seq data, only a small fraction, typically ranging from 1% to 10%, of the expected accessible chromatin regions are detected, compared to 10% to 45% of detected expressed genes in scRNA-seq [195]. Furthermore, while the quantification of mRNA from scRNA-seq is typically performed at the gene level, the definition of informative features for scATAC-seq is less straightforward. Common analysis pipelines quantify sequencing read counts in peaks of accessibility, inferred from pseudo-bulk aggregates using methods developed for bulk ATAC-seq data [195]. Accessibility can also be assessed at specific genomic regions such as transcription factor (TF) binding sites or gene promoters. Alternatively, the data can be summarized using sequence features of accessible regions, such as k-mer frequencies or TF motifs. The resulting cell-feature matrix may either be used directly to infer cell types and states or provided as input to methods for dimensional reduction, such as cisTopic [206], (based on latent dirichlet allocation) or latent semantic indexing (LSI) [68, 195].

VAE generative model

Suppose $\mathbf{X}$ summarizes the observed open chromatin profiles at M peaks of accessibility in N cells. That is, x_{nm} corresponds to the number of reads overlapping region m in cell n. Typically, the number of reads in each region will be low. In fact, it is common practice to binarize $\mathbf{X}$ such that $x_{nm} \in \{0, 1\}$, which was found to reduce technical noise associated with the sequencing process [195]. Let s_n be a batch identifier for cell n. I assume the following generative model for binary accessibility data:

$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_K) \tag{3.32}$$

$$l_n \,|\, s_n \sim \text{LogNormal}(l_\mu(s_n), l_{\sigma^2}(s_n)) \tag{3.33}$$

$$\boldsymbol{\rho}_n = f_\rho(\mathbf{z}_n, s_n) \tag{3.34}$$

$$x_{nm} \,|\, \rho_{nm}, l_n \sim \text{Bernoulli}\Big(1 - (1 - \rho_{nm})^{l_n}\Big). \tag{3.35}$$

Here, $\mathbf{z}_n$ are latent cell states and l_n is a cell-specific size-factor variable capturing effects associated with variable sequencing depth. As in the scVI model [131] discussed in section 2.2.3, l_n follows a LogNormal distribution with parameters $l_\mu(s_n), l_{\sigma^2}(s_n)$, corresponding to the empirical mean and variance of the log-library size (log of total non-binary counts) per batch. By conditioning the generative process on (latent) size factors and batch identifiers, the model is encouraged to remove technical nuisance variation from the cell latent space. The latent representation $\mathbf{z}_n$ and batch identifiers are mapped to $\boldsymbol{\rho}_n$ using a fully-connected neural net f_ρ with softmax activation, such that ρ_{nm} is the relative activity of peak m. The distribution over binary accessibility profiles is a function of the scaling factor l_n and relative peak activities $\boldsymbol{\rho}_n$. Intuitively, if l_n were the true (discrete) number of of reads in cell m, $1 - (1 - \rho_{nm})^{l_n}$ is the probability of observing at least one read in peak m. However, to simplify inference, l_n is modeled as a continuous variable.

The developmental dataset considered later in this chapter additionally includes coarse temporal labels. *Drosophila Melanogaster* embryos were collected during different time windows after egg laying, to study changes in accessibility at major developmental stages. I model the observed time stamps as noisy realizations of an underlying continuous process. Suppose the time label y_n associated with cell n takes on values in $\{1, 2, \ldots, T\}$. I model the relative order of cells along the developmental trajectory as a function of $\mathbf{z}_n$ and draw y_n from an ordinal distribution [207],

$$p(y_n \,|\, \mathbf{z}_n) = \Phi\Big(w_{y_n} - f_y(\mathbf{z}_n)\Big) - \Phi\Big(w_{y_n - 1} - f_y(\mathbf{z}_n)\Big), \tag{3.36}$$

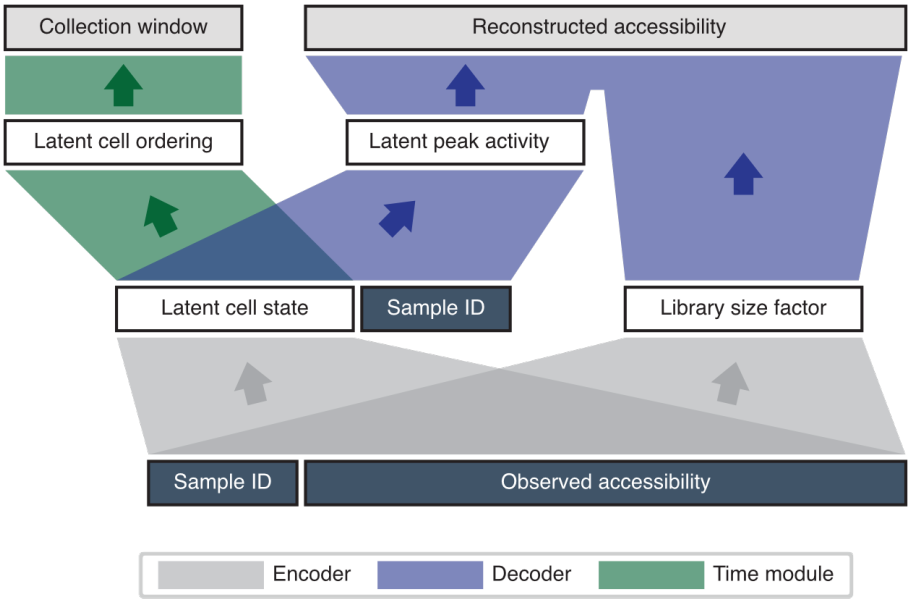

Figure 3.2. Cell-state VAE overview. The encoder network maps observed chromatin accessibility profiles to a lower dimensional cell state space and latent size factors, while removing variation associated with sample IDs or other categorical variables, e.g., batch. The decoder maps latent variables to the relative peak activities and reconstructs observed profiles under a size-factor adjusted Bernoulli likelihood. The temporal module infers a continuous temporal ordering of cell states from few observed discrete time labels using an ordinal likelihood model.

where Φ is the cumulative distribution function of the standard normal distribution. The parameters w_t divide the real line, where I fix $w_0 = -\infty$, $w_T = \infty$ and $w_1, \ldots, w_{T-1}$ are optimized subject to $w_t < w_{t+1}$. The function f_y models the latent time as a function of the cell state $\mathbf{z}_n$. Ordinal labels t are identified with consecutive intervals (w_{t-1}, w_t), allowing for varying rates of cell state changes across time. A draw from this model corresponds to the unique interval containing a noisy realization $f_y(\mathbf{z}_n) + \epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$ of the actual latent time $f_y(\mathbf{z}_n)$. To improve model interpretability, I choose $f(\mathbf{z}_n)$ to be a linear function, thereby constraining the developmental changes in cell states to align with a single axis of variation in the latent space.

Variational approximation

The marginal distribution under the non-linear generative model 3.32-3.35 and 3.36 are intractable. I use stochastic variational inference [144] to approximate the true posterior distribution over latent variables and optimize the model parameters (section 2.2.1). Following [131], I use a mean-field factorization for the approximate latent posterior,

$$q(\mathbf{z}_n, l_n \,|\, \mathbf{x}_n, s_n) = q(\mathbf{z}_n, \,|\, \mathbf{x}_n, s_n) q(l_n \,|\, \mathbf{x}_n, s_n) \tag{3.37}$$

where $q(\mathbf{z}_n, \,|\, \mathbf{x}_n, s_n)$ and $q(l_n \,|\, \mathbf{x}_n, s_n)$ are chosen to be multivariate normal and LogNormal, respectively. Model and variational parameters can be optimized by stochastic gradient ascent on the variational lower bound

$$\log p(\mathbf{x}_n) \geq \mathbb{E}_{q(\mathbf{z}_n \,|\, \mathbf{x}_n, s_n) q(l_n \,|\, \mathbf{x}_n, s_n)} [\log p(\mathbf{x}_n \,|\, \mathbf{z}_n, l_n, s_n) + \log p(y_n \,|\, \mathbf{z}_n)] \tag{3.38}$$

$$- \mathrm{KL}(q(\mathbf{z}_n \,|\, \mathbf{x}_n, s_n) \,||\, p(\mathbf{z}_n)) - \mathrm{KL}(q(l_n, \,|\, \mathbf{x}_n, s_n) \,||\, p(l_n)). \tag{3.39}$$

The complete generative model and variational approximation is visualized in **Fig.** 3.2.

3.1.6 Application to scATAC-seq from developing Drosophila embryos

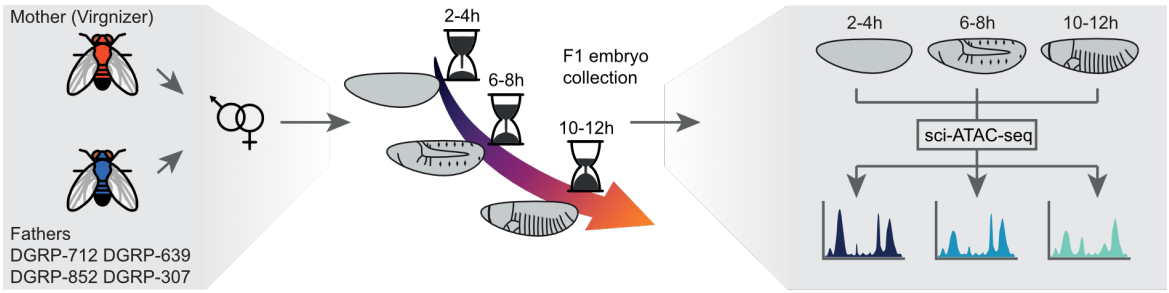

Figure 3.3. scATAC-seq of developing *Drosophila Melanogaster* embryos. F1 embryos from crosses of the same mother and four genetically distinct fathers were collected at three different timepoints (2-4h, 6-8h and 10-12h after egg laying). Chromatin accessibility was profiled using the sci-ATAC-seq protocol.

I applied the scDALI model to study allele-specific patterns of genomic accessibility in F1 hybrid embryos of representing the first filial generation of offspring from the same mother and four genetically distinct fathers [191] (**Fig.** 3.3) Embryos from these four crosses were

collected at three different time windows after egg laying (2-4h, 6-8h and 10-12h), approximately corresponding to major developmental stages where cells are multipotent, undergo lineage commitment and or tissue differentiation, respectively. Chromatin accessibility in single-cells was profiled using sci-ATAC-seq, a scATAC-seq protocol based on single-cell combinatorial indexing (sci-) [68, 96]. Briefly, sci-ATAC-seq splits cells in multiple wells and tags nucleic acids using molecular barcodes specific to each well. Cells are then pooled and the procedure is repeated, such that the probability of observing cells with colliding barcodes is below a desired threshold. A detailed description of the experimental process is included in appendix A.1.

Processing of raw sci-ATAC-seq data

I processed the raw data generated by the sequencing machines based on the pipeline[2] developed by Cusanovich et al. [68]. BCL files were converted to fastq files using the bcl2fastq tool, v.2.16 (Illumina). To correct sequencing and PCR amplification errors I matched each read barcode against all possible barcodes produced by the split-pool procedure. Approximate matches (Levenshtein distance < 3 and distance to next best match > 2) were fixed to the closest matching reference barcode. Ambiguous and unknown barcodes were discarded. Reads were trimmed using Trimmomatic [208] and aligned to the dm6 reference genome using Bowtie 2 [209] with options `-X 2000 -3 1`. Subsequently, PCR duplicates were removed. To separate barcodes corresponding to genuine cells from background noise, I fit a two-component Gaussian mixture model to the log-transformed read counts per barcode (**Fig.** 3.4). Barcodes were classified as noise, when the posterior probability of belonging to the mixing component with higher read depth was below 95%. The histogram of DNA fragment sizes for the processed data (determined from the paired-end sequencing reads) exhibited the typical nucleosome banding pattern corresponding to the length of DNA wrapped around a single nucleosome (**Fig.** 3.4). Furthermore, the read count distribution for each of the 12 sequenced libraries was consistent with existing sci-ATAC-seq data of time-matched *Drosophila Melanogaster* embryos from a reference strain published by Cusanovich et al. [68].

[2] `https://github.com/shendurelab/fly-atac/)`

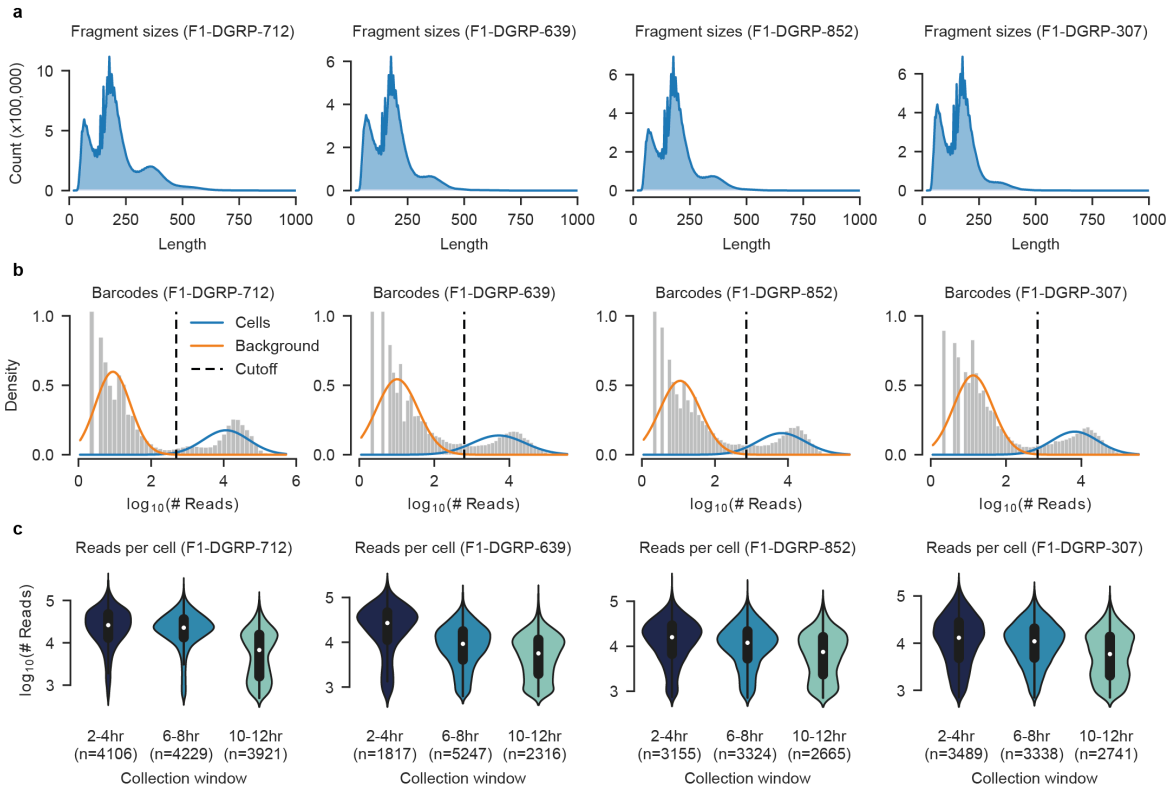

Figure 3.4. QC for *Drosophila Melanogaster* sci-ATAC-seq data. **(a)** Fragment length measured using paired-end sequencing reads for all crosses. Peaks correspond to the expected nucleosomal banding pattern. **(b)** Distribution of log-total read counts associated with each barcode. To separate barcodes corresponding to real cells from background noise, a two-component Gaussian mixture model is fitted (orange and blue lines). Barcodes to the left of the dotted line, corresponding to the 95% posterior probability of belonging to the foreground mixture component, were discarded. **(c)** Violin plots of the distribution of read counts for all cells, stratified by the time window of sample collection (2-4, 6-8 and 10-12 hours after egg laying).

Cell state inference from total counts

Chromatin accessibility was quantified in single cells by computing the number of reads overlapping any of 53,133 genomic regions previously identified as peaks of accessibility in the matched Cusanovich et al. dataset [68] and lifted[3] to the dm6 *Drosophila Melanogaster* reference genome. Aggregated pseudo-bulk counts for each collection window were highly correlated, both between crosses, as well as when compared to the published reference dataset (**Fig.** 3.5). To exclude biases in chromatin accessibility associated with sex chromosomes [68], I restricted all further analyses to peaks located on autosomes. Cells were filtered based on the cell-count distribution in each cross and collection window, retaining only those cells whose counts were within the 10% and 99% quantiles. This resulted in 35,485 high-quality cells in total. I further limited the feature set to the 25,000 top most accessible peaks.

The binarized accessibility matrix was used to train the cell-state variational autoencoder model described in section 3.1.5. One-hot-encoded labels for each cross were used as 'batch' variables, to remove broad inter-individual and technical effects from the cell state space. Encoder and decoder networks were implemented using batch normalization [210] and ReLU activation functions. Parameters were optimized using ADAM [211] for 30 epochs, using the 1cycle learning rate policy proposed in [212] with a maximum learning rate of 10^{-2}. The conditional likelihood in 3.38-3.33 was approximated using 5 Monte Carlo samples. To dampen the influence of the KL divergence on the latent space inference at the early stages of training, the KL term was scaled by a factor of $i/25$ when computing the ELBO in epoch i. The number of hidden nodes, layers and the dimension of the latent space were tuned by maximizing the held-out log likelihood on 20% of the cells. Parameter choices for the best-performing model can be found in **Table** 3.1.

Visual inspection of a two-dimensional UMAP (Uniform Manifold Approximation and Projection) [213] projection of the inferred cell latent space (**Fig.** 3.6a-d) confirmed that cells from different crosses were well-mixed and the embedding captured progressive changes across embryonic development. Cell clusters in the VAE latent space were identified using the Leiden algorithm [100, 214] with a resolution of 1.2, resulting in 28 clusters (**Fig.** 3.6e). I trained logistic regression models to discriminate each cluster based on the relative peak activity inferred by the VAE model [100, 215]. In order to link clusters to known cell types

[3] `https://github.com/FlyBase/bulkfile-scripts`

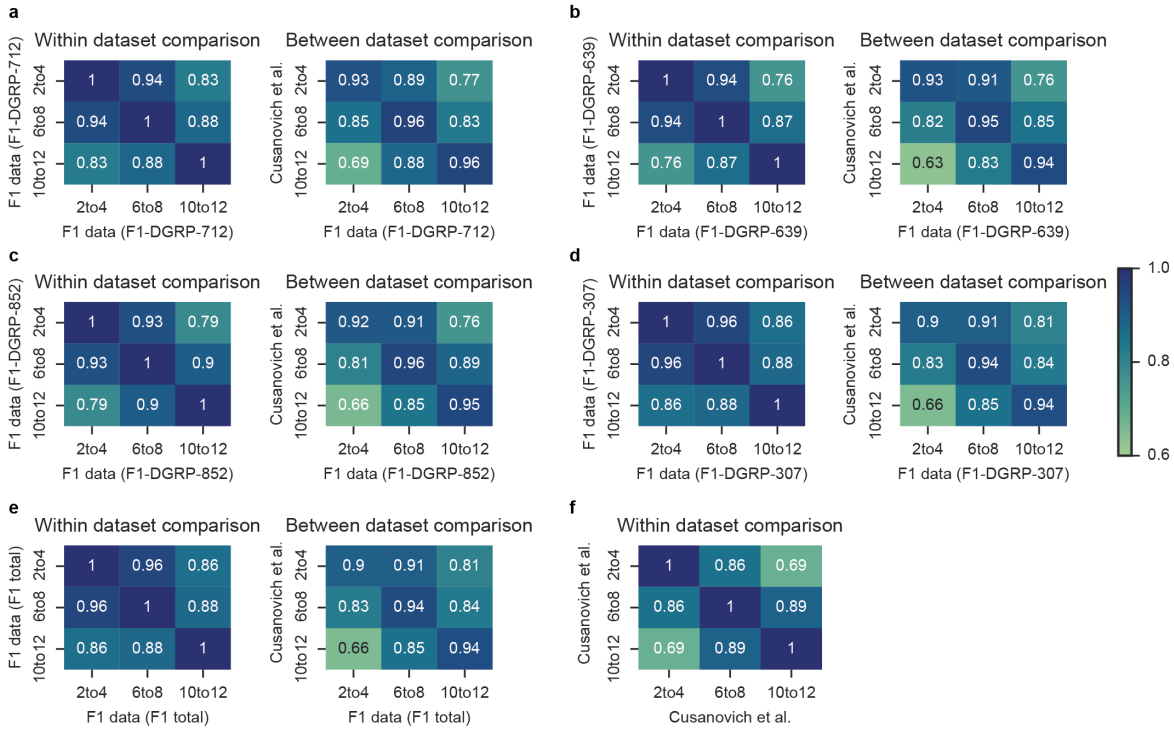

Figure 3.5. Comparison with published sci-ATAC data. Pearson correlation between pseudo-bulk aggregates of sci-ATAC-seq reads. **(a-d)** Correlation between different sample collection windows, both for crosses within our dataset as well as compared to published, time-matched sci-ATAC-seq data from a reference strain [68]. Matched timepoints are highly correlated. **(e)** Pearson correlation for the combined dataset (aggregated over all four crosses). **(f)** Within dataset comparison for the Cusanovich et al. data.

Table 3.1. Cell-state VAE hyper-parameters

Parameter	Value
Cell state dimension k	8
Hidden layers for the encoder $q(\mathbf{z}_i \mid \mathbf{x}_i, c_i)$	[256, 128]
Hidden layers for the encoder $q(l_i, \mid \mathbf{x}_i, c_i)$	[256]
Hidden layers for the decoder $f_\rho(\mathbf{z}_i, c_i)$	[64, 128]
Hidden layers for time module $f_z(\mathbf{z}_i)$	None

and tissues, peaks were ranked in each cluster based on the inferred regression coefficients, followed by an enrichment analysis for enhancer elements with validated *in vivo* spatio-temporal activity in specific tissues during embryogenesis (CAD4 database [68]) and tissue-

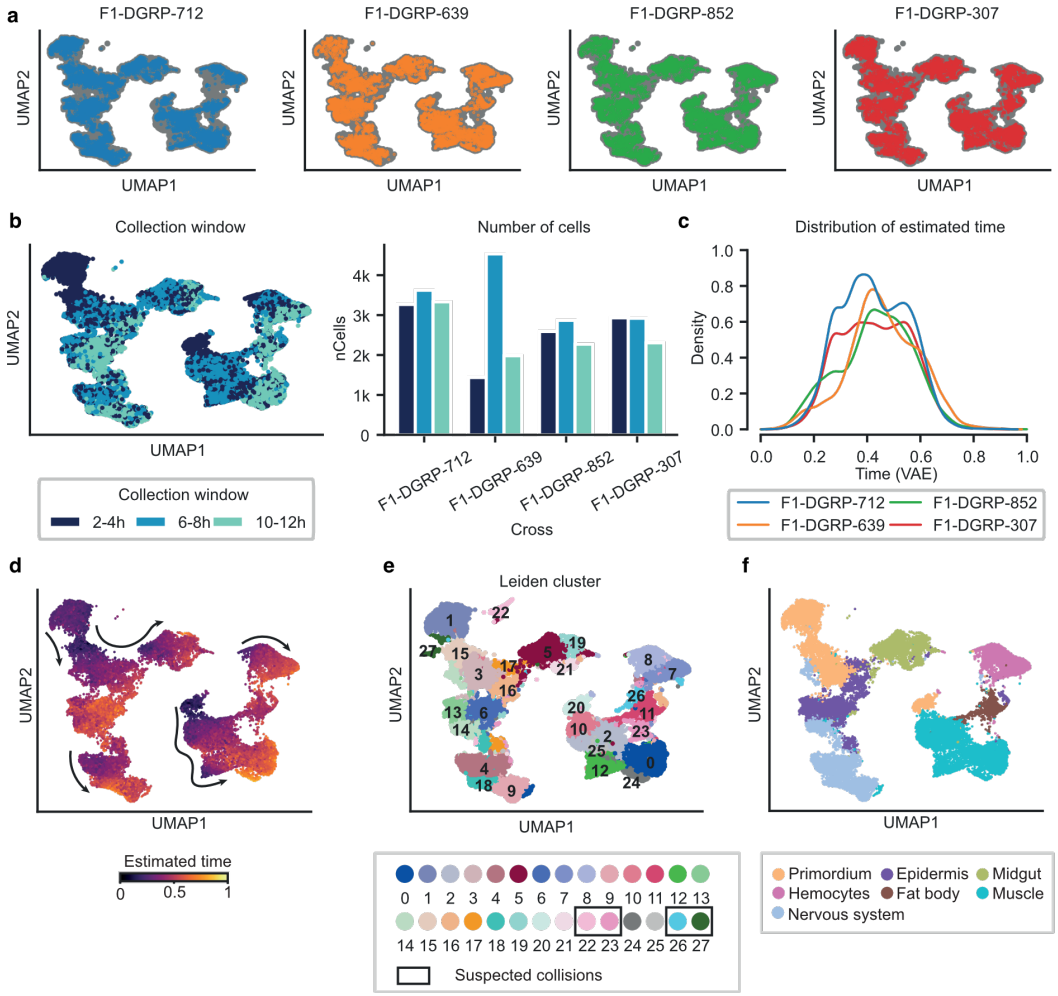

Figure 3.6. Cell-state inference from VAE embedding. **(a)** 2-dimensional UMAP embedding of VAE latent space. Cells do not cluster by cross, but are uniformly integrated. **(b)** Distribution of cells associated with three different sample collection windows. **(c)** Distribution of the estimated continuous temporal ordering. **(d)** Cell latent space colored by estimated latent time. **(e)** Leiden clusters inferred from VAE embedding. Clusters 22, 23, 26, 27 were hypothesized to be the result of barcode collisions and removed from later analyses. **(f)** An enrichment analysis of differential peaks in each cluster for known enhancer elements and genes with tissue-specific expression resolves major cell lineages.

specific genes (tissue-specific expression of the nearest gene based on in situ hybridization data from the Berkeley Drosophila Genome Project[4] and FlyBase gene expression annotations[5]) using Fisher's exact test. Based on these enrichments, Stefano Secchia generated an assignment of clusters to one of seven cell populations. This annotation resolved major embryonic lineages, including muscle, nervous system and ectoderm (**Fig.** 3.6f), demonstrating that the cell-state VAE inferred a biologically meaningful latent space. Four clusters (1,432 cells) could not be annotated unambiguously, and were hypothesized to be the result of barcode collisions (**Fig.** 3.6e). Following the removal of these four clusters, I obtained a final dataset of 35,485 cells for downstream analyses.

Generation of allele-specific counts

I re-processed the filtered sequence alignments to quantify chromatin accessibility on an allele-specific level. Mapping artifacts are an important source of confounding in allele-specific analyses [188]. When aligning reads to a reference genome, a genetic mutation not present in the reference may be mistaken for a sequencing error, causing the read to be discarded. In some cases, such a read may even map to a completely different position in the genome. Conversely, sequencing reads from the reference allele tend to be aligned with higher confidence. If not accounted for, reference mapping biases can lead to false positive discoveries in allele-specific analyses. Using existing genotyping data for the parental strains [191], I created cross-specific VCF files containing heterozygous genetic variants present in the F1 generation. I then employed the WASP pipeline[6] [188] to mitigate allelic mapping artifacts. Briefly, for all reads overlapping a particular sequence variation, WASP checks if altering the variant allele affects the read alignment. If this is the case, the read is discarded. The WASP filter led to the exclusion of approximately 7-8% of mapped reads from the original alignment (**Fig.** 3.7a). To quantify allele-specific chromatin accessibility, I adapted the original WASP code for count generation from bulk data to the single cell setting[7]. I focused on 1 kilobase (kb) windows centered on each of the 53,133 peaks identified by Cusanovich et al. [68] to mitigate the inherent sparsity of the data. Reads were allocated to specific alleles if they overlapped at least one heterozygous single-nucleotide variant. In cases where reads overlapped with multiple variants, one variant was randomly selected to

[4] `http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl`

[5] `http://flybase.org/)`

[6] `https://github.com/bmvdgeijn/WASP/tree/master/mapping`

[7] `https://github.com/tohein/scai_utils`

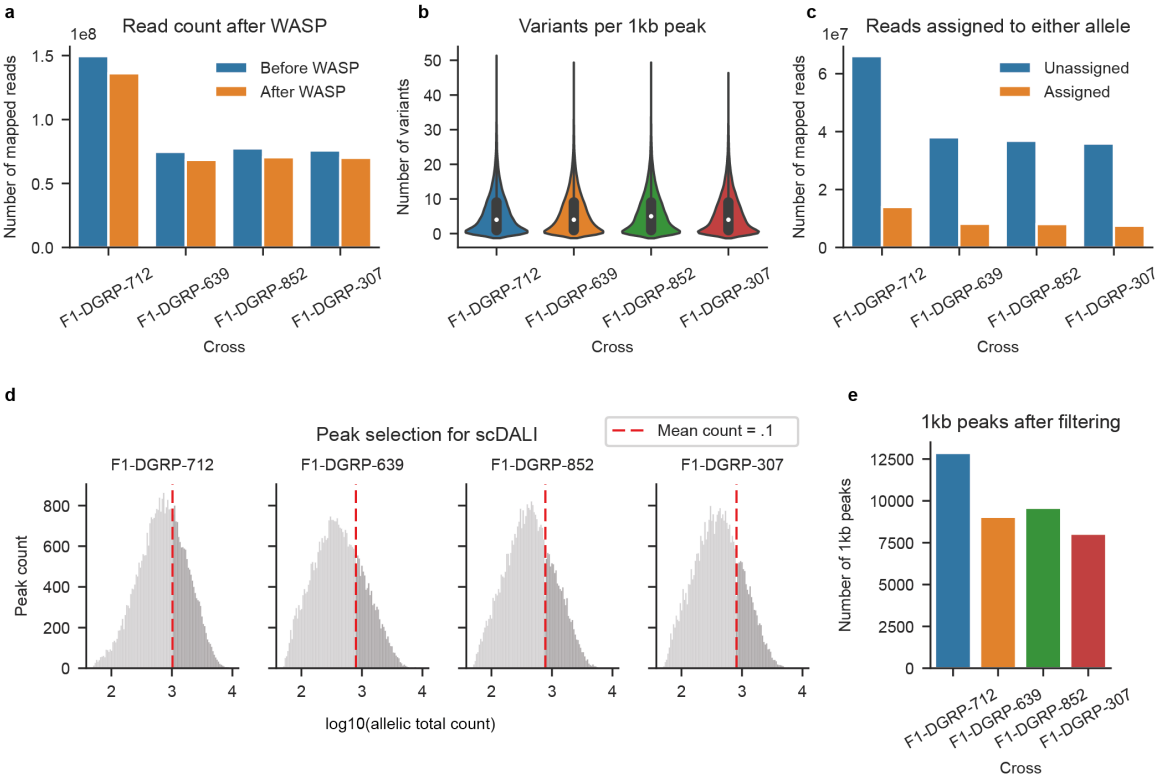

Figure 3.7. QC for allele-specific quantifications. **(a)** Number of sequencing reads before and after applying the WASP pipeline to correct for reference mapping biases. **(b)** Distribution of the number of variants in 1kb windows centered on peaks of accessibility in each cross. **(c)** Number of reads that could or could not be assigned to either allele. Approximately 20% of reads can be used to quantify allele-specific chromatin accessibility. **(d)** Histogram of allelic log-total read counts (number of reads that could be assigned to either allele across cells). High quality peaks were filtered by requiring that the average allelic total count was greater than .1 (red dotted line) **(e)** Number of 1kb peaks after filtering, resulting in a combined set of 39,530 peaks to be tested for allelic imbalance.

determine the allele of origin. Using this process, 20% of reads could be asssigned to either haplotype (using on 5–6 variants per region on average, **Fig.** 3.7b, c). Due to the difficulty of accurately estimating the allelic base rate for sex chromosomes (corresponding to the proportion of female embryos in the sample), I restricted all further analysis to autosomes. Additionally, I filtered windows in each cross based on allelic coverage (mean count of reads assigned to either allele < 0.1), resulting in between 8040 and 12,861 peaks per cross and a combined set of 39,530 peaks to be tested for allelic imbalance (**Fig.** 3.7d, e).

Empirical validation of scDALI on simulated data

I initially validated the scDALI approach by simulating from the model 3.5-3.7 using the inferred cell state representations (VAE embedding $\mathbf{C}_{\text{VAE}} \in \mathbb{R}^{N\times 8}$ and one-hot-encoded Leiden clusters derived from the VAE embedding $\mathbf{C}_{\text{Leiden}} \in \mathbb{R}^{N\times 24}$) and observed allelic total counts for one of the F1 crosses (F1-DGRP-712, N=10220 cells, 12,861 peaks). Using a linear covariance function, simulation cell-state covariance matrices were defined as (**Fig.** 3.8)

$$\mathbf{K}_{\text{VAE}} = \mathbf{C}_{\text{VAE}}\mathbf{C}_{\text{VAE}}^T \tag{3.40}$$

$$\mathbf{K}_{\text{Leiden}} = \mathbf{C}_{\text{Leiden}}\mathbf{C}_{\text{Leiden}}^T \tag{3.41}$$

and normalized (section 2.1.7.

I assessed the degree of extra-binomial variation in the observed allele-specific data, by fitting a Beta-Binomial model using no additional cell state information (**Fig.** 3.9a). Based on the histogram of estimated dispersion parameters for all peaks, I ran all simulations at $\theta \in \{2, 5\}$. First, I evaluated the calibration of all three scDALI tests, by simulating from their respective null models. When simulating neither heterogeneous nor homogeneous imbalances, the p-values from both scDALI-Joint (with $\mathbf{K}_{VAE}$) and scDALI-Hom approximately followed the expected uniform distribution (**Fig.** 3.9b). I then simulated different levels of homogeneous imbalance $\alpha \sim \mathcal{N}(0, \sigma^2_{hom}$, $\sigma^2_{hom} \in \{0.01, 0.05, 0.1)\}$ and assessed the calibration of scDALI-Het (testing for heterogeneous effects using $\mathbf{K}_{VAE}$) as well as two baseline candidates: a one-way ANOVA test to compare empirical allelic rate between clusters and variant of scDALI using a Binomial rather than Beta-Binomial likelihood. While both the scDALI-Het and ANOVA tests were found to be calibrated for all simulated levels of overdispersion and homogeneous imbalance, using the Binomial likelihood model led to an inflated p-value distribution (**Fig.** 3.9c). Additionally, I compared scDALI-Het with

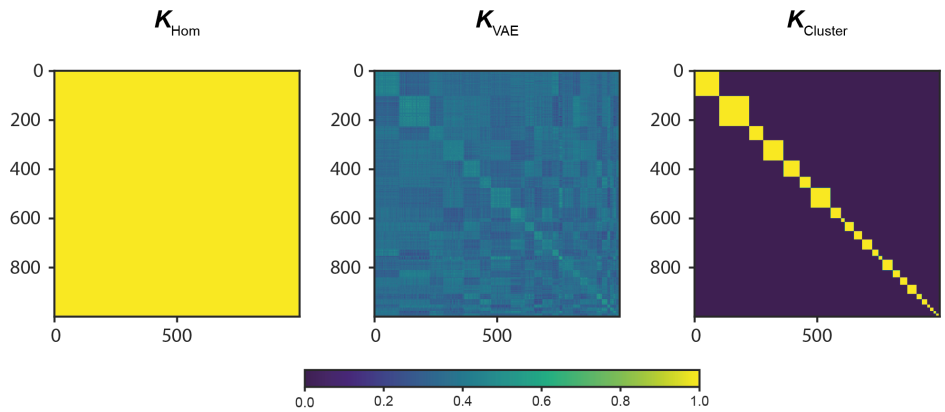

Figure 3.8. Cell-state covariance matrices for the simulation procedure. Left: Homogeneous effects (constant matrix of ones). Middle: Linear covariance based on VAE latent space inferred from real sci-ATAC-seq data, restricted to 10,200 cells from cross F1-DGRP-712. Right: Block-diagonal matrix indicating cell membership to one of 24 Leiden clusters in the VAE latent space. All matrices have been subsampled to 1,000 cells for the purpose of visualization.

$\mathbf{K}_{Leiden}$ and ANOVA to an ordinary linear regression model for empirical allelic rates, which incorporated one-hot-encoded Leiden clusters as fixed effect covariates (multiple-degrees-of-freedom likelihood-ratio test, OLS-LRT). Models were fitted to varying numbers of cells ($N \in \{250, 500, 1000, 5000\}$) and considering an increasing number of cell-state dimensions ($K \in \{5, \ldots, 24\}$). Allele-specific counts were simulated for a subset of 1,000 peaks. **Fig.** 3.10d shows the inflation factor,

$$\frac{\log_{10}(\text{median}P)}{\log_{10}(0.5)} \tag{3.42}$$

quantifying the deviation from the expected median p-value under the null (averaged across 25 random initializations). Consistent with previous results on multiple-degrees-of-freedom tests in fixed effect models [184], I found the OLS-LRT to produce inflated test statistics when the number of tested cell state dimensions was large compared to the samples size. Neither scDALI-Het nor ANOVA suffered from the same issue. I therefore excluded OLS-LRT from further experiments.

Next, I evaluated the statistical power to detect homogeneous vs. heterogenoeus effects for the scDALI tests. I simulated allelic counts from the scDALI model, using the alternative formulation 3.29. For each combination of $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, the relative ex-

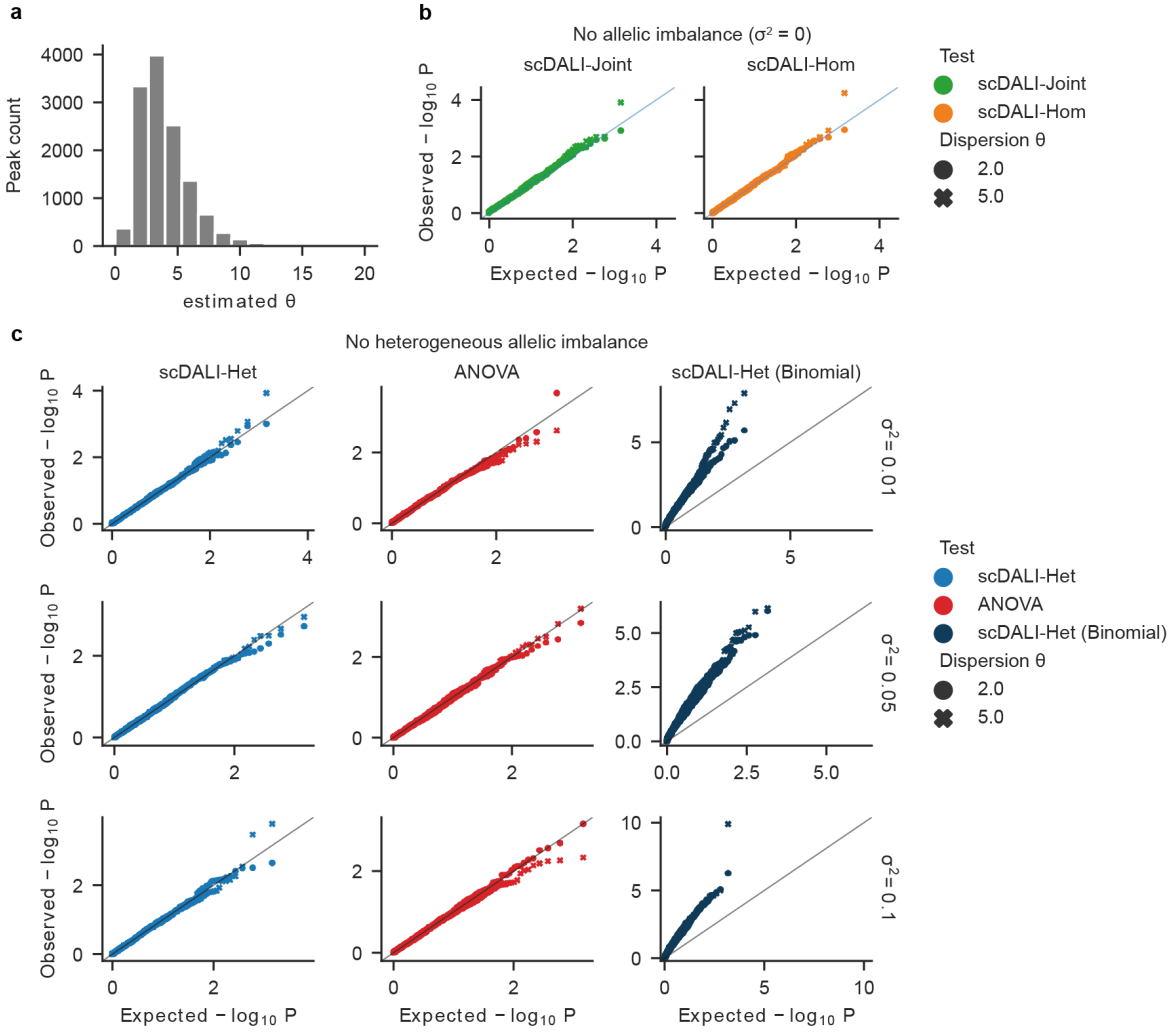

Figure 3.9. Statistical calibration of scDALI and alternative methods on simulated data. **(a)** Histogram of estimated dispersion parameters θ under a Beta-Binomial model for peaks of accessibility in real sci-ATAC-seq data from F1 cross F1-DGRP-712. Models were fitted without leveraging cell-state information. **(b, c)** Q-Q plots comparing the distribution of observed p-values to a uniform distribution when simulating different levels of homogeneous imbalance $\alpha \sim \mathcal{N}(0, \sigma^2)$ (on a logit scale) and overdispersion $\theta \in \{2, 5\}$. All synthetic datasets were generated without additional heterogeneous imbalance, based on observed allelic total read counts for 5,000 cells and 1,000 peaks from cross F1-DGRP-712. Shown are results for the scDALI tests, scDALI-Het (Binomial) a variant of scDALI-Het with Binomial likelihood and a one-way ANOVA model testing for differences in allelic rates between Leiden clusters. All scDALI variants were provided with the linear covariance matrix based on the VAE latent space inferred from real data (**Fig.** 3.8).

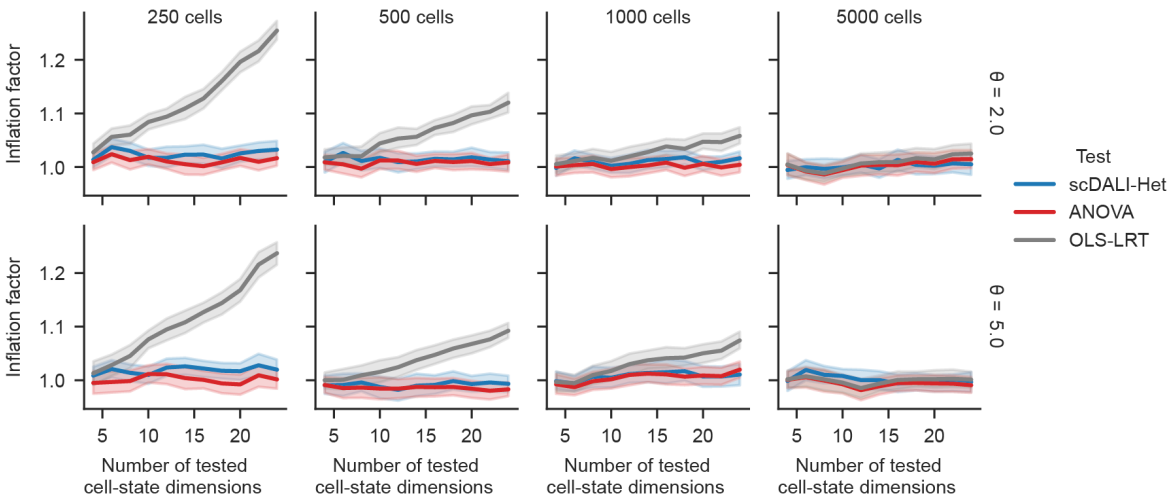

Figure 3.10. Statistical calibration as a function of the number of tested cell state dimensions (Leiden clusters) and sample size. Shown is the p-value inflation factor $\log_{10}(\text{median}P)/\log_{10}(0.5)$, averaged across 25 random seeds for the simulation procedure. Considered were scDALI-Het, one-way ANOVA and a multiple-degrees-of-freedom likelihood-ratio test based on a linear regression model which incorporated cell state variables as fixed effects (OLS-LRT). Both ANOVA and OLS-LRT were fitted to empirical allelic rates. While ANOVA and the scDALI tests were calibrated across all simulation scenarios, the OLS-LRT produced inflated test statistics when considering a large number of cell state dimensions compared to the sample size.

tent of heterogeneous imbalance, and $\sigma^2_{tot} \in \{0.01, 0.05, 0.1\}$, the total variance explained by allele-specific effects, I simulated data for 5,000 cells, using the observed allelic total counts for 1,000 ATAC peaks randomly chosen from the real sci-ATAC-seq data. Both the simulation procedure as well as the scDALI tests were run using $\mathbf{K}_{VAE}$ as a cell-state covariance matrix. I evaluated statistcal power at a significance level of 0.05 across 25 random seeds (**Fig.** 3.11a). As expected, scDALI-Joint successfully detected effects from both categories, allowing for a general assessment of both heterogeneous and homogeneous allelic imbalance from single cell data.

Lastly, I compared detection power for discrete vs. continuous heterogeneous effects. I simulated allelic counts counts assuming continuous cell states, discrete cell clusters derived from these states and a weighted combination thereof,

$$\mathbf{K} = \eta \mathbf{K}_{Leiden} + (1 - \eta)\mathbf{K}_{VAE} \tag{3.43}$$

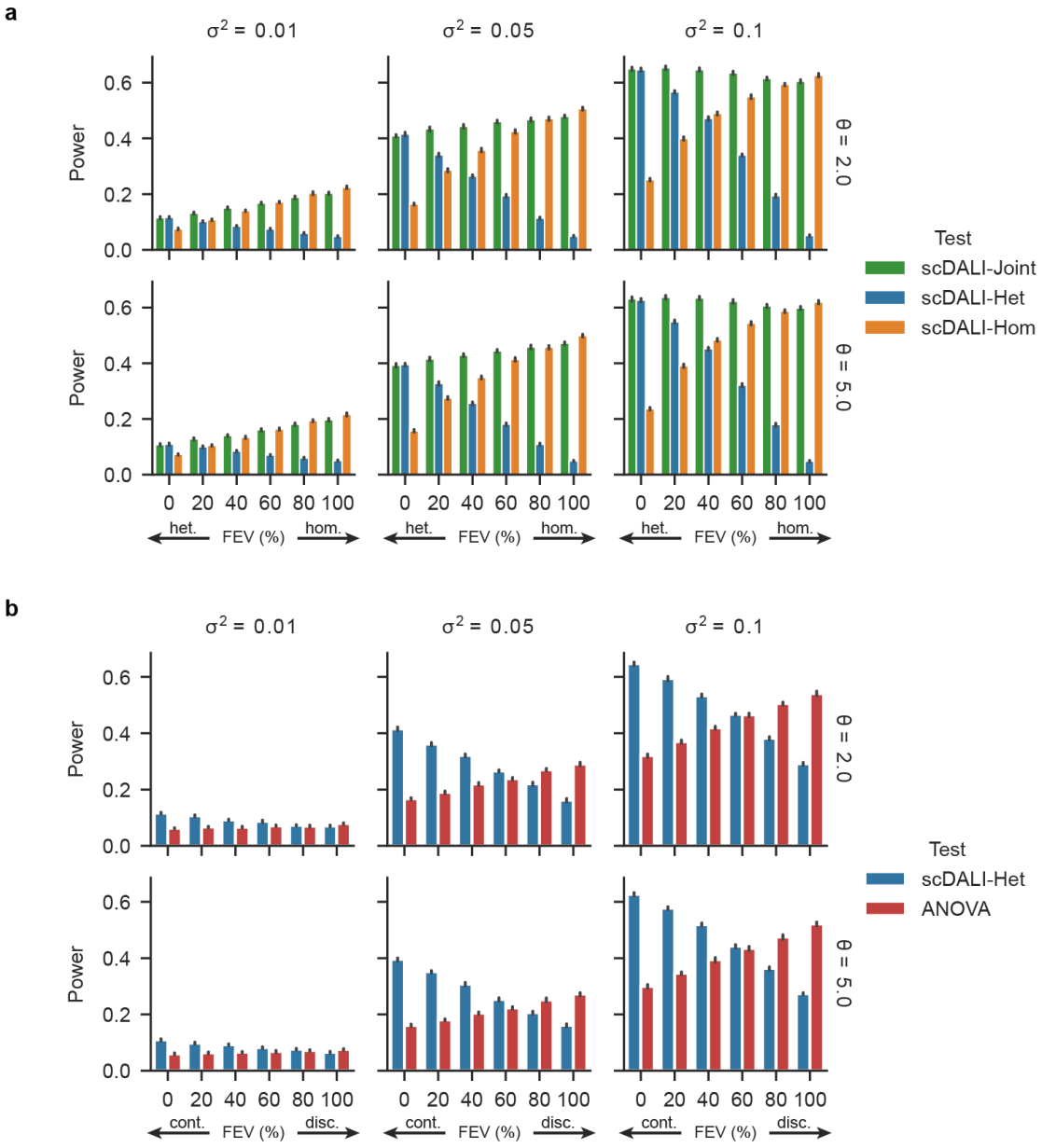

Figure 3.11. Power assessment on simulated data. Alternative counts were sampled from the scDALI model, considering different scaling factors σ^2 for the simulation covariance matrix (columns) and levels of overdispersion θ (rows). Data was generated for 5,000 cells, using observed allelic total counts for 1,000 ATAC peaks sampled from real sci-ATAC-seq data. All scDALI models used the linear covariance matrix based on the VAE latent space (**Fig.** 3.8). Power was evaluated as the fraction of positive findings, averaged across 25 random initializations. **(a)** Power to detect heterogeneous vs. homogeneous allelic imbalance for all scDALI tests. **(b)** Power to detect allele-specific variation associated with discrete vs. continuous cell states. scDALI-Het is compared to a one-way ANOVA model, assessing differences in empirical allelic rates between Leiden clusters.

Data was simulated from the scDALI model 3.5-3.7, varying the mixing coefficient $\eta \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and scaling parameter $\sigma^2_{het} \in \{0.1, 0.05, 0.1\}$ while assuming no additional homogeneous effects ($\alpha = 0$). I compared scDALI-Het, trained using $\mathbf{K}_{VAE}$, to the one-way ANOVA model based on discrete Leiden clusters. As before, data was simulated for 5,000 cells and 1,000 peaks and results were averaged across 25 random initializations (**Fig.** 3.11b). scDALI-Het offered substantial power advantages in the presence of strong to medium levels of continuous effects, whereas ANOVA was best suited to detect purely discrete effects.

Empirical runtime analysis

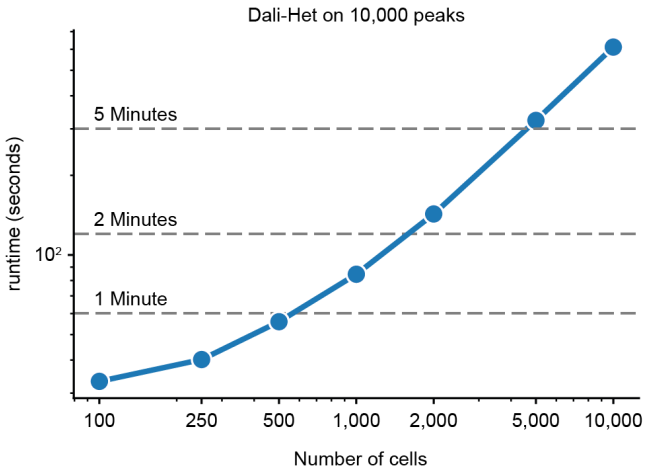

Figure 3.12. scDALI-Het runtime analysis. 10,000 randomly-chosen ATAC peaks were tested for heterogeneous allelic imbalance, using varying numbers of cells (with non-zero allelic total counts) from cross F1-DGRP-712. scDALI-Het scales linearly with the sample size. Runtimes were evaluated on a 2018 MacBook Pro with 2,3 GHz Quad-Core Intel Core i5 processor.

To evaluate the empirical runtime of scDALI-Het, I ran the model using 10,000 randomly selected peaks, using an increasing number of cells with non-zero allelic total counts from cross F1-DGRP-712. Runtimes were assessed on a 2018 MacBook Pro with 2,3 GHz Quad-Core Intel Core i5 processor (**Fig.** 3.12). scDALI-Het scaled linearly with the number of cells, enabling the analysis of large datasets with up to tens of thousands of cells.

Mapping allelic imbalance across development

Having validated scDALI using simulations, I applied the framework to detect allele-specific effects in 39,530 peaks of accessibility identified from the real sci-ATAC-seq profiles of *Drosophila Melanogaster* embryos, jointly considering cells from all developmental stages (**Fig.** 3.13 and **Fig.** 3.14). Test statistics for scDALI-Joint and scDALI-Het were constructed using the VAE latent space coordinates. To counteract the problem of multiple testing, p-values were adjusted using the Benjamini-Hochberg procedure [141] (section 2.1.9), to control the false discovery rate. Approximately 20% of tested peaks (7,823) showed evidence for allelic imbalance (scDALI-Joint, FDR < 0.1). The majority of these peaks were also identified by scDALI-Hom (83%). However, scDALI-Het discovered 415 peaks of accessibility with heterogeneous effects that could not be detected using scDALI-Hom, showing how cell-state-specific imbalances may be missed by methods that assume exclusive homogeneous effects such as (pseudo-) bulk approaches. As an example, a peak on chromosome 3 (chr3R:20310056-20311056) in cross F1-DGRP-307 showed strong evidence for heterogeneous imbalances, identified by both scDALI-Het ($P = 5.45 \times 10^{-8}$) and scDALI-Joint ($P = 1.93 \times 10^{-8}$), but these effects cancelled out when considering average deviations from allelic balance globally across all cells (scDALI-Hom $P = 0.81$, **Fig.** 3.13b).

I then assessed the robustness of scDALI-Het with respect to different cell state representations. I considered two alternative methods for inferring low-dimensional embeddings of scATAC-seq data, latent semantic indexing (LSI) [68], using the leading components 2 to 20 (the first component was excluded due to correlation with total counts per cell) and cisTopic [206] (50 topics). This comparison showed (**Fig.** 3.15a, b) that significant associations could be reliably identified when using either cell-state representation (VAE, LSI and cisTopic). To evaluate the possibility that my analysis of heterogeneous effects may be confounded by variation in total accessibility, I applied the scDALI-Het test using a covariance matrix defined by the outer product of the total count vector for each peak. Reassuringly, this test did not identify any significant associations for the vast majority of peaks (**Fig.** 3.15c, d).

Properties of regions with heterogeneous allelic imbalance

For each of the 415 peaks of accessibility with evidence for heterogeneous allelic imbalance identified by scDALI-Het (P < 0.1 FDR), I estimated the posterior distribution of allelic

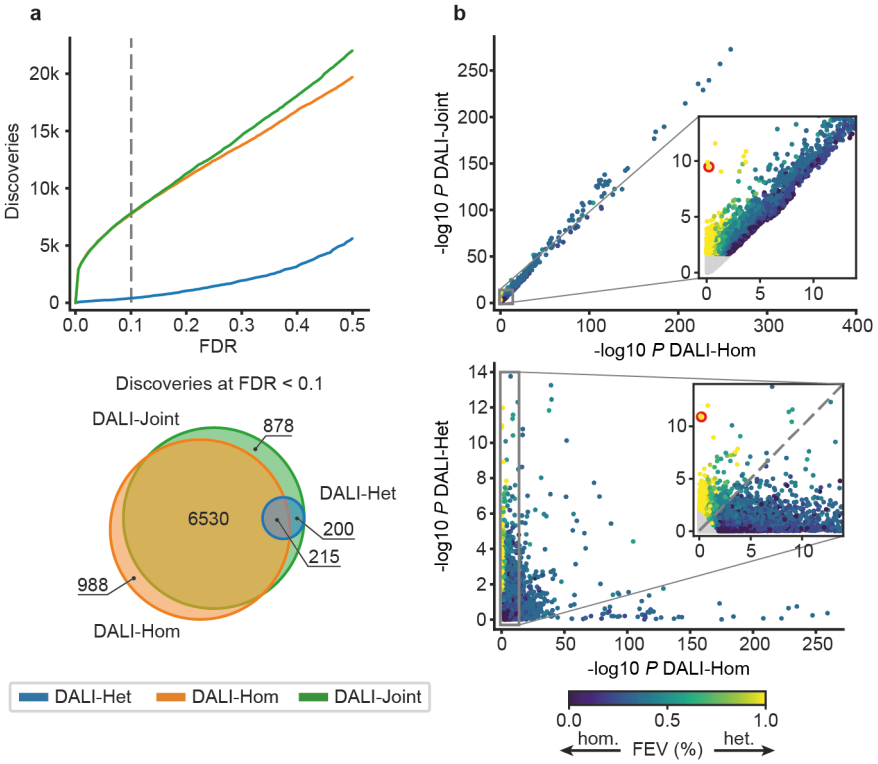

Figure 3.13. scDALI discoveries. **(a)** Number of peaks with allelic imbalance identified by scDALI-Joint, scDALI-Hom and scDALI-Het across all F1 crosses and timepoints. Top: Number of discoveries as a function of FDR (controlled using the Benjamini-Hochberg procedure). Bottom: Overlaps between discoveries by all three scDALI tests (FDR < 0.1). **(b)** Scatter plots of negative log p-values produced by the scDALI tests, comparing scDALI-Joint vs. scDALI-Hom (top) and scDALI-Het vs. scDALI-Hom (bottom). Points are colored by the estimated extent of allelic imbalance driven by heterogeneous effects; non-significant peaks are colored grey (adjusted scDALI-Joint p-value > 0.1). Highlighted in red is the peak chr3R:20310056- 20311056, a region with significant heterogeneous cell-state-specific effects but no discernible global imbalance.

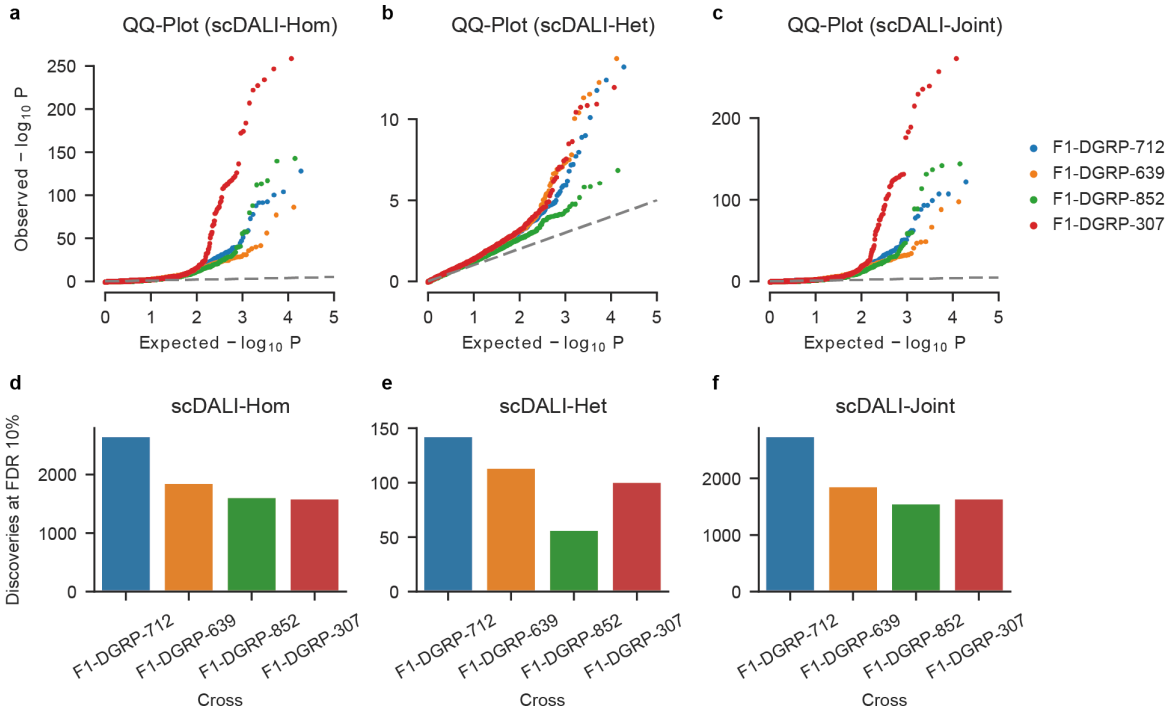

Figure 3.14. scDALI discoveries by cross. **(a-c)** Q-Q plots for p-values from scDALI-Hom, scDALI-Het and scDALI-Joint. **(d-f)** Number of discoveries for each test (FDR < 0.1) stratified by cross.

rates under the scDALI model. Models were trained using sparse variational inference implemented in GPflow [201], using a maximum of 1,000 inducing points (depending on the total number of cells with non-zero read counts for a given peak). Cell-state specific effects were then annotated using two different strategies. I stratified the estimated allelic rates (posterior mean) in 7 annotated developmental lineages (see **Fig.** 3.6f), to identify lineages with distinctive allele-specific effects. Furthermore, by ordering cells by their estimated posterior mean allelic rate and computing the difference between the top and bottom 10% quantiles (Qdiff10), I define a measure of the effect size of heterogeneous allele-specific imbalances, which captures the variation in allelic rates between the most extreme populations (**Fig.** 3.16). As an alternative approach, Stefano Secchia generated transcription factor (TF) activity scores for each cell using chromVAR [216] (v.1.10.0), based on a curated set of 65 TFs with known DNA binding motives from [183]), to identify TFs whose activity strongly correlated with the estimated allelic rates. In principle, this approach allows for pinpointing particular regulatory processes linked to allelic imbalance without needing to define distinct

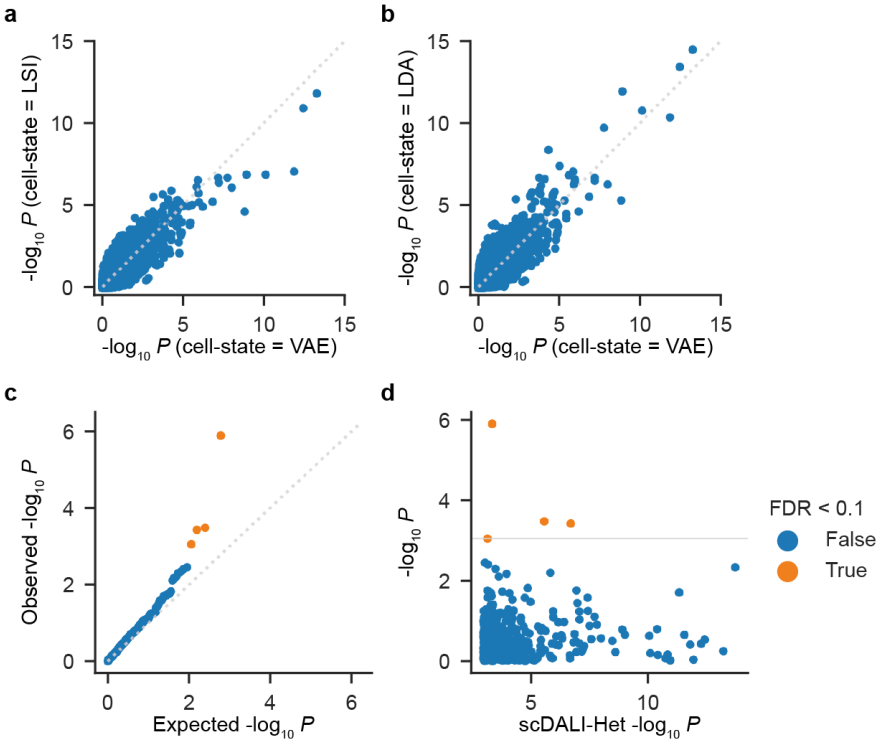

Figure 3.15. scDALI-Het diagnostics. **(a, b)** Comparison of negative log p-values using alternative cell-state representations for scDALI-Het (data from cross, F1-DGRP-712 with 10,220 cells and 12,861 peaks). Covariance matrices were constructed from the VAE embedding (default), **(a)** latent semantic indexing (LSI) and **(b)** cisTopic. **(c, d)** Testing for associations between allele-specific quantifications and total counts in 415 peaks of accessibility with evidence for heterogeneous allelic imbalance (identified using scDALI-Het). **(c)** Q-Q plot, comparing p-values produced by the association test to a uniform distribution. **(d)** Scatter plot of p-values from scDALI-Het (using the VAE cell-state representation) and the association test as in **(c)**. Only 4 out of 415 peaks with heterogeneous allelic imbalance also show evidence for an association between total and allele-specific read counts (FDR < 0.1).

cell groupings *a priori*.

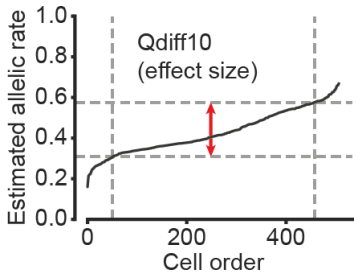

Figure 3.16. Qdiff10 measures the difference between the top and bottom 10% quantiles of the distribution of allelic rates estimated by scDALI.

In several cases allelic imbalance affected known lineage-specific regulatory elements. For example, region chr3R:22877489-22878489 (scDALI-Het $P = 2.7 \times 10^{-5}$) has been previously identified as a neuronal-specific DNase Hypersensitive Site (DHS) [217] and has been demonstrated to function as a nervous system enhancer in vivo (CAD4 database [68]). Accordingly, this region is identified as predominantly accessible in the nervous system (**Fig.** 3.17a). In addition, while cells from other lineages show no appreciable allelic imbalance, accessibility in the nervous system is strongly biased for the paternal allele (Qdiff10 of 0.24 = 0.24, **Fig.** 3.17b, c). In accordance with the allelic imbalance identified by scDALI at this locus, the assessment of TFs associated with heterogeneity in allelic effects identified known nervous system regulators, such as Tramtrack (ttk) and Hairy (h) (**Fig.** 3.17d, e).

Interestingly, I found a number of regulatory regions that show opposing allelic imbalances in different lineages. For example, region chr2R:13675707-13676707 has only a small maternal bias (estimated overall mean rate 0.61) when considering the global allelic rate but is identified as a site with pronounced allelic heterogeneity by scDALI (scDALI-Het $P = 1.5 \times 10^{-8}$, **Fig.** 3.18a). This region has previously been identified as a neuronal and muscle-specific DHS [217] and accordingly shows increased accessibility in the nervous system and muscle in the data. However, accessibility is biased for the maternal allele in the muscle and the paternal allele in the nervous system (Qdiff10 = 0.29, **Fig.** 3.18b, c). This pattern of opposing allelic imbalance is also reflected in the correlation with the activity of TFs active in these tissues. For example, known muscle regulators, such as Twist (twi) and Tinman (tin) are correlated with the maternal allelic rate, while factors active in the nervous system, for example, Tramtrack (ttk), Disconnected (disco), and Kruppel (Kr), are correlated with the paternal rate (**Fig.** 3.18d, e).

Another example is chr3R:20310056-20311056 (scDALI-Het $P = 5.45 \times 10^{-8}$), a region spanning an intron of the gene CG42668. The total accessibility of this region largely coin-

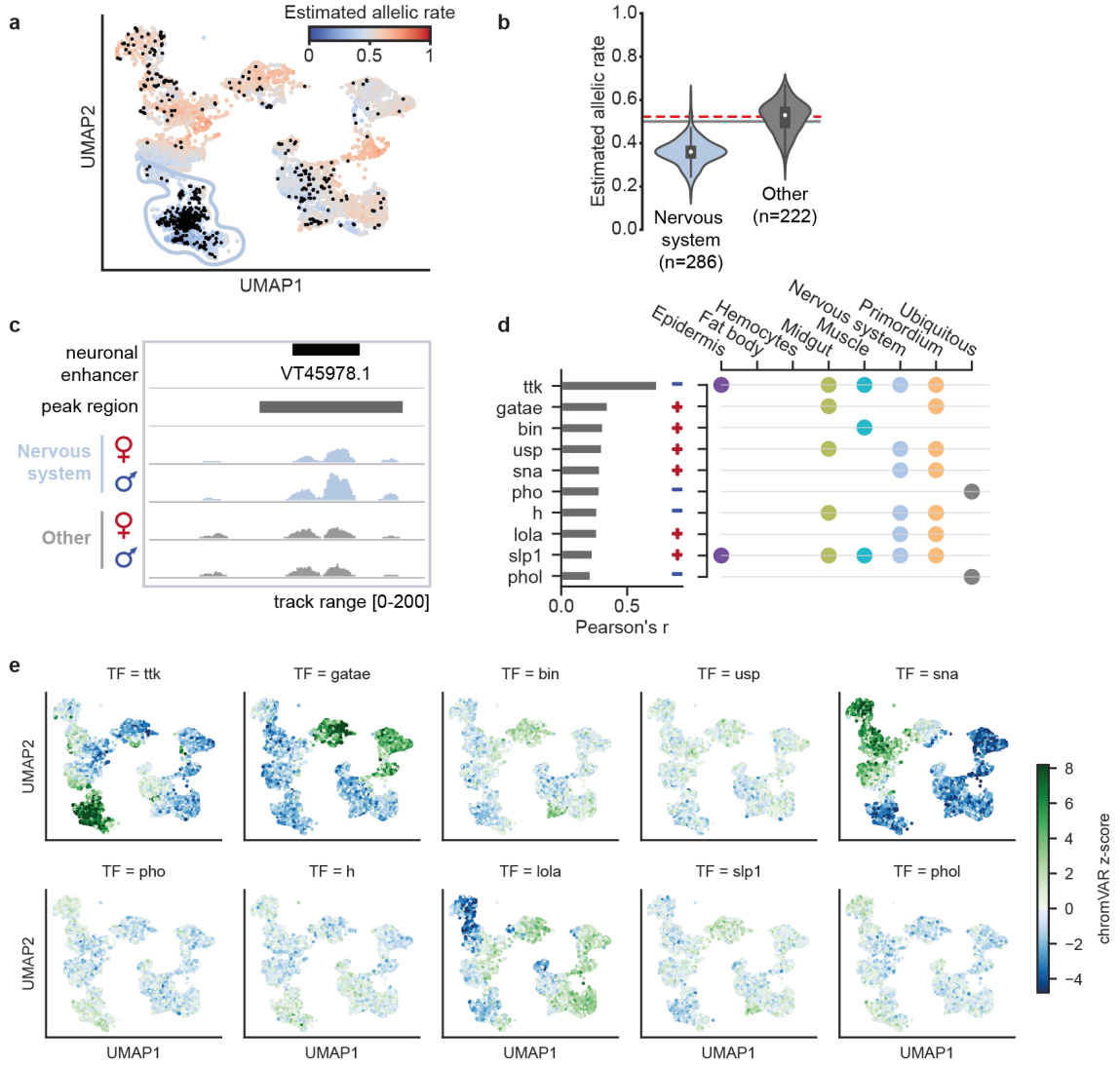

Figure 3.17. Exemplary analysis of region chr3R:22877489-22878489 in cross F1-DGRP-639, an ATAC peak of accessibility with heterogeneous allelic imbalance. **(a)** UMAP visualization of the estimated allelic rate (maternal accessibility relative to total accessibility). Cells with observed allele-specific counts (training data) are highlighted in black. **(b)** Violin plots showing the distribution of estimated allelic rates in selected lineages. Solid grey lines indicate allelic balance (rate = 0.5), while the dotted red line shows the estimated mean allelic rate across all cells. **(c)** Genome browser tracks for region chr3R:22877489-22878489 illustrating allele-specific aggregate accessibility for the nervous system and other populations. **(d)** Left: Correlation between estimated allelic rates and chromVAR transcription factor (TF) activity scores in individual cells. Shown are the ten strongest associations, with plus and minus signs indicating the direction of correlation. Right: Curated lineage annotation for each TF. **(e)** chromVAR transcription factor activity scores (z-scores) based on the total accessibility of associated motives for TFs shown in **(d)**.

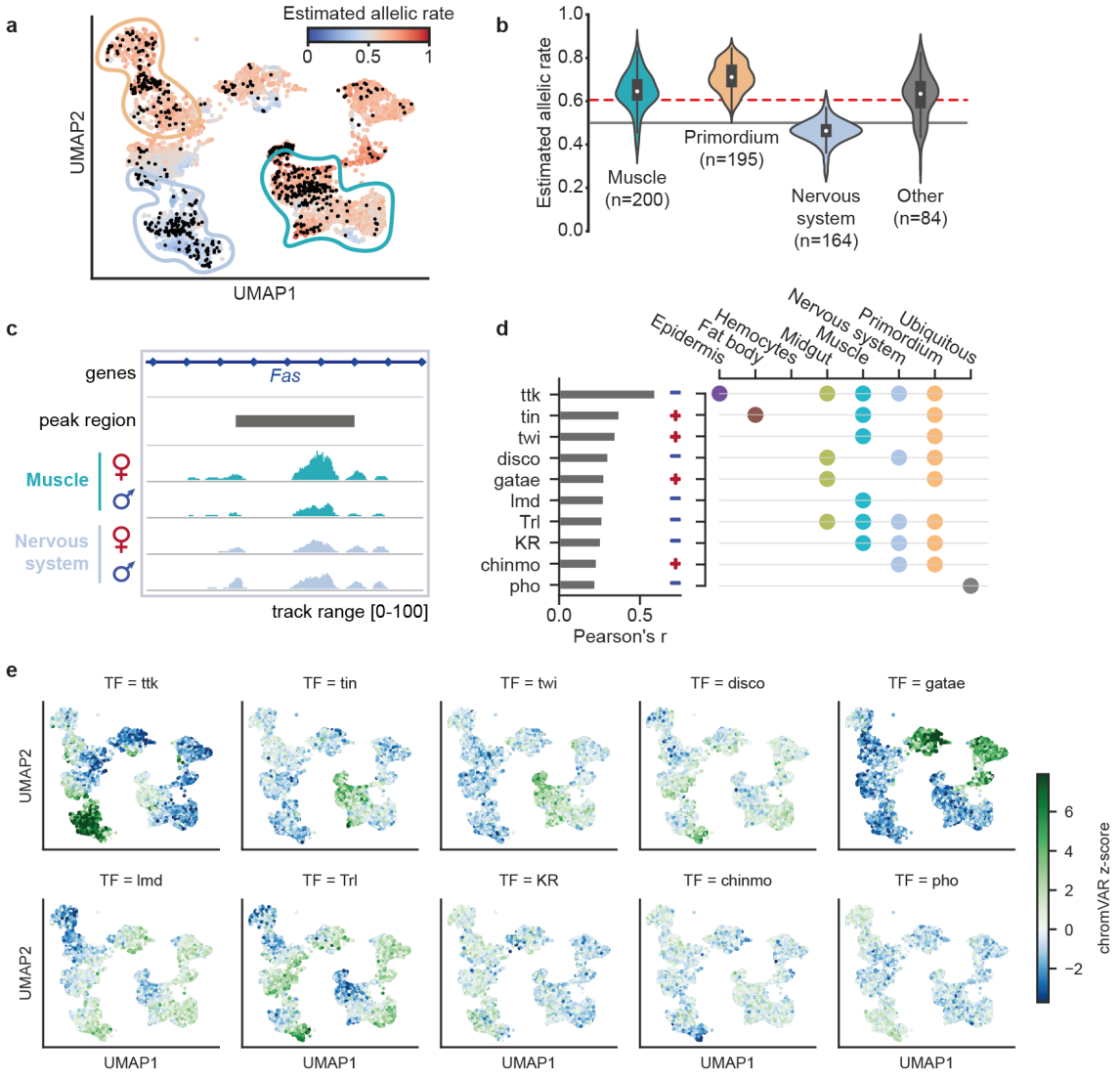

Figure 3.18. scDALI analysis for peak chr2R:13675707-13676707 in cross F1-DGRP-639 , revealing opposing effects in the nervous system and muscle lineage. Panels as in **Fig.** 3.17.

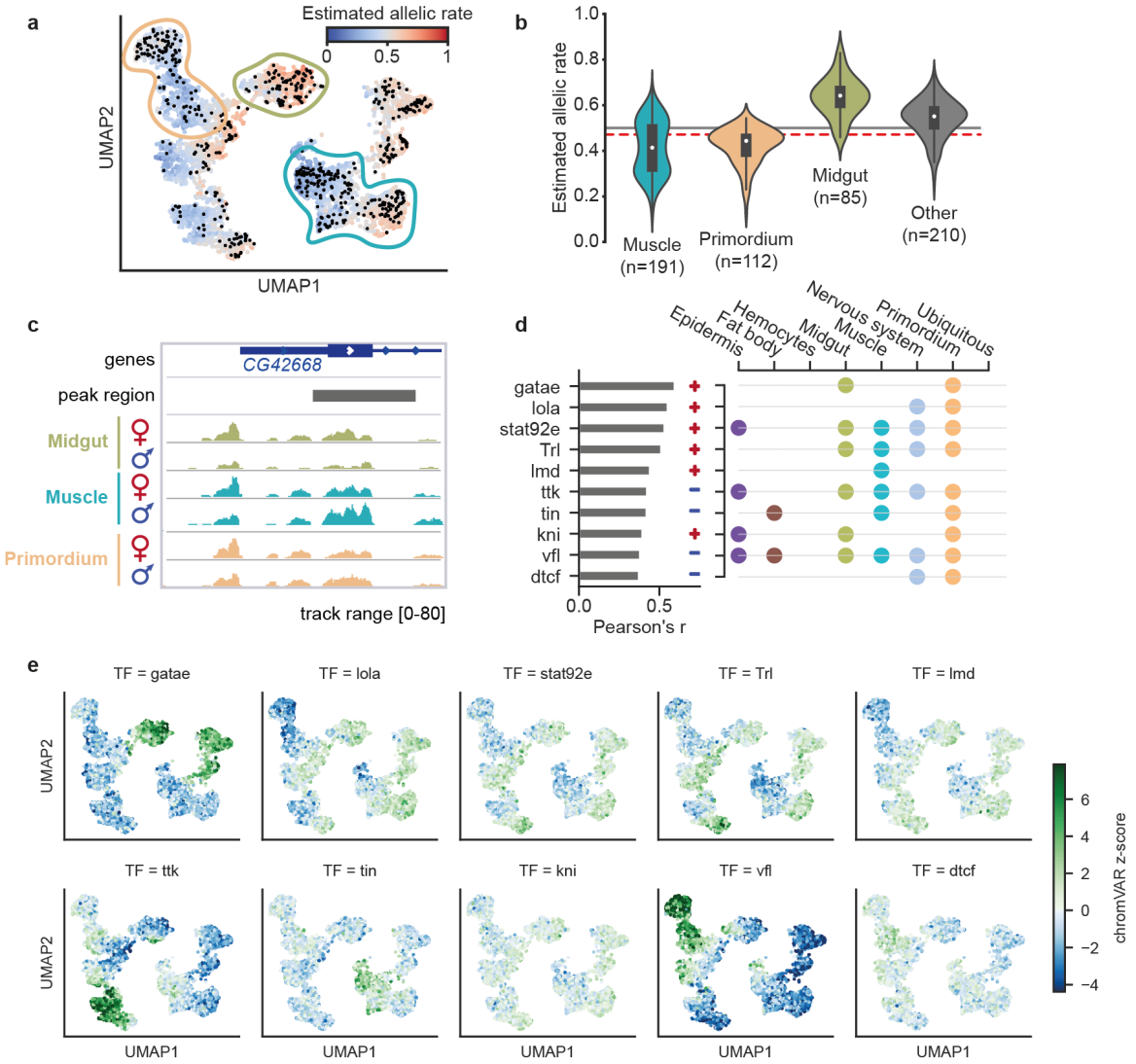

Figure 3.19. scDALI analysis for peak chr3R:20310056-20311056 in F1-DGRP-307, revealing lineage-specific differences in allelic rates for the muscle, primordium, and midgut, as well as intra-lineage variation within the muscle population. Panels as in **Fig.** 3.17.

cides with the known tissue-specific gene expression of CG42668 in the cells of the midgut and visceral muscle. The allele-specific analysis revealed differential allele-specific effects in both tissues, suggesting distinct regulatory programs orchestrating the tissue-specific activity of CG42668 (**Fig.** 3.19a). Furthermore, muscle cells showed additional intra-lineage variation, resulting in a bi-modal distribution of allelic rates (**Fig.** 3.19b,c). Despite the presence of strong inter- and intra-lineage variation (quantile difference 0.39), this effect is obscured in a bulk-level analysis (scDALI-Hom P = 0.81). The activity score of GATAe, a known midgut TF, was highly correlated (Pearson r > 0.5) with the maternal rate, while Zelda (vfl), which has a role in zygotic genome activation and early developmental patterning in the embryo primordium, with the paternal rate, consistent with the allelic bias observed in these cell populations (**Fig.** 3.19d,e). The temporal intra-lineage variation within the muscle population was also reflected in the correlation with the activity of known early and late muscle TFs. Twist (twi) and Tinman (tin) are active in the early muscle primordium (mesoderm) where they direct the specification of the muscle lineages, and concordantly their activity scores were correlated with the paternal allelic rate observed in the early muscle cells. TF Lameduck (lmd) was instead correlated with the maternal rate, as it is required during later stages of muscle formation for the proper specification of the somatic and visceral muscle (**Fig.** 3.19d,e).

More globally, allele-specific effects were stronger at distal regulatory elements (potential enhancers) compared to promoter-proximal regions, both for peaks with heterogeneous (onesided Mann-Whitney U test, $P = 5.6 \times 10^{-5}$) as well as homogeneous (onesided Mann-Whitney U test, $P = 2.17 \times 10^{-26}$) imbalance (**Fig.** 3.20a). Furthermore, imbalances were significantly more common at distal versus proximal regions (**Fig.** 3.20b), similar to what has been observed in bulk ATAC-seq data at time-matched developmental stages [191]. These differences between distal and proximal sites were less pronounced when considering discoveries from scDALI-Hom (two-sided Binomial test $P = 0.02$), with about 61% of significant regions being found at proximal regions compared to 62% of all tested peaks. Interestingly, however, I found this effect to be markedly more prominent for heterogeneously imbalanced regions (two-sided Binomial test $P = 2.15 \times 10^{-10}$), with only 47% of peaks discovered by scDALI-Het being located near gene promoters.

To further characterize heterogeneous imbalances, I used scDALI to assess differential lineage effects, testing for differences in mean allelic rates between each lineage and all remain-

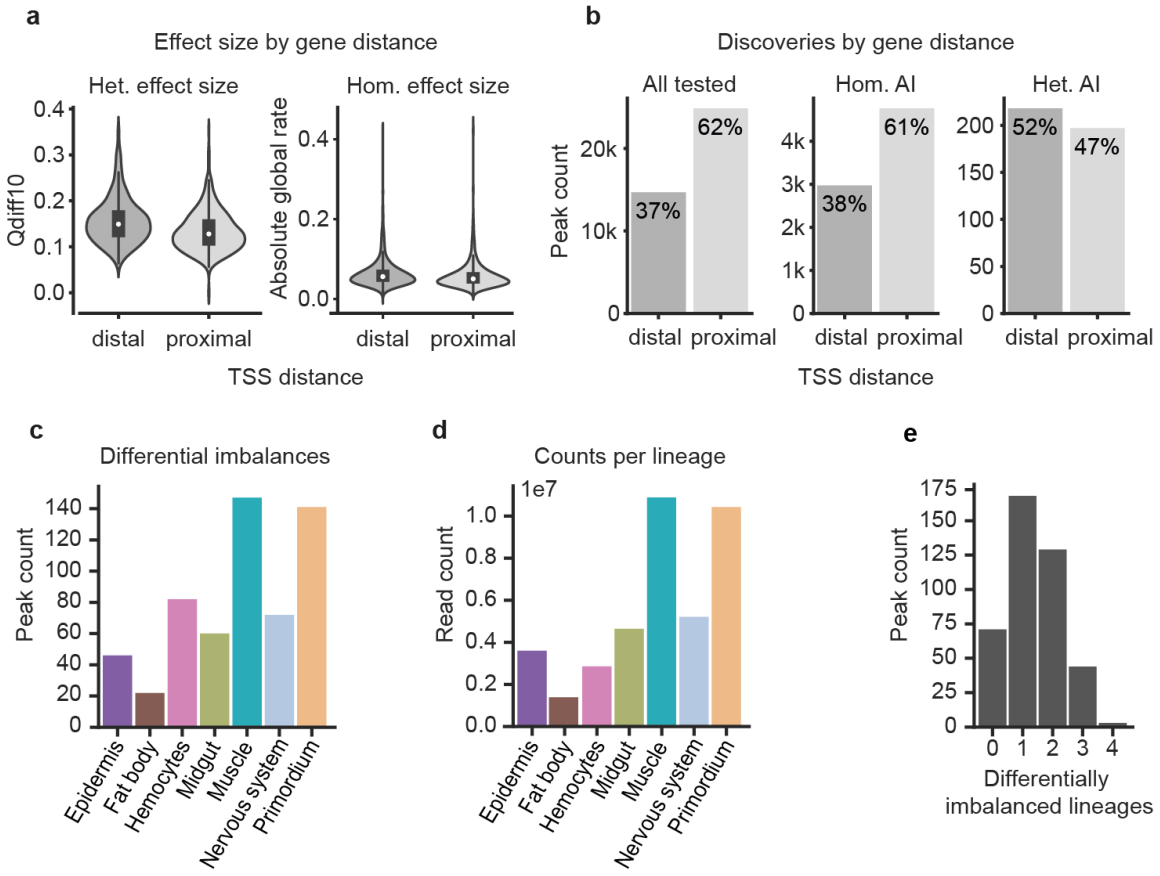

Figure 3.20. Analysis of allelic effects. **(a)** Effect size distribution for heterogeneously (Qdiff10) and homogeneously imbalanced (absolute deviation from 0.5) peaks, considering distal and promoter-proximal regions separately. **(b)** Total number of peaks tested and peaks with allelic imbalance identified using alternative tests (FDR < 0.1), stratified by the peak distance to the transcription start site (TSS) of the closest gene. Heterogeneously imbalanced peaks are markedly more common at distal regions. **(c-e)** By-lineage analysis of allelic imbalance using scDALI-Het for peaks with significant heterogeneous imbalances. **(c)** Distribution of the number of differentially imbalanced lineages per peak. **(d)** Read count distribution across lineages. **(e)** Distribution of the number of peaks with increasing numbers of differentially imbalanced lineages. The majority of peaks show imbalance in a single lineage

ing cells in each of the 415 peaks identified by scDALI-Het. This test was also formulated under the scDALI-Het framework, replacing the continuous cell state kernel with a blockdiagonal matrix to indicate lineage membership. Unsurprisingly, the frequency of significant imbalances by lineage (FDR < 0.1) largely resembled the overall read count distribution, which influences the detection power for allelic imbalance (**Fig.** 3.20c, d). For the majority of peaks, allele-specific variation was attributable to one or two differentially imbalanced lineages (72%); however, 11% of peaks showed differences between three or four lineages (**Fig.** 3.20e). Interestingly, for 17% of scDALI-Het discoveries, allele-specific effects do not differentiate any single lineage, indicating the presence of significant intra-lineage variation, for example due to developmental time.

Allelic imbalance across developmental time

Developmental time is a major driver of variation in the data and therefore a promising predictor of allele-specific changes within lineages. I applied scDALI to test for time-specific allelic imbalances within muscle, the lineage with the largest number of cells, using the pseudo-temporal ordering estimated by the VAE model as a cell state representation (**Fig.** 3.21). Leveraging the scDALI framework, I designed a covariance matrix capturing both linear and nonlinear (polynomial) temporal dependencies. Specifically, I set $\mathbf{K}_{Time} = \mathbf{C}_{Time}\mathbf{C}_{Time}^{T}$,

$$\mathbf{C}_{Time} = \begin{bmatrix} \vdots & \vdots & \vdots \\ f_y(\mathbf{z}_n) & f_y(\mathbf{z}_n)^2 & f_y(\mathbf{z}_n)^3 \\ \vdots & \vdots & \vdots \end{bmatrix} \tag{3.44}$$

where $f_y(\mathbf{z}_n)$ is the continuous developmental ordering inferred by the VAE model (see eq. 3.36). Out of 363 peaks with significant heterogeneous allelic imbalance that are accessible in muscle (mean total allelic count within lineage < 0.1), scDALI identified 69 (19%) peaks with significant time-specific effects (FDR < 0.1; **Fig.** 3.21a, b). Notably, 27% of these peaks with time-specific allelic imbalance did not show any lineage-specific effects (**Fig.** 3.21c). As an example, region chr2R:13675707-13676707 discussed above (**Fig.** 3.18) does indeed exhibit strong time-specific imbalances (**Fig.** 3.21d-f), consistent with the observed intra-lineage variation specifically in muscle cells.

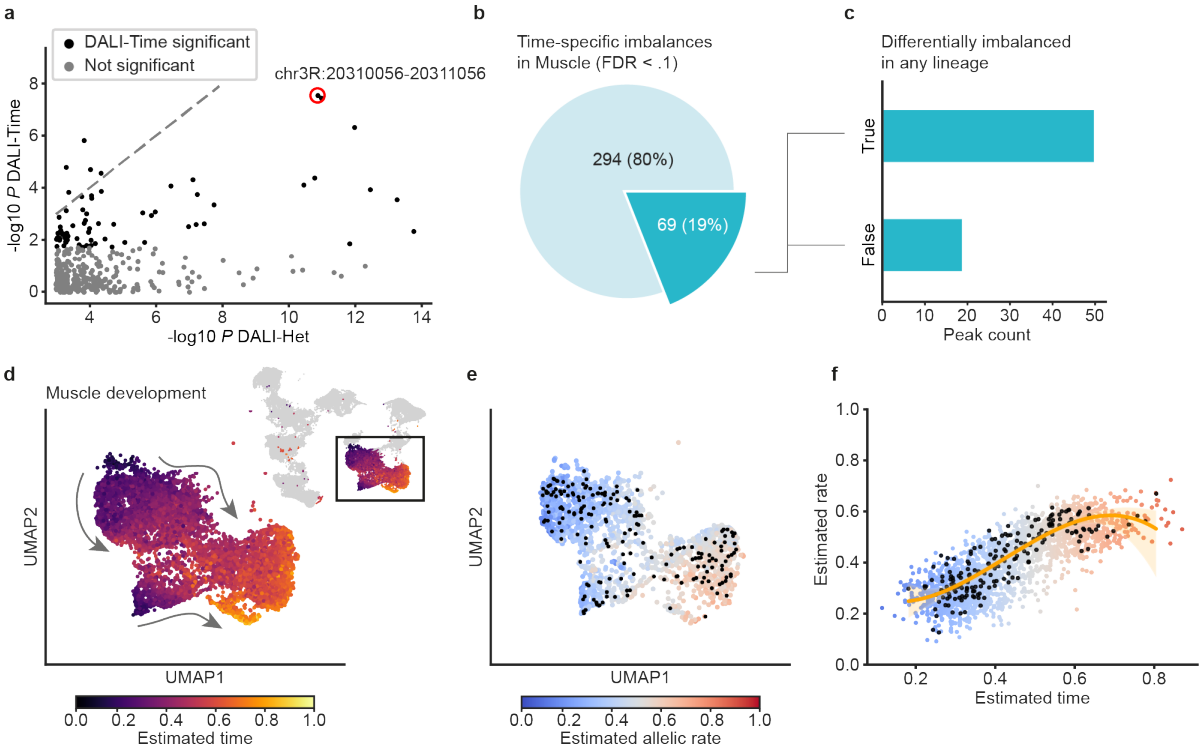

Figure 3.21. scDALI-Het identifies time-specific intra-lineage variation. **(a)** Scatter plot of negative log p-values for scDALI-Het versus a scDALI test for time-specific variation in the muscle population. Red circle highlights region chr2R:13675707-13676707. **(b)** Of 363 peaks identified by scDALI-Het that are accessible in the muscle, 19% showed significant temporal effects (FDR < 0.1). **(c)** The majority of peaks with a time-specific effect in the muscle did not show significant differential allelic imbalance between lineages. **(d)** Temporal order for the muscle lineage estimated by the variational autoencoder model. **(e, f)** Estimated allelic rates across time for region chr2R:13675707-13676707. Black dots denote cells with observed allele-specific counts in this region

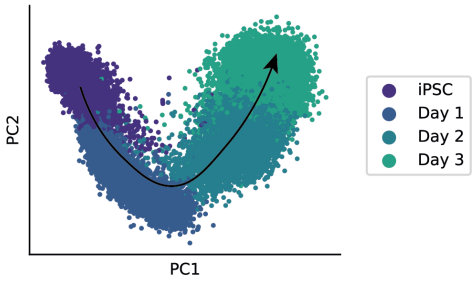

Figure 3.22. scRNA-seq of differentiating iPS cells. Principal component analysis (PCA) of total gene expression counts for 34,254 cells. Shown are cell embeddings using the first two principal components, colored by the day of sample collection.

3.1.7 Application to scRNA-seq of human IPSCs

To demonstrate that scDALI is also applicable to single-cell RNA-seq, I considered a recently published multi-donor single-cell RNA-seq dataset of human induced pluripotent stem cells (iPSCs) differentiating towards definitive endoderm [84]. Samples were profiled using a full-length sequencing protocol (Smart-seq2, [64]), allowing for the quantification of gene expression in haplotype-resolved manner and thus providing the basis for an analysis using scDALI. The study spans single-cell profiles from 125 donors at four time points of iPSC differentiation (day 0: iPSCs, day 1, day2, and day 3 of differentiation towards definite endoderm). Total gene expression counts for 34,254 cells and all genes, as well as allele-specific quantifications for 4,470 previously identified SNP-gene pairs (4,422 eQTL lead variants) were obtained as described in the primary publication [84]. Reads were initially mapped to reference and alternative alleles for each heterozygous SNP in every cell and subsequently assigned relative to the genotype of each chromosome using haplotype assignments estimated in the primary publication. Allele-specific read counts were aggregated at the gene level, by summing up the counts for each SNPs contained in exonic regions. Finally, for each eQTL (gene-SNP pair), gene-level allele-specific counts were interpreted relative to the eQTL variant to obtain a consistent definition of ASE across cells from different donors that were heterozygous for that variant. SNP-gene pairs were filtered by requiring at least 50 cells with nonzero allele-specific counts, leading to 3,966 pairs to be tested using scDALI-Het. I performed principal component analysis (PCA) of total gene expression counts from 34,254 cells and used the leading k principal components (PCs) and a linear covariance function to define cell state covariance matrices. I chose $k = 1$ to focus on time-specific allelic imbalance (**Fig.** 3.22) while $k = 10$ was used to model more general cell-state effects.

While allelic rates are generally less susceptible to confounding variables such as batch ef-

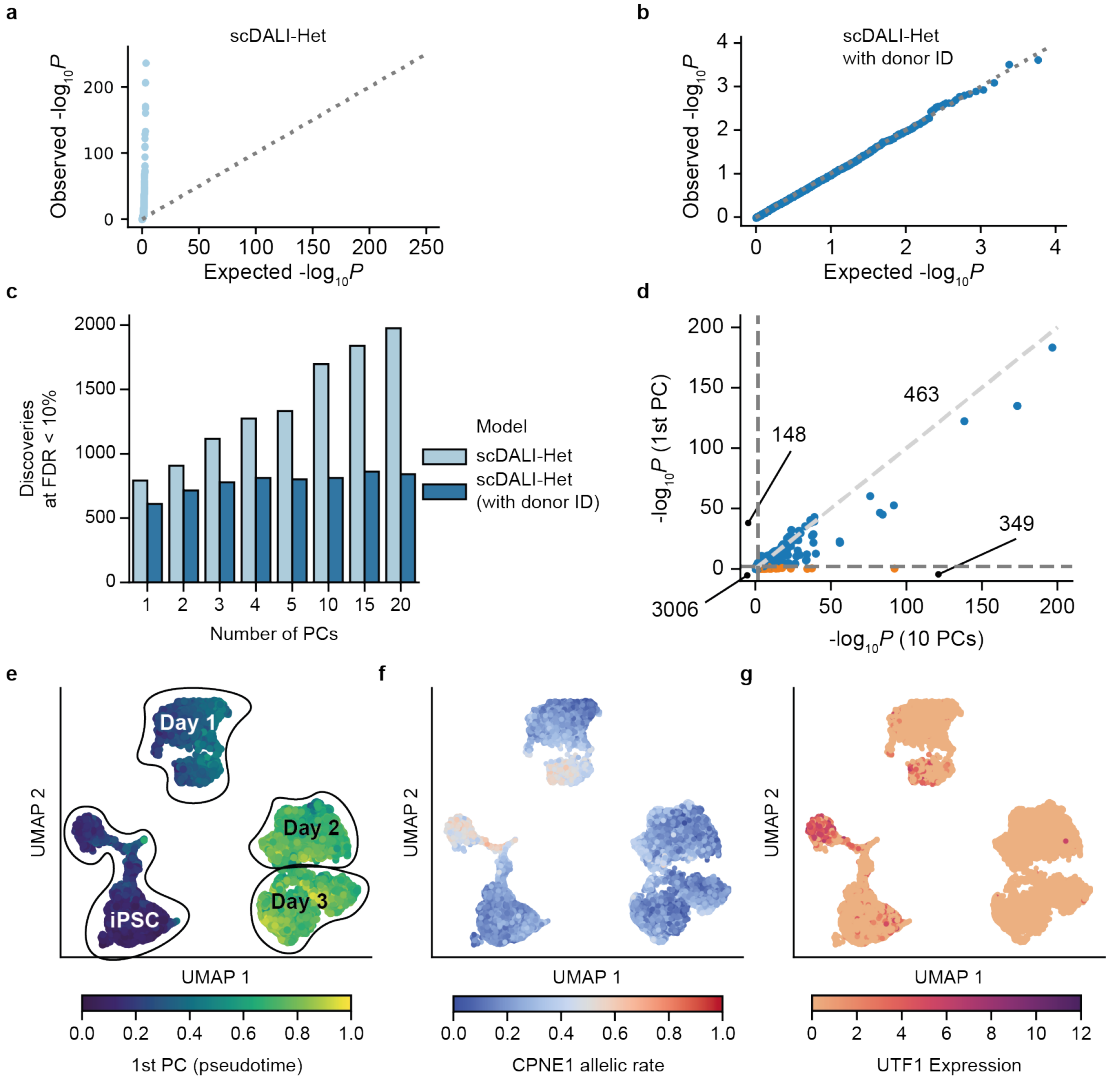

Figure 3.23. scDALI-Het applied to scRNA-seq of differentiating iPSCs reveals cell-state specificity of eQTL. **(a,b)** Q-Q plot of scDALI-Het p-values when permuting the cell-state coordinates of cells from the same donor. A model that does not account for the donor identity yields inflated p-values **(a)**, whereas scDALI-Het with donor identities as fixed effects yields calibrated results **(b)**. **(c)** Number of discoveries for varying numbers of principal components (PCs) used to define the cell state kernel. If donor identity is not accounted for, using a larger number of PCs for the cell-state definition leads to an increasing number of discoveries. **(d)** Scatter plot of negative log p-values, comparing a model using only the leading PC versus scDALI-Het with 10 PCs. Orange dots are discoveries that are exclusively identified by the general cell-state test (10 PCs). Indicated are the number of significant discoveries in each quadrant (10% FDR). **(e)** UMAP visualization of collectionand pseudotime. **(f)** Estimated allelic rates for the eGene-QTL pair (CPNE1, chr20:34344225 T/A). **(g)** Expression of UTF1, a previously identified marker for neuronal differentiation success.

fects, donor-specific read mapping biases as well as differences in the representation of cell types and cell states could lead to spurious signals of heterogeneous allelic variation. To assess the effect of donor-specific effects on test calibration, I permuted the leading 10 PC coordinates among cells from the same donor before constructing the cell-state covariance matrix. I then compared two implementations of scDALI-Het that either did or did not account for the donor background using a one-hot-encoded representation of the donor identities (donor IDs) for each cell as additional fixed-effect covariates. This analysis confirmed the need to account for donor identities in order to retain calibrated test statistics (10 PCs, **Fig.** 3.23a, b). I then assessed the number of eQTL with heterogeneous imbalance discovered by scDALI-Het when varying the number of principal components used to construct the cell-state kernel, finding that more complex kernels yielded a larger number of discoveries, which however saturated for five or more components (**Fig.** 3.23c). For example, a model using the first PC to define a cell state kernel (which primarily captured differentiation, **Fig.** 3.22) identified 611 eQTL with heterogeneous allelic imbalance compared to 812 eQTL when using 10 components (**Fig.** 3.23d, FDR < 0.1). This indicates that although variation in gene expression in this data is predominantly explained by the differentiation state (**Fig.** 3.23e), the remaining sources of variation drive a substantial fraction of distinct genetic regulation. One example of such an effect is an eQTL with heterogeneous ASE for CPNE1 ($P = 3 \times 10^{-9}$, scDALI-Het). CPNE1 has been shown to play a role in neuronal progenitor cell differentiation [218]. Intriguingly, the pattern of allelic imbalance is confined to a distinct subpopulation of iPS cells, which is marked by expression of UTF1. Notably, this UTF1-positive iPS subpopulation has recently been associated with differentiation efficiency towards a midbrain neural fate [85] (**Fig.** 3.23f).

3.1.8 Discussion

The majority of disease associated variants impact non-coding regions, disrupting the function of regulatory elements such as enhancers and promoters. As enhancers regulate when and where genes are expressed, genetic variation within enhancers naturally has cell type-specific effects. However, capturing and understanding these genetic effects is an enormous challenge. Resolving these effects to specific cell types using classical quantitative trait loci (QTL) mapping would require FACS sorting different cell types from a heterogeneous tissue across a large panel of individuals, a huge task that is often impossible as specific markers for cell isolation are not available for many cell types and transitions. To address this, I

developed scDALI, a computational framework to characterize the cell-type specificity of genetic effects from single-cell sequencing data in an unbiased fashion. The model provides a principled strategy for exploiting two independent signals that can be obtained from the same sequencing experiment, whether that is gene expression or epigenetic data: (1) total counts, which I use to derive cell types and states, and (2) allele-specific quantifications of genetic effects within genomic features such as genes or ATAC peaks of accessibility. Combining these two measurements enabled scDALI to discover for both pervasive, homogeneous imbalance and cell-state-specific heterogeneous effects, without the need to define cell types or cell states a priori.

I applied scDALI to newly generated scATAC-seq profiles from an F1 cross design, assaying dynamic and discrete changes in allele-specific chromatin accessibility of developing Drosophila melanogaster embryos, a naturally very heterogeneous sample. I designed a novel variant of the variational autoencoder framework, to infer cell states from sparse, high-dimensional open chromatin measurements while integrating coarse-grained information on sampling times. The learned latent representation successfully separated known cell types and developmental lineages, providing a comprehensive description of cell states in the data. scDALI discovered thousands of regions with allelic imbalance, hundreds of which show distinct cell state-specific effects. About half of the regulatory regions with allele-specific effects in specific cell types were not detectable in a pseudo-bulk analysis, as opposing effects canceled out across the cell state space. Although the total number of discoveries with heterogeneous imbalances was relatively modest, increasing sample sizes will likely improve power to detect these effects. For example, the more recent sci-ATAC-seq3 method now allows for generating datasets with millions of cells [219], two orders of magnitude larger than the data considered here. Nevertheless, our analysis identified genetic effects at a number of characterized tissue-specific developmental enhancers. scDALI estimates allele-specific effects in individual cells, which allows dissecting this heterogeneity at different resolutions. I showed how this map can be used to identify the underlying regulatory programs by associating differential allelic imbalance with pathway or transcription factor activity scores. Alternatively, it is possible to aggregate allelic rates at the level of known (discrete) clusters, thereby assessing the distribution of estimated allelic activity both between and within lineages or cell types. I found that developmental time is an important contributor to intra-lineage variation of allelic imbalance, pinpointing developmental stage-specific enhancers. Furthermore, the analysis revealed that allele-specific effects are

significantly stronger and more common at distal elements (putative enhancers) compared to promoter-proximal regions. Notably, these differences are markedly more pronounced among peaks with heterogeneous (tissue-specific) imbalances compared to homogeneous effects, confirming and extending previous results on bulk-sequencing data [191]. I then applied scDALI to a published scRNA-seq dataset from 125 human iPS cell lines and demonstrated how the model can be used to discover context-specific genetic effects of known eQTL and characterize the associated cellular subpopulations. While the approach uncovers many novel putative enhancers, it also has its limitations. The focus of this work lies on the characterization of cell-state-specific effects for known quantitative trait loci and the mapping of genetic effects from few available individuals or even a single sample. In particular, I do not test for interactions between cell states and the presence of genetic variants, which prevents the model from discovering potential causal loci associated with cell-state-specific allelic imbalance. While in principle, it is possible to combine allelic analyses with genotype data to identify causal variants [40, 187, 188, 192], this requires larger numbers of unique genotypes. The required multi-individual single-cell sequencing studies are only beginning to emerge and scDALI could be extended to leverage such variation. Understanding to what degree allele-specific effects replicate at different molecular layers remains another important direction of future research. In this study, we have demonstrated that scDALI can be flexibly applied to both single-cell RNA-seq and ATAC-seq data. However, new multi-omics methods can obtain both DNA accessibility and RNA measurements from the same single cell [220]. The integration of these different dimensions of allelic imbalance across both modalities will be an important area for future work that may help to relate the functional impact of genetic variation in enhancers to their target gene's expression.

3.2 CellRegMap: mapping context-specific eQTL

Seminal studies have shown that it is possible to identify expression quantitative trait loci from single-cell RNA-sequencing data [83, 84, 97]. These studies not only recovered known eQTL previously discovered using bulk sequencing methods, but also demonstrated that accounting for the molecular context captured using scRNA-seq provides increased resolution to map genetic effects [83, 85, 221]. Most existing workflows extract eQTL separately from each of multiple discrete cell populations, by modeling genetic effects on aggregate expression profiles (chapter 1). In principle, this approach could be improved using established

methods for multi-tissue eQTL analysis (e.g., [58, 59, 222–228]) to jointly model data from multiple populations. However, both approaches remain limited in their ability to account for subtle cell states and continuous transitions, that can not be accurately captured by discretization of single-cell profiles. Alternatively, more flexible interaction models exist and have been applied in the context of bulk-eQTL mapping (e.g., [83, 184, 229]), but do not effectively account for multi-level covariance structure at the level of individuals (genetic relatedness) and cells (repeated sampling from the same individual). Consequently, these approaches do not fully leverage the resolution provided by single-cell data, potentially failing to detect changes in allelic regulation across more subtle cell subtypes.

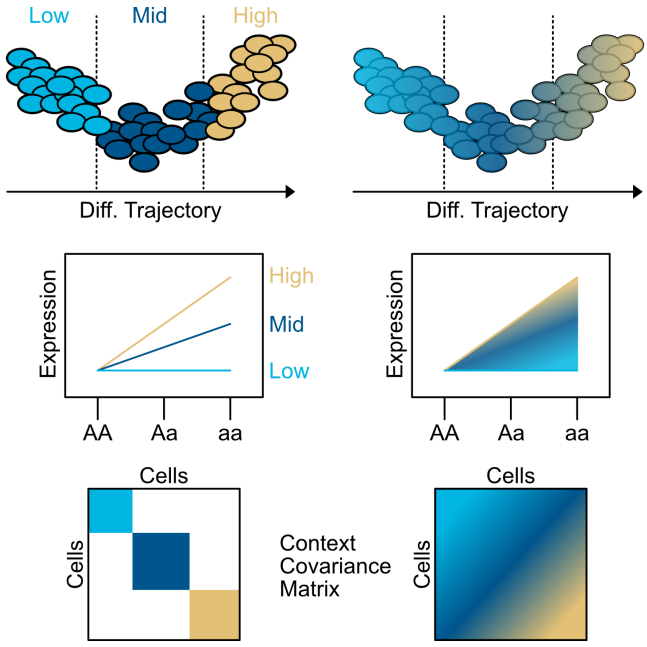

Figure 3.24. CellRegMap is a statistical model to identify genetic effects on gene expression in single cells. Instead of discretizing the cell-state space, CellRegMap uses a cell covariance matrix to capture continuous interactions between molecular contexts and genetic regulation. Adapted from [186]; original figure by Anna Cuomo.

This section introduces the Cellular Regulatory Map (CellRegMap), a framework for mapping regulatory variants in an unbiased manner across cell types and cell states as obtained from scRNA-seq profiles. Cellregmap uses the formalism developed for scDALI to avoid any discretization of cells into cell types and instead uses a multi-dimensional cell state manifold estimated from single-cell transcriptome profiles to define cellular contexts in a continuous and unbiased manner. CellRegMap then allows to test for and characterize in-

teraction effects between individual genetic variants and cellular context on gene expression traits (**Fig.** 3.24). The primary use case of CellRegMap is to reanalyze eQTL variants with known additive effects, however the model can in principle also be used for variant discovery. To validate CellRegMap, I develop a semi-synthetic simulation framework that leverages real cell states, expression profiles and genotypes. The synthetic data is used to assess statistical calibration, power and computational requirements. Lastly, I discuss an illustrative example application of CellRegMap to the scRNA-seq data set of differentiating human iPSCs already introduced in the previous section.

Acknowledgements and contributions

This work was supervised by Oliver Stegle and John C. Marioni. The model was developed and implemented by Anna Cuomo, with contributions from Danilo Horta. Anna Cuomo performed the analysis of eQTL variants from the iPSC data. I evaluated alternative methods for defining cellular contexts from the iPSC data, contributed to the software implementation and formal analysis and developed the validation procedure. The figures in this section are my own work, unless noted otherwise.

3.2.1 Background: StructLMM model for genetic interactions

The CellRegMap can be viewed as an extension of StructLMM [184], a linear mixed model developed for identifying genotype-environment interactions on physiological traits from population data. I briefly review the basic StructLMM model, before introducing the full CellRegMap model in the following section. Let $\mathbf{y} \in \mathbb{R}^N$ be a vector of expression levels of a particular gene of interest, measured in N samples (typically using bulk RNA sequencing). Furthermore, let $\mathbf{X}$ be a matrix of covariates and $\mathbf{g}$ be the N-dimensional genotype vector under the allelic dosage model discussed in section 2.1.9. The StructLMM model assesses if a genetic locus interacts with any of K different environmental variables $\mathbf{E} \in \mathbb{R}^{N \times K}$, such as life style factors (dietary factors, physical activity, alcohol intake, etc.), to shape gene expression,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{g}\beta_G + \mathbf{g} \odot \boldsymbol{\beta}_{GxE} + \mathbf{u} + \boldsymbol{\epsilon}, \tag{3.45}$$

where $\odot$ denotes the Hadamard (element-wise) product and

$$\boldsymbol{\beta}_{GxE} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{GxE}\mathbf{K}) \tag{3.46}$$

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{E}\mathbf{K}) \tag{3.47}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\boldsymbol{\mathcal{I}}_N), \tag{3.48}$$

for a given environmental covariance matrix $\mathbf{K} = \mathbf{K}(\mathbf{E})$ (typically $\mathbf{K} = \mathbf{E}\mathbf{E}^T$). While persistent genetic effects are modeled as fixed effects with coefficient β_G, StructLMM allows for sample-specific effect sizes $\boldsymbol{\beta}_{GxE} \in \mathbb{R}^N$ for the genotype-environment (GxE) interaction term. Here, the elements of $\boldsymbol{\beta}_{GxE}$ are not estimated explicitly, but are marginalized under a multivariate normal prior distribution defined by the environmental covariance matrix.

Notably, the model does not incorporate additional random effect terms to account for other sources of sample covariance, such as genetic relatedness or repeated measurements. Instead, additional variables (e.g. PCs of a genetic kinship matrix) have to be included as fixed effects, which often fails to effectively control for more subtle relatedness or repeat structure (see also section 2.1.9).

3.2.2 The CellRegMap model

To adapt the StructLMM model for the purpose of single-cell eQTL testing, CellRegMap follows the scDALI concept and replaces environmental variables with a low-dimensional embedding $\mathbf{C} \in \mathbb{R}^{N\times K}$ of the observed $N \times M$ gene expression matrix for M genes to represent the molecular context. Analogous to StructLMM, the embedding $\mathbf{C}$ is then used to construct a cell-state covariance matrix $\mathbf{K} = \mathbf{K}(\mathbf{C})$, in order to detect genotype-context (GxC) interactions. However, population-scale single-cell experiments sample multiple cells from the same individual, introducing additional structure in the data. The CellRegMap model accounts for this structure using a modified variance component and can be cast as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{g}\beta_G + \mathbf{g} \odot \boldsymbol{\beta}_{GxC} + \mathbf{u} + \boldsymbol{\epsilon}, \tag{3.49}$$

where now

$$\boldsymbol{\beta}_{GxE} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{GxC}\mathbf{K}) \tag{3.50}$$

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{C}\mathbf{K} + \sigma^2_{RxC}\mathbf{R} \odot \mathbf{K}) \tag{3.51}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\boldsymbol{\mathcal{I}}_N). \tag{3.52}$$

Here, $\mathbf{R} \in \mathbb{R}^{N \times N}$ denotes a relatedness matrix of individuals (see section 2.1.9) expanded to all cells based on the known assignment of cells to individuals. Note that if $\sigma^2_{RxC} = 0$, one recovers the original StructLMM model. In cases where genetic relatedness is not expected to be a major confounding factor, $\mathbf{R}$ is simply used to encode the cell-individual map, such that $\mathbf{R}_{ij} = 1$ if cells i and j were sampled from the same individual and $\mathbf{R}_{ij} = 0$ otherwise. The additional variance component $\mathbf{R} \odot \mathbf{K}$ accounts for possible interactions between this relatedness matrix and the molecular context. Similar ideas have been considered previously, to model polygenic interactions with environmental variables for heritability estimation [230]. As I will show later using simulations, this term is crucial to ensure calibrated test statistics for genotype-context interaction tests.

3.2.3 Statistical hypothesis testing

Note that the Hadamard product $\mathbf{g} \odot \boldsymbol{\beta}_{GxC}$ in eq. 3.49 can be expressed as the matrix-vector product $\text{diag}(\mathbf{g})\boldsymbol{\beta}_{GxC}$. Therefore, the marginal likelihood of $\mathbf{y}$ under the CellRegMap model defined in eq. 3.49 is given by

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\alpha} + \mathbf{g}\beta_G, \sigma^2_{GxC}\text{diag}(\mathbf{g})\mathbf{K}\text{diag}(\mathbf{g}) + \sigma^2_C\mathbf{K} + \sigma^2_{RxC}\mathbf{R} \odot \mathbf{K} + \tau^2\boldsymbol{\mathcal{I}}_N). \tag{3.53}$$

To test for context-specific genetics effects, one needs to evaluate if the scaling factor σ^2_{GxC} is different from zero,

$$H_0 : \sigma^2_{GxC} = 0 \text{ vs. } H_1 : \sigma^2_{GxC} > 0. \tag{3.54}$$

Following section 2.1.6, a score-based test statistic can be defined as

$$Q = \frac{1}{2}\mathbf{y}^T\hat{\mathbf{P}}_0\text{diag}(\mathbf{g})\mathbf{K}\text{diag}(\mathbf{g})\hat{\mathbf{P}}_0\mathbf{y}, \tag{3.55}$$

where

$$\hat{\mathbf{P}}_0 = \hat{\mathbf{V}}_0^{-1} - \hat{\mathbf{V}}_0^{-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\hat{\mathbf{V}}_0^{-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\hat{\mathbf{V}}_0^{-1}, \tag{3.56}$$

$\tilde{\mathbf{X}} = \begin{bmatrix}\mathbf{X}\ \mathbf{g}\end{bmatrix}$ and $\hat{\mathbf{V}}_0$ is the estimated marginal covariance under the null $\sigma^2_{GxC} = 0$,

$$\hat{\mathbf{V}}_0 = \hat{\sigma}^2_C\mathbf{K} + \hat{\sigma}^2_{RxC}\mathbf{R} \odot \mathbf{K} + \hat{\tau}^2\boldsymbol{\mathcal{I}}_N. \tag{3.57}$$

Under the null, Q follows a mixture of chi-squared distributions with one degree of freedom, $z_i^2 \sim \chi^2(1)$,

$$Q \sim \sum_i \psi_i z_i^2, \tag{3.58}$$

where ψ_i are the non-zero eigenvalues of $\frac{1}{2}\mathbf{V}_0^{-1/2}\text{diag}(\mathbf{g})\mathbf{K}\text{diag}(\mathbf{g})\mathbf{V}_0^{-1/2}$. Assuming a linear cell-state covariance, $\mathbf{K} = \mathbf{C}\mathbf{C}^T$, the ψ_i can equivalently be obtained as the non-zero eigenvalues of $\frac{1}{2}(\mathbf{C}\text{diag}(\mathbf{g}))^T\mathbf{V}_0^{-1}(\mathbf{C}\text{diag}(\mathbf{g}))$ (see eq. 3.25). To evaluate the limiting distribution, Davies exact method [114] is used, switching to the modified moment matching approximation method [113, 115] when this fails to converge.

In order to obtain (restricted) maximum-likelihood estimates of all model parameters under the null, the FaST-LMM algorithm proposed by Lippert et al [139] is used, following a particular reparameterization of the marginal covariance under the null $\sigma^2_{GxC} = 0$,

$$\mathbf{V}_0 = \nu^2\mathbf{M} + \tau^2\boldsymbol{\mathcal{I}}_N, \tag{3.59}$$

where $\mathbf{M}$ admits a low rank factorization to enable an efficient singular value decomposition. A detailed derivation can be found in the published article [186].

3.2.4 The CellRegMap association test

The primary focus of CellRegMap is to reanalyze previously identified eQTL variants (e.g., from bulk-sequencing studies), in order to detect and characterize context-specific effects. However, as part of the CellRegMap software suite, it is also possible to test for persistent genetic effects, while appropriately accounting for the cellular context. This association test is based on the following simplified variant of CellRegMap,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{g}\boldsymbol{\beta}_G + \mathbf{u} + \boldsymbol{\epsilon}, \tag{3.60}$$

where

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2_C\mathbf{K} + \sigma^2_R\mathbf{R}), \tag{3.61}$$

and all other variables are defined below eq. 3.49. The evidence for persistent genetic effects is assessed under the null and alternative hypotheses

$$H_0 : \beta_G = 0 \text{ vs. } H_1 : \beta_G \neq 0, \tag{3.62}$$

using a likelihood-ratio test (section 2.1.5). Maximum-likelihood estimates for the null and alternative models are again computed using the FaST-LMM algorithm.

3.2.5 GxC effect size estimation

For each SNP-gene pair, CellRegMap can be used to estimate the per-allele genetic effect due to GxC in individual cells, corresponding to $\boldsymbol{\beta}_{GxC}$. Note that parameter estimation for

CellRegMap is based on the marginal model likelihood, 3.53 and $\boldsymbol{\beta}_{GxC}$ is not estimated explicitly. Instead, $\boldsymbol{\beta}_{GxC}$ can be obtained from the predictive distribution for $\mathbf{y}$ under a Gaussian process interpretation of the CellRegMap linear mixed model (for details, see [186]). For a given cell let $y^*(\text{ref})$ and $y^*(\text{alt})$ be the estimated expected expression of the target gene for the reference $g^* = 0$ and alternative allele $g^* = 1$ of the QTL variant, respectively, assuming no persistent effects ($\beta_G = 0$). Then, β^*_{GxC} can be estimated as

$$\beta^*_{GxC} = y^*(\text{alt}) - y^*(\text{ref}). \tag{3.63}$$

Let p be the minor allele frequency (MAF) of the variant, that is, the frequency at which the alternative allele ($g = 1$) occurs in the population,

$$g \sim \text{Bin}(2, p). \tag{3.64}$$

Under the assumption of Hardy-Weinberg equilibrium (such that allele frequencies remain constant from one generation to the next), the average heritability (here, fraction of phenotypic variation) explained by the QTL variant can be computed as (see for example [231])

$$1/\sqrt{2p(1-p)}\beta^*_{GxC}, \tag{3.65}$$

where the scaling factor accounts for the variance of g under the binomial model 3.64.

3.2.6 A semi-synthetic simulation framework

To validate the calibration of CellRegMap and assess statistical power, I applied the model to simulated data. In particular, I developed a semi-synthetic simulation procedure, which builds on empirically observed genotypes, gene expression data & cellular contexts from the scRNA-seq dataset of differentiating human iPS cells [84] (see section 3.1.7). In order to evaluate the statistical calibration of the proposed test procedure one needs to obtain realistic gene expression profiles under the null, i.e., in the absence of (context-specific) genetic effects. One option is to permute the observed genotypes, while maintaining the true cell-to-individual assignment. However, this approach would also remove the possible confounding influence of subtle interaction effects between genetic relatedness (kinship) and the molecular context ($\sigma^2_{RxC}\mathbf{R} \odot \mathbf{K}$ in the CellRegMap model, eq 3.49 and below). I therefore propose an alternative strategy, that does not rely on permuted data. First, SNPs and target genes are sampled uniformly from different chromosomes, thereby avoiding the possibility of confounding the simulated eQTL with existing *cis* eQTL in the data. Furthermore, *trans*

eQTL effects are generally very small [50], making it unlikely that such an association (if present) could be detected at the sample sizes considered here. It is therefore reasonable to expect that the measured expression of the chosen target gene is largely independent of the variant allele. To simulate true positive genetic effects, I then sampled synthetic counts for each cell using a conventional linear interaction model with Poisson likelihood,

$$\tilde{y}_n \sim \text{Poisson}(\lambda_n), \quad \lambda_n = \exp\left(y_n + \sum_k g_n \odot c_{n,k}(\beta_{GxC})_k + g\beta_G\right), \tag{3.66}$$

where

- y_n is the log-transformed observed (background) gene expression for a given gene and cell n in the reference dataset,
- g_n is the variant genotype from the reference dataset,
- $c_{n,k}$ denotes the k-th context variable,
- $(\beta_{GxC})_k \sim \mathcal{N}(0, \sigma^2\rho_{GxC})$ is the interaction effect size for context k,
- $\beta_G \sim \mathcal{N}(0, \sigma^2(1-\rho_{GxC}))$ is the effect size of the persistent genetic effect,
- σ^2 is the total genetic variance and ρ_{GxC} is the fraction of genetic variance explained by GxC.

Notably, possible confounding factors such as read count distribution (dropout, overdispersion), batch effects or context-specific expression variation present in the observed expression counts do not need to be simulated using a parametric or model-based approach.

Synthetic data for 500 gene-SNP pairs was generated using real genotypes (50 individuals), background gene expression profiles (100 cells per individual) and cellular contexts obtained using factor analysis (MOFA [130]) of the full expression matrix. I primarily focused on simulating continuous effects by constructing a cell covariance matrix from the observed MOFA factors. Additionally, as part of the power assessment, I also considered discrete contexts.

Test calibration

I assessed the statistical calibration of the proposed tests, CellRegMap and CellRegMap-Association, as well as three alternative models (**Fig.** 3.25a and appendix A.2, **Fig.** A.1):

- StructLMM [184],
- SingleEnv-LRT, a fixed-effect version of CellRegMap, where we test for GxC interactions with individual context dimensions using a likelihood ratio test and report the minimum p-value across all contexts (Bonferroni-adjusted for the number of contexts), similar to [229],
- MultiEnv-LRT, a fixed-effect version of CellRegMap with a multiple-degree-of-freedom likelihood ratio test for GxC effects.

Both SingleEnv-LRT and MultiEnv-LRT share the same null model as CellRegMap. Data were simulated assuming only persistent ($\rho_{GxC} = 0, \sigma^2 = 0.025$, A.2, **Fig.** A.1) or no genetic effects ($\sigma^2 = 0$, **Fig.** 3.25a) and testing for GxC effects using either 10 (**Fig.** 3.25a) or 20 (**Fig.** A.2, **Fig.** A.1) MOFA factors. All models control for the same number of background contexts as tested (additive effects of environmental context and context-repeat-structure interaction).

I confirmed the statistical calibration of CellRegMap in all simulated scenarios. StructLMM produced strongly inflated p-values, indicating that accounting for kinship and repeat structure is key to limiting false positive discoveries. As shown before [184], MultiEnv-LRT does not retain calibration for larger numbers of context variables and was therefore excluded from other simulation experiments (**Fig.** A.2).

Statistical power

Next, I evaluated statistical power for CellRegMap, CellRegMap-Association and SingleEnv-LRT in three different settings (all simulations assume $\sigma^2 = 0.025$, **Fig.** 3.25b). Initially, I varied ρ_{GxC}, the fraction of genetic variance explained by GxC (0, 0.25, 0.5, 0.75, 1.0) for 10 tested and simulated contexts. The power of both GxC tests increased as the fraction of the genetic effect explained by GxC increases, noting that CellRegMap was substantially better powered than the SingleEnv-LRT test. In addition to the CellRegMap interaction test, I assessed the CellRegMap-Association test which as expected is best powered to identify variants with primarily association signals. As a second parameter, I varied the number of cellular contexts that are simulated to contribute to GxC (out of 20 included in both tests). The results of this analysis show that CellRegMap outperformed the corresponding SingleEnv-LRT GxC test when larger numbers of cellular contexts were simulated to contribute to GxC

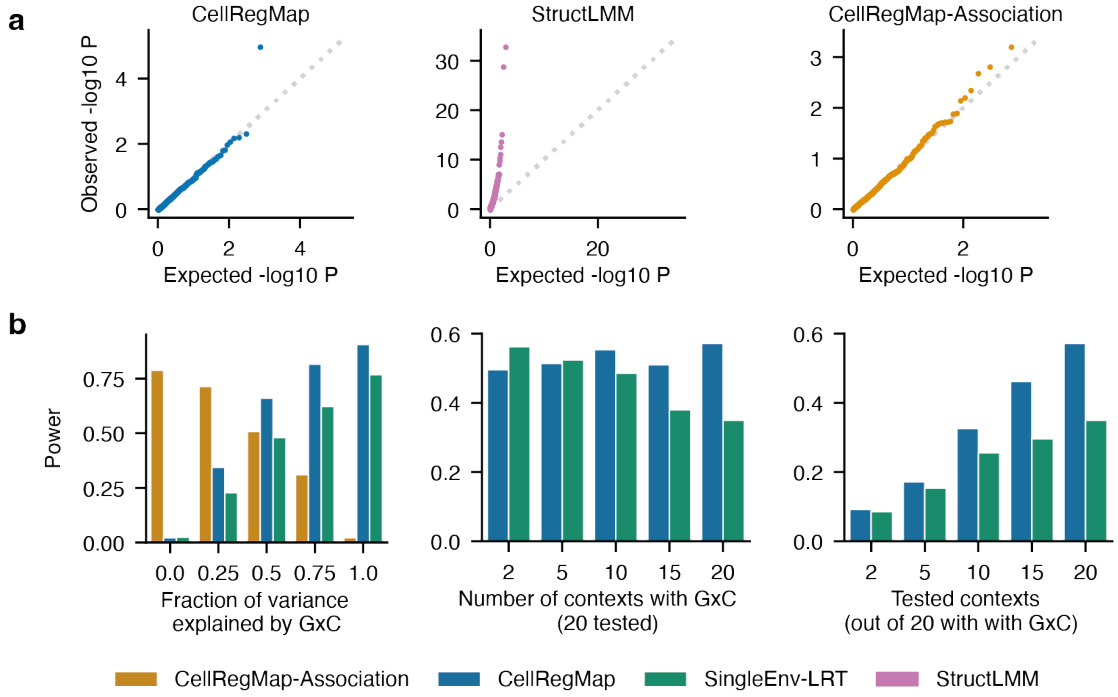

Figure 3.25. CellRegMap validation using simulated data. Test performance for 500 simulated semi-synthetic eQTL based on real expression profiles and genotypes. **(a)** Test calibration under the null hypothesis (without any genetic effects). StructLMM, a model that does not account for the repeat structure in single-cell sequencing data yields inflated test statistics. p-values from CellRegMap and CellRegMap-Association, a variant of CellRegMap for detecting persistent genetic effects only (Materials and Methods), follow the expected uniform distribution. **(b)** Power at significance level $\alpha = 0.01$ as a function of the fraction of genetic variance explained by GxC (left), the number of simulated contexts with GxC (middle) and the number of tested contexts (out of 20 all contributing to GxC, right). Compared are CellRegMap, CellRegMap-Association (where applicable) and a fixed-effect likelihood-ratio-test for single contexts (minimum p-value across all contexts, Bonferroni-adjusted for the number of tested contexts).

(> 5 contexts). I also varied the number of cellular contexts tested in the model, again finding that CellRegMap offers advantages for larger numbers of contexts.

Furthermore, I considered discrete cellular contexts derived using Leiden clustering [100, 214] (based on 20 MOFA factors, resolution of 0.5 and 1.0, resulting in 12 and 24 clusters, respectively) or based on the observed sample collection timepoints (Day 1-4). Compared a continuous context definition (based on 20 MOFA factors) these representations led to a significant reduction in discovery power (**Fig.** 3.26). Taken together, these results demonstrate power advantages and robustness of CellRegMap, compared with existing methods, particularly when multiple cellular contexts contribute to GxC.

Finally, I used simulated data to assess the impact of expression level and expression variance on the power to detect genuine GxC effects, finding that the power to identify GxC effects is increased for genes with higher overall expression level mean and lower variance (10 tested and simulated contexts, **Fig.** 3.27)

3.2.7 Runtime complexity

Using the LMM efficient implementation described by Lippert et al [139] one can show (see [186] and [184]) that the runtime scales linearly with the minimum of the number of cells and the product of (number of unique individuals × the number of cellular contexts).

I additionally evaluated the empirical runtime using observed expression profiles and contexts (MOFA factors) from the iPSC differentiation dataset [84] and simulated individuals / genotypes. I assessed the runtime of CellRegMap and CellRegMap-Association as a function of either the total number of cells (5,000, 7,500, 10,000, 12,500 or 15,000 cells sampled without replacement from the full dataset), the number of individuals (50, 75, 100, 125, 150) or the number of contexts tested for GxC effects (2, 5, 10, 15 or 20 leading MOFA factors; using the same number as background effects). Nonvarying parameters were set to a default of 10,000 cells, 100 donors and 10 tested contexts. All experiments were run on an Intel Xeon CPU E5-2660 v4 with 2.00GHz and averaged across 125 simulated eQTL (**Fig.** 3.28).

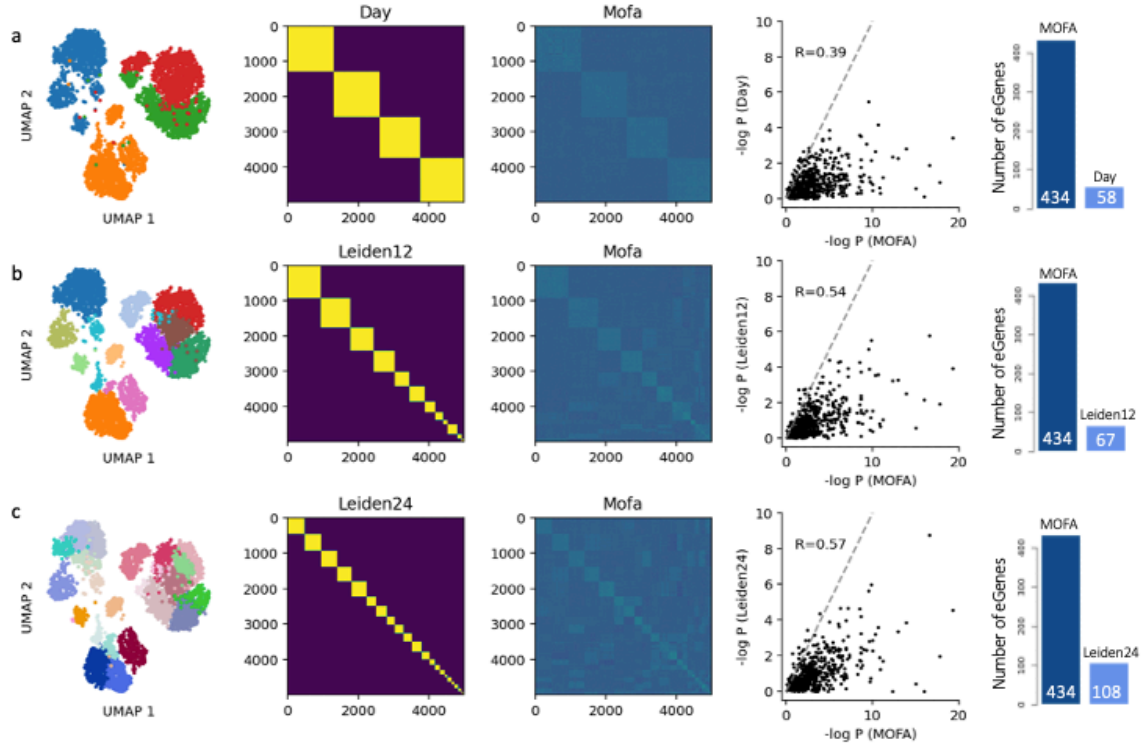

Figure 3.26. CellRegMap performance using discrete cell contexts. Left: 2-dimensional visualisation of the gene expression data and discrete clusters. Middle: Discrete and continuous covariance matrices (based on MOFA factors) sorted by cluster membership. Right: Result comparison; negative log p-values comparing models using continuous (x-axis) vs discrete (y-axis) contexts, as well as bar plots showing the number of significant GxC eQTL identified using the different context-covariance matrices. **(a-c)** Results for 500 semi-synthetic eQTL, considering discrete contexts at three different resolutions: **(a)** the original 4 sampling time points (Day), **(b)** 12 and **(c)** 24 Leiden clusters. Number of significant eGenes is out of 500 tested (FDR < 10%).

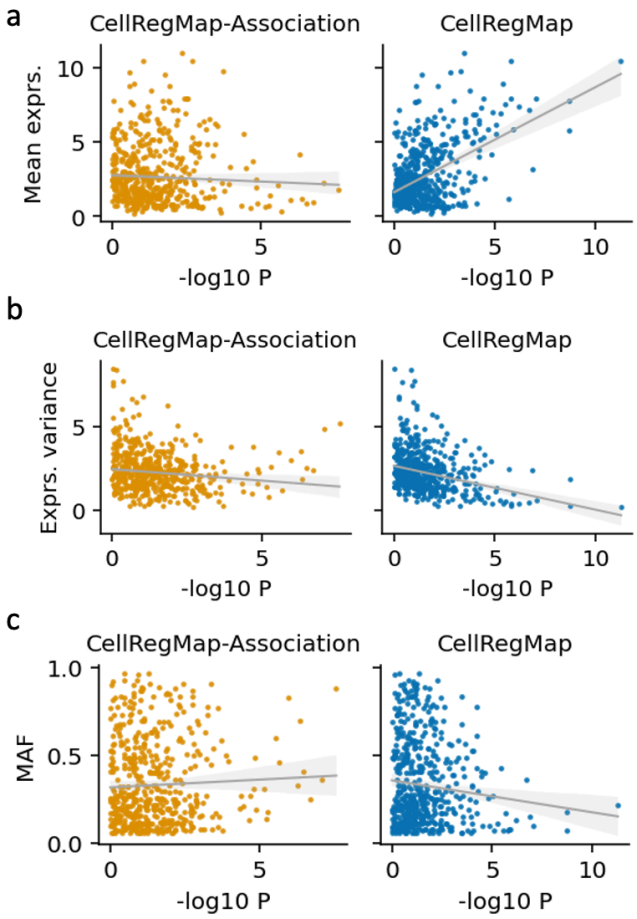

Figure 3.27. CellRegMap p-values stratified for gene properties on simulated data. Results for simulated genetic effects, showing **(a)** the mean observed gene expression of the simulated eGene (prior to adding the genetic effect), **(b)** the prior gene expression variance and **(c)** the minor allele frequency as a function of the P-value estimated by CellRegMap (blue) and CellRegMap-Association (orange). Both models are fitted using the same set of ground truth context variables as used in the simulation. Lines show the regression fit and shaded areas 95% confidence intervals.

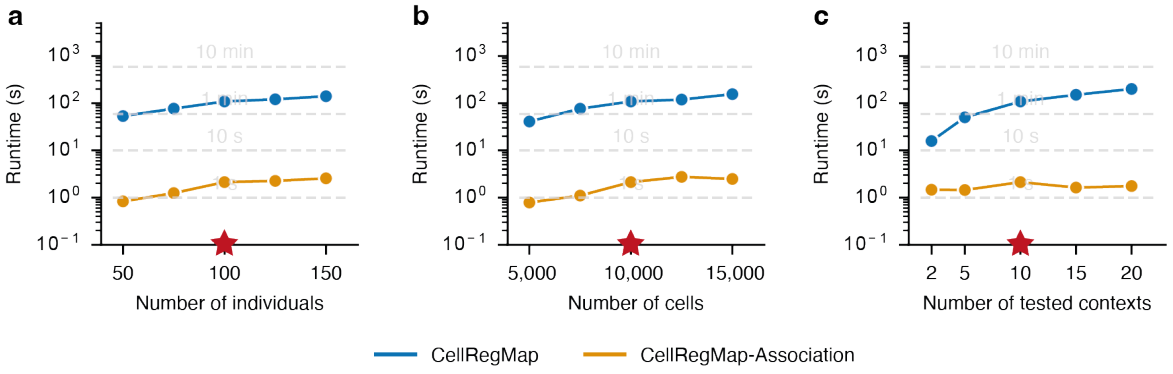

Figure 3.28. Runtime assessment of CellRegMap and CellRegMap-Association test. Shown are empirical runtimes for simulated data averaged across 150 eQTL. Shown are runtimes (y-axis) for testing a single eQTL as a function of the number of individuals (a), the total number of cells (b) and the number of context variables (c); runtimes are evaluated for both the CellRegMap model (interaction test, blue), and the corresponding association test (orange). Stars highlight default values for fixed parameters. All parameters retained at their default parameter values except the parameter that is indicated on the x axis.

3.2.8 Example application to differentiating human iPSCs

The analysis in this section was performed by Anna Cuomo and is included as an illustrative example. Further details and additional application studies can be found in the published paper [84].

We applied our model to map context-specific eQTL from single-cell RNA-seq profiles of differentiating human iPSCs from 125 individuals [84]. As previously discussed (see section 3.1.7), cell differentiation is the dominant cellular context in this study, and hence this dataset is an ideal test case to assess the ability of CellRegMap to identify continuous changes of allelic effects across a cellular trajectory.

Count data were processed as in the primary paper [84], where counts were normalized using scran [232] and log-transformed. The log-normalized count data for the top 500 highly variable genes was as input for MOFA [130] to estimate latent factors that explain variation in gene expression in the data. The inferred representation captured both differences in major cell types across the differentiation trajectory, but also more subtle cell states. The first factor (MOFA 1) primarily explained the differentiation axis, with cells transitioning

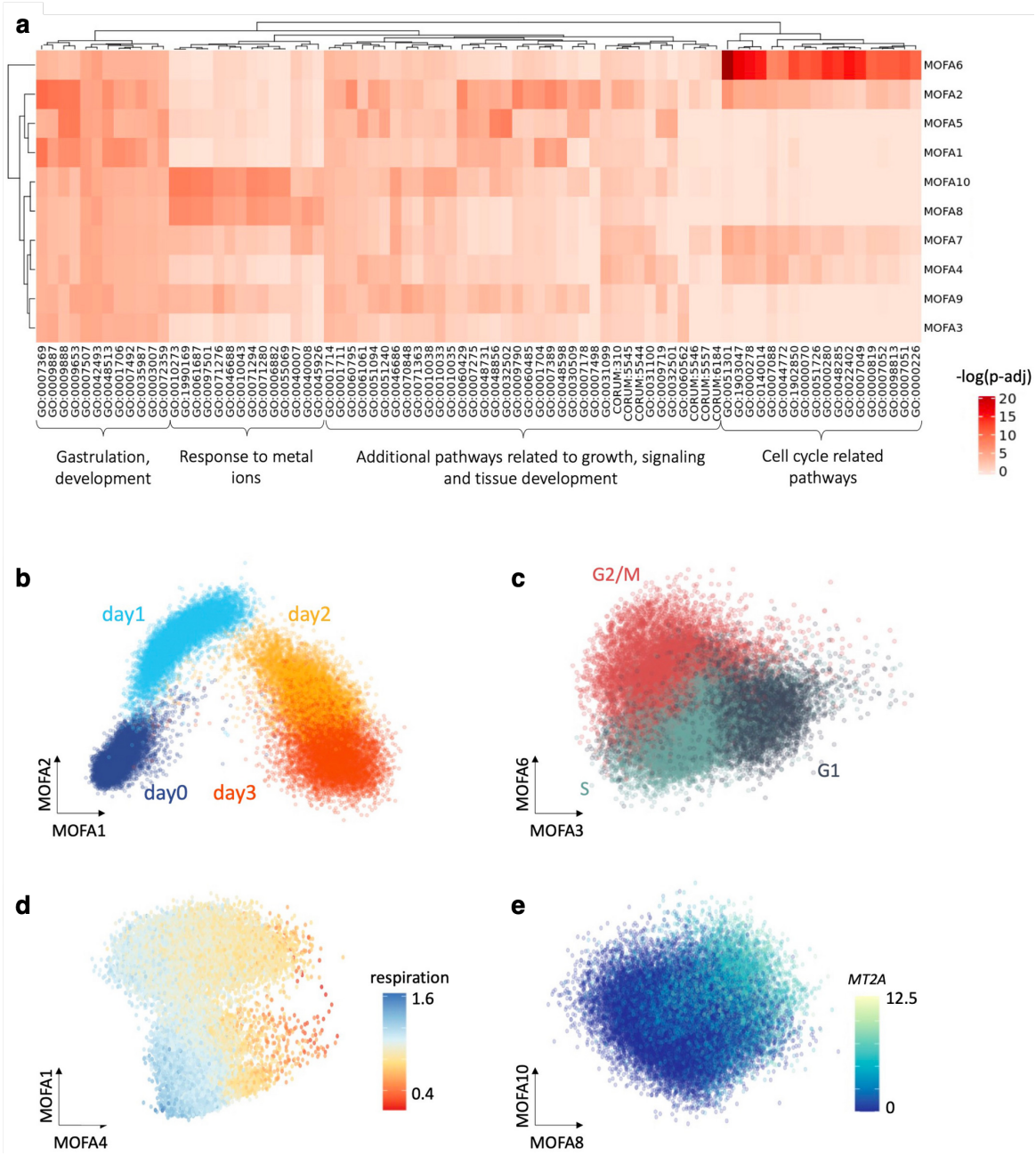

Figure 3.29. Annotation of the MOFA factors from the iPSC differentiation data [84] **(a)** Heatmap displaying negative log p-values from gene ontology enrichment analyses based on the absolute values of the loadings of individual MOFA factors. **(b)** Scatter plot of MOFA factors 1 & 2, with color corresponding to the time point of collection (day 0,1,2 & 3 of endoderm differentiation). **(c)** Scatter plot of MOFA factors 3 & 6, colored by estimated cell cycle phase (G1, G2/M, S; estimated using Seurat). **(d)** Scatter plot of MOFA factors 1 & 4 capturing respiration. **(e)** Scatter plot of MOFA factors 8 & 10, capturing a signature linked to response to metal ions, colored by expression of gene with top loadings, MT2A.

between a pluripotent state and the definitive endoderm fate. Higher order factors captured other cellular contexts, including cell cycle phase (MOFA 3 and 6), respiration (MOFA 4) and others (**Fig.** 3.29).

We applied CellRegMap to test for GxC effects at 4,470 eQTL variant/gene pairs that were previously reported in the primary analysis of the dataset using a conventional eQTL mapping workflow that did not account for GxC interactions [84]. Log-transformed gene expression measurements were quantile-normalized to better fit the Gaussian distribution assumed by the model. We compared CellRegMap when only using the first MOFA factor to define the cell context covariance, which is similar to the approach taken in the primary analysis [84], to a model that leverages the information contained in the leading 10 MOFA factors. The model with 10 components yielded a substantially larger number GxC effects (322 vs. 183, FDR < 0.05; **Fig.** 3.30a), indicating that despite cell differentiation being the major driver of expression variation, other more subtle cellular states also manifest in GxC interactions on gene expression.

Next, we set out to characterize specific cellular contexts that are associated with the identified GxC interactions. We used CellRegMap to estimate the GxC allelic effects in each cell, thereby recovering the continuous landscape of the GxC component of genetic effects across the cell–context manifold. This analysis identified a range of allelic patterns, including GxC effects that are primarily governed by cellular differentiation but also more complex patterns that involve multiple cellular contexts and higher-order cellular factors. For example, the eQTL variant rs113520162 for IER3 had a GxC effect that reflects variation across cell differentiation explained by the first MOFA component (**Fig.** 3.30b, middle). Other eQTL, such as rs11180470 for GLIPR1L1, had GxC effects that were associated with two MOFA factors (**Fig.** 3.30b, right). More generally, we observed that higher order MOFA components capture changes in cellular contexts beyond cellular differentiation, including the cell cycle (**Fig.** 3.30c) and cellular respiration (**Fig.** 3.30d). Collectively these results illustrate how CellRegMap can be used to uncover different cellular contexts that manifest in GxC interactions.

3.2.9 Alternative context definitions

To assess the robustness of the identified GxC effects, I considered alternative latent variable methods to capture cellular contexts. I compared the MOFA workflow to principal compo-

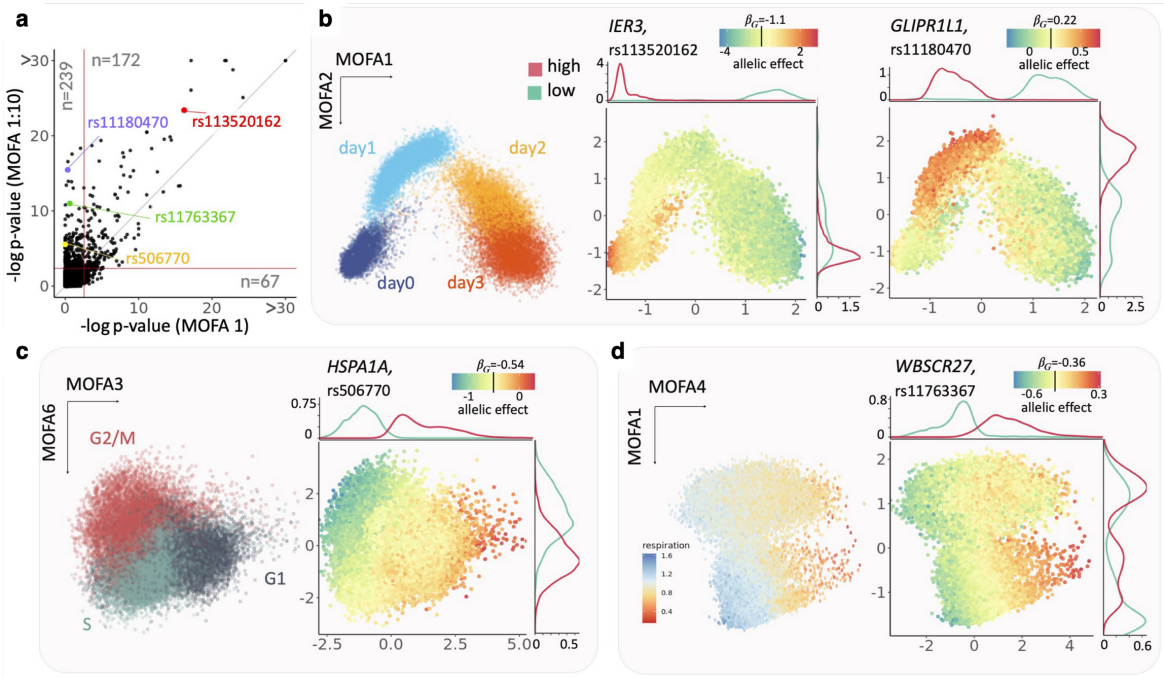

Figure 3.30. Application to human iPSCs. **(a)** Scatter plot of negative log p-values obtained from CellRegMap when using either the first MOFA factor or the leading 10 factors to define the cell context covariance (4,470 eQTL variants and genes). Horizontal and vertical lines denote the FDR < 5% significance threshold (Benjamin-Hochberg adjusted). Shown in each quadrant is the number of eQTL with evidence for a GxC effect. **(b-d)**. Examples of eQTL with GxC interaction. **(b)** Left: scatter plot of the first two MOFA factors (capturing cell differentiation as context) with color denoting the time point of collection; middle: identical scatter plot with color encoding the estimated allelic effect for the eQTL variant rs113520162 for the gene IER3; right: allelic effect for the eQTL at rs11180470 for the gene GLIPR1L1. Shown are allelic effects (β_{GxC}) for individual cells centered on the persistent effect. Marginal densities highlight cells that have either increased (high, red) or decreased (low, cyan) allelic effects (corresponding to the bottom and top 10% quantiles, respectively). Whereas the GxC effect for the eQTL for IER3 is primarily explained by the first MOFA component, the GxC effect for GLIPR1L1 is captured by the combination of the first two MOFA factors. **(c)** Analogous to **(b)**; scatter plot between MOFA factors 3 and 6 with cells colored by alternative annotations. Left: inferred cell cycle phase; Right: allelic effects for an eQTL at rs506770 for HSPA1A (yellow). **(d)** As in **(b, c)** scatter plot of MOFA factors 4 and 1. Left: cells colored by cellular respiration (Materials and Methods); Right: allelic effects for the eQTL at rs11763367 for WBSCR27 (green). Figure credit: Anna Cuomo.

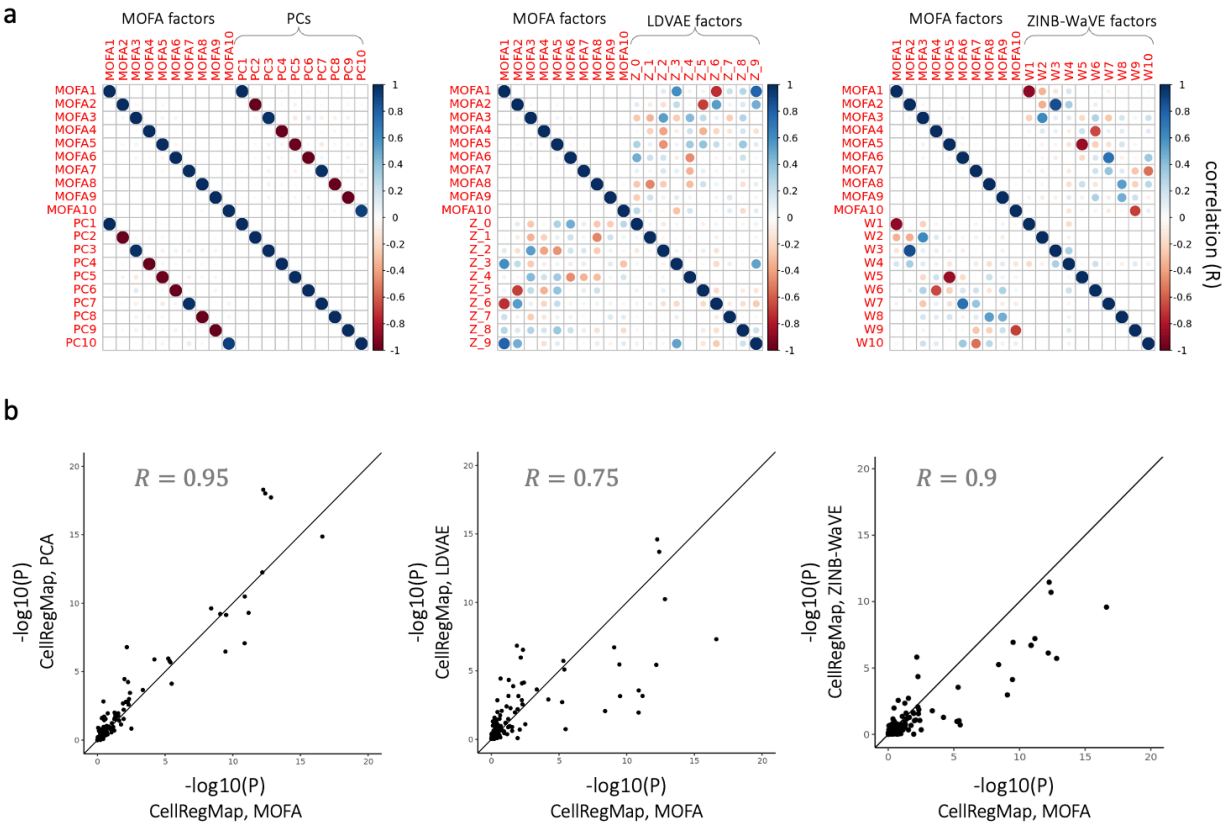

Figure 3.31. Comparison of alternative methods to define cellular contexts from the iPSC data. Compared were the MOFA workflow, principal component analysis (left), LDVAE [173] (middle) and ZINB-WaVE [233]. **(a)** Correlation matrix between the leading 10 factors identified by MOFA and the respective alternative method. **(b)** Scatter plot of negative log p-value of the CellRegMap interaction test when using MOFA (x-axis) vs. alternative methods (y-axis) for defining cell contexts.

nent analysis, linearly-decoded scVI [131,173], as well as ZINB-WaVE [233], a factor analysis method with zero-inflated negative binomial likelihood. Considered were CellRegMap test results for 121 SNP gene pairs from 88 unique genes on chromosome 22. Negative log p-values were highly correlated (Pearson's R 0.75 to 0.95, **Fig.** 3.31), and highly significant associations could be replicated using either representation.

3.2.10 Discussion

Here, I described the Cellular Regulatory Map (CellRegMap), a linear mixed model for the identification and characterization of context-specific eQTL that is applicable to cellular states derived from scRNA-seq. Building on the methodology developed for scDALI, CellRegMap uses cell manifolds derived from single-cell transcriptome profiles to estimate cellular contexts in an unbiased manner to then test for genetic interaction effects.

Conceptually, CellRegMap is also related to and extends StructLMM, a model that was originally designed to identify genotype-environment interactions in population cohorts [184]. CellRegMap adapts these concepts to single-cell genomics, by including an additional relatedness component in the model to account for dependencies across cells that are assayed from the same individual. CellRegMap retains calibrated test statistics and enjoys power benefits compared with conventional fixed-effect interaction tests. Additionally, framework is complemented with a genetic association test designed specifically for single-cell sequencing data (CellRegMap-Association), allowing to efficiently generate sets of candidate eQTL to be tested for GxC.

To illustrate the model, CellRegMap was applied to a single-cell dataset of iPS cells from 125 individuals across differentiation towards a definitive endoderm fate [84]. The dominant source of variation in this dataset is a continuous differentiation signal, which manifests in dynamic eQTL across differentiation. Nevertheless, CellRegMap also identifies eQTL associated with other dimensions of transcriptome variation, including factors associated with cell-cycle phase or respiration. These results highlight how CellRegMap can be used to map heterogeneity in genetic effects even in seemingly homogeneous systems and pinpoint eQTL signals to specific subpopulations and molecular processes.

Although we demonstrated that CellRegMap is broadly applicable to different datasets and scRNA-seq technologies, the model is not free of limitations. Currently, application of the

CellRegMap model to single-cell count data requires appropriate processing steps (e.g., variance stabilization and quantile-normalization) to provide cell-level or pseudo-cell expression estimates that approximately follow a Gaussian distribution. Although our results indicate that this approximation is acceptable in practice and retains statistical calibration, explicit modeling of count data could provide additional power benefits, in particular in the regime of lowly expressed genes. Furthermore, for computational reasons, the primary focus of CellRegMap lies on the annotation of known eQTL variants rather than variant discovery. An analogous two-stage strategy is used for mapping genotype-environment interactions at known GWAS loci in population cohorts. Such procedures build on the assumption that the persistent genetic effect signal is sufficiently strong to enable discovery. The CellRegMap-Association test implemented as part of the software can be used to define an end-to-end workflow in the cases where eQTL are not known a priori. However, future extensions of CellRegMap could focus on improving computational scalability in order to enable the discovery of eQTL variants while accounting for GxC.

Chapter 4

LIVI: identifying context-specific genetic effects on gene modules from population-scale single-cell RNA-seq data

Expression quantitative trait loci studies, which test for associations between genetic variants and inter-individual variation in gene expression, hold promise to elucidate the function of disease-associated variants and their role in disease initiation and progression. The previous chapter has focused on the development of methods for the identification of *cis*-regulatory effects, where quantitative traits such as chromatin accessibility or gene expression are affected by proximal (typically < 1Mb) genetic variants. However, given the complexity of gene regulatory networks (GRNs), genetic effects are expected to also regulate groups of genes in *trans* (see chapter 1.2; **Fig.** 1.1), mediated via regulatory dependencies and networks. Indeed, *trans* eQTL effects have previously been identified in bulk tissue samples using latent factor or topic models [234–236], providing maps of genetic regulation of gene expression in different human tissues. Under the premise that genes which co-vary may also be co-regulated, these methods test for associations between genetic variants and latent factors of gene expression data, representing structured changes across many different genes. This approach significantly reduces the multiple testing burden incurred by testing at the gene level, which is of particular importance for *trans*-effect mapping where effect

sizes tend to be much smaller compared to *cis* eQTL [50]. Recent advances in single-cell RNA sequencing (scRNA-seq) and their application to population-scale cohorts open up the possibility to identify genetic effects that are specific to individual cell types or subtypes [50, 84–86]. However, there exist no methods that leverage single-cell resolution to identify persistent and context-specific *trans*-regulatory effects.

Here, I propose ***L**atent **I**nteraction **V**ariational **I**nference* (LIVI), which combines scalable stochastic variational inference, with the interpretability of linear latent factor models, allowing for fast mapping of both persistent genetic effects and context-specific effects that arise from interactions between genetic variants and continuous single-cell states. LIVI builds on a Variational Autoencoder [144] with linear decoder [173] to jointly model single-cell gene expression measurements from multiple donors in a population cohort. The model disentangles canonical cell state variation from donor-specific effects using an adversarial approach [237, 238] to then explicitly reintroduce donor-specific effects in the latent space. LIVI captures both discrete and continuous cell states from single-cell expression profiles in an unsupervised manner and does not require predefined cell type annotations. I validate LIVI on simulated data and apply it on a real dataset of more than one million cells from a thousand donors [86].

Acknowledgements and contributions

This work was supervised by Oliver Stegle. LIVI was co-developed by Danai Vagiaki and myself. I implemented the model and performed the evaluation on simulated data. Danai Vagiaki applied the model to the OneK1K dataset [86].

4.1 Previous work

Among methods that consider single-cell readouts for identification of *trans* effects, LIVI is related to [236], which also seeks to estimate latent representations of scRNA-seq data informed by genetics. However, the downstream testing procedure considers individual genes, incurring a higher multiple testing burden. Furthermore, genetic variants are used during the inference of the latent space, which necessitates a computationally expensive permutation scheme to obtain calibrated test statistics. The approach does not use a noise model tailored to single-cell sequencing data and relies on aggregation of single-cell readouts to ”pseudo-cells” as a pre-processing step. Another related line of work uses latent factor/topic models

to conduct genetic analyses in the latent space [234, 235], which however do not distinguish between interaction and persistent genetic effects. Conceptually most closely related to the work presented here is MrVI [239], which also uses a VAE model to assess context-specific and persistent sample effects on scRNA-seq data. A key innovation of LIVI is that the model employs an adversarial approach to explicitly disentangle donor effects from canonical cell states, which I show to improve statistical power on simulated data. Additionally, the hypothesis test proposed in [239] requires discrete sample covariates, whereas LIVI allows for rapid testing using arbitary genetic or non-genetic donor covariates.

4.2 The LIVI model

LIVI is a probabilistic model for identifying context-specific effects of inter-donor variation on latent gene expression factors. The model projects cells and donors into the same latent space, summarizing population-level effects in single vectors for each donor (**Fig.** 4.1). These donor embeddings can then be used in downstream analyses, both in an explorative manner as well as to rapidly test for associations with genotypes and clinical covariates.

4.2.1 Modeling population-scale data

In the population scRNA-seq setting, gene expression profiles $\mathbf{x} \in \mathbb{R}^D$ for each cell are paired with an donor label y. LIVI builds on the basic VAE model and incorporates this grouping structure to capture donor-specific effects on latent gene expression factors.

For each donor $y = 1, \ldots, M$, let $\mathbf{u}_y, \mathbf{v}_y \in \mathbb{R}^K$ denote trainable embedding vectors of a K-dimensional latent space. LIVI then decomposes the cell latent space for donor y as the sum of shared and donor-specific variation:

$$\mathbf{z} = \mathbf{c} + \mathbf{c} \odot \mathbf{u}_y + \mathbf{v}_y \tag{4.1}$$

where $\mathbf{c} \sim \mathcal{N}(0, 1)$ and $\odot$ denotes the Hadamard product. Here, the latent variables $\mathbf{c} \in \mathbb{R}^K$ model variation that is shared across all donors, e.g. canonical cell types and states (contexts). The vectors u_y and v_y represent dynamic and persistent donor effects on the latent dimensions. That is, the distribution of $\mathbf{z}$ is Gaussian, with donor-specific mean and variances,

$$\mathbf{z} \sim \mathcal{N}(\mathbf{v}_y, \text{diag}(\mathbf{u}_y)^2). \tag{4.2}$$

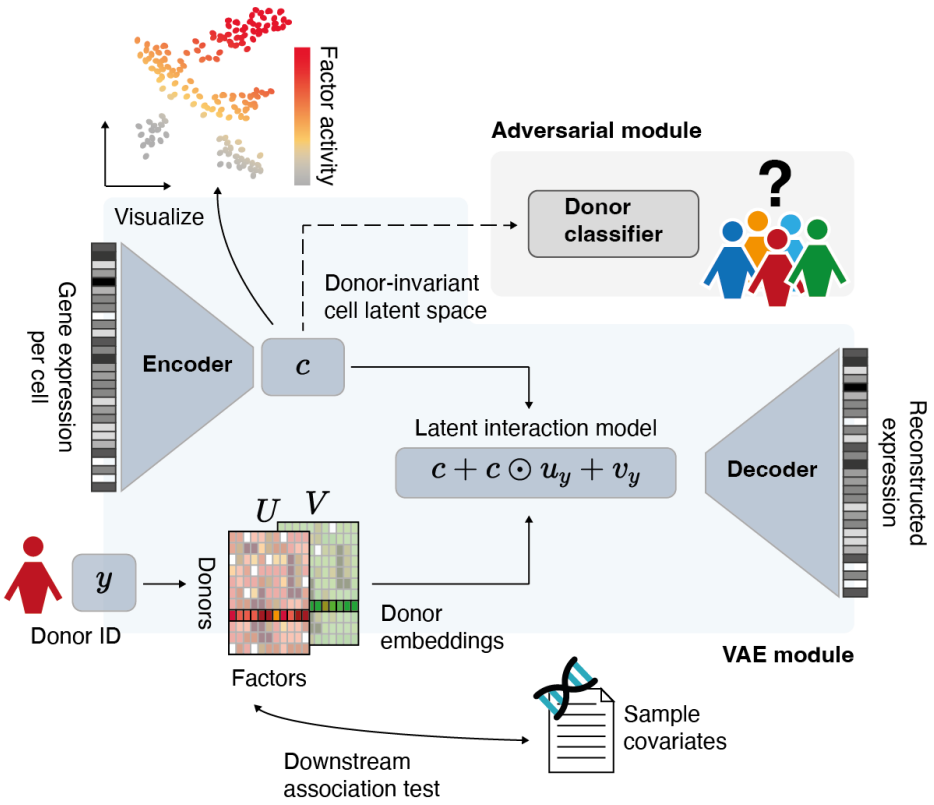

Figure 4.1. LIVI overview. Gene expression vectors for each cell are mapped to base cell states $\mathbf{c}$, capturing canonical cell types and states. An adversarial classifier encourages removal of sample-specific variation from base states. For each donor y, the model infers two embedding vectors, $\mathbf{u}_y, \mathbf{v}_y$, summarizing cell-state-specific and additive effects on latent expression factors. Trained models can be used for efficient association and interaction tests and effect visualization.

To allow for interpretability of the latent factors, LIVI employs a linear decoder [173] to map latent variables to expression frequencies

$$\boldsymbol{\rho} = \text{Softmax}(\mathbf{W}^T\mathbf{z} + \mathbf{b}) \in [0,1]^D. \tag{4.3}$$

Gene expression profiles are assumed to be drawn from a size-factor-adjusted negative binomial distribution [131], to account for technical noise due to the sampling process and residual over-dispersion

$$\mathbf{x} \,|\, \mathbf{z} \sim \text{NB}(\boldsymbol{\rho} l, \boldsymbol{\theta}), \tag{4.4}$$

where $\boldsymbol{\theta} \in \mathbb{R}^D$ is a trainable vector of dispersion parameters for each gene. scVI [131] models the size factor l as an unobserved latent variable. In practice, however, l typically

converges to the total observed transcript count, $\sum_{d=1}^{D} x_d$. Therefore l is treated as observed, $l = \sum_{d=1}^{D} x_d$.

4.2.2 Adversarial penalty and variational inference

Conditional VAE models have been successfully used to integrate data across different sequencing technologies and batches [131] (see also section 2.2.2). Contrary to batch effects, however, genetic variation explains only a small fraction of the overall expression variance. To further constrain the model to factorize all donor-specific effects from the invariant latent space, an adversarial approach [237, 238] is used. Let $f : \mathbb{R}^K \mapsto [0,1]^M$ be a multi-layer fully-connected network with softmax activation as a final layer. The network f is trained to predict the donor label from the latent representation $\mathbf{c}$, by minimizing the cross entropy loss ℓ_{adv},

$$\ell_{adv}(f(\mathbf{c}), y) = -\log f(\mathbf{c})_y. \tag{4.5}$$

Assuming a sufficiently flexible parameterization for f, the performance of the trained classifier provides a measure of the amount of information on the donor label y preserved in the latent representation $\mathbf{c}$ of a particular cell. Conversely, the VAE model is trained to fool the classifier, thereby removing donor-specific variation from the shared latent space. That is, the VAE parameters are updated to maximize the compound loss,

$$\underbrace{\mathbb{E}_{q(\cdot|\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{c}, y)p(\mathbf{c})}{q(\mathbf{c}|\mathbf{x})}\right]}_{\text{ELBO}} + \lambda_{adv}\mathbb{E}_{q(\cdot|\mathbf{x})}\left[\ell_{adv}(f(\mathbf{c}), y)\right] \tag{4.6}$$

where the parameters of the classifier f are being held fixed, and λ_{adv} is a hyperparameter. The first term corresponds to the standard variational objective, where I approximate the true posterior using a probabilistic encoder,

$$q(\mathbf{c} \mid \mathbf{x}) = \mathcal{N}(\mathbf{c} \mid \boldsymbol{\mu}(\mathbf{x}), \text{diag}(\sigma^2(\mathbf{x}))), \tag{4.7}$$

analogous to the standard VAE (section 2.2). That is, $\boldsymbol{\mu}$ and σ^2 are multi-layer fully-connected networks mapping $\mathbf{x}$ to the parameters of the Gaussian variational distribution. Using the approximate posterior, the evidence lower bound (ELBO) on the marginal likelihood $p(\mathbf{x})$ can be evaluated and optimized using the reparameterization trick as discussed in section 2.2. The donor embeddings $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{M \times K}$ are viewed as part of the observation model $\mathbf{x} \mid \mathbf{c}$ and can be optimized using stochastic gradient ascent on compound objective 4.7, together with the decoder parameters $\mathbf{W}, \mathbf{b}, \boldsymbol{\theta}$ and posterior parameters.

The full training procedure alternates between training the classifier f, while keeping the VAE parameters fixed and updating the VAE (eq. 4.7).

4.2.3 Testing for genetic and covariate effects

The matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{M \times K}$ of donor embeddings summarize interaction and persistent effects on gene expression factors at the level of donors rather than cells. As a result, they enable fast downstream hypothesis testing of covariate associations. Let $\mathbf{g} \in \mathbb{R}^M$ denote a covariate vector, e.g. encoding variant genotypes or a clinical variable such as disease status. To identify interaction effects on factor k (LIVI (int)), I use a simple linear mixed model

$$\mathbf{u}_k = \mathbf{g}\beta + \mathbf{H}\boldsymbol{\alpha} + \mathbf{u} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\mathcal{I}}_M), \tag{4.8}$$

and assess $\beta \neq 0$ using a likelihood ratio test. Here, $\mathbf{H}$ is an optional matrix of fixed effects covariates, e.g., to model batch effects, and the random effect term $\mathbf{u}$ accounts for population structure. Notably, by testing donor embeddings that have been trained to satisfy 4.1, interaction effects can in principle be detected using a simple additive model. An analogous procedure can be used to test for persistent effects using $\mathbf{V}$ (LIVI (add)).

4.3 Evaluation on synthetic data

To assess the influence of data generating parameters such as genetic effect sizes and the degree of "discreetness" of the latent cell states, I evaluated our model on synthetic data with known groundtruth. Let N denote the number of cells and D the number of genes. I simulated genetic effects of L variants on latent factors using a latent decomposition $\mathbf{Z} \in \mathbb{R}^{N \times K}$ analogous to the LIVI model:

$$\mathbf{Z} = \mathbf{C} + \mathbf{C} \odot \mathbf{GU} + \mathbf{GV} \tag{4.9}$$

Here, $\mathbf{C} \in \mathbb{R}^{N \times K}$ is the donor-invariant latent space, $\mathbf{G} \in \mathbb{R}^{N \times L}$ a cell-level genotype matrix and $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{L \times K}$ the effect sizes for additive and interaction effects, respectively. All simulations assumed that each variant affects at most a single factor. Latent variables were mapped to high-dimensional simulated mean expression profiles using a linear decoder, $\mathbf{X} = \mathbf{ZW}$.

4.3.1 Estimation of effect sizes and explained variance

The magnitude of genetic effects on simulated gene expression profiles depends on the factor variance, the genetic effect sizes as well as the loading matrix $\mathbf{W} \in \mathbb{R}^{K \times D}$. To gauge the effect strength of SNP-factor associations, I therefore evaluated the explained variance at the gene level. Here, I focus on the interaction term, however, an analogous argument can be used to set effect sizes for additive effects, $\mathbf{V}$.

Under a linear decoder model, one can defer the evaluation of the genetic effect to the gene space as follows. Let $\mathbf{u}_l$ be the column vector corresponding to the l-th row of $\mathbf{U}$. Then,

$$(\mathbf{C} \odot \mathbf{GU})\mathbf{W} = \sum_{l=1}^{L} \left(\mathbf{C} \odot \mathbf{g}_l \mathbf{u}_l^T \right) \mathbf{W} = \sum_{l=1}^{L} \underbrace{\mathbf{C}\text{diag}(\mathbf{u}_l)\mathbf{W} \odot \mathbf{g}_l \mathbf{1}_D^T}_{=:\mathbf{X}_l^{\text{int}}} \tag{4.10}$$

where $\mathbf{1}_D$ is the D-dimensional vector of ones. I quantify the total explained variance by a matrix $\mathbf{X}$ as the sum of gene (column) sample variances[1]

$$\text{Var}_{\text{Total}}(\mathbf{X}) := \sum_{d=1}^{D} \text{Var}_s(\mathbf{x}_d). \tag{4.11}$$

Let $\sigma^2_{\text{int}} = \sum_{l=1}^{L} \text{Var}_{\text{Total}}(\mathbf{X}_l^{\text{int}})$ be a total desired variance of the interaction term at the gene level. Suppose k is the (single) factor affected by variant l. The associated effect size, the entry u_{lk} in $\mathbf{U}$, is chosen as follows. Note that the column of $\mathbf{X}_l^{\text{int}}$ corresponding to gene d can be written as

$$u_{lk} w_{kd} \mathbf{c}_k \odot \mathbf{g}_l, \tag{4.12}$$

and therefore

$$\text{Var}_{\text{Total}}(\mathbf{X}_l^{\text{int}}) = \sum_{d=1}^{D} \text{Var}_s(u_{lk} w_{kd} \mathbf{c}_k \odot \mathbf{g}_l) \tag{4.13}$$

$$= u_{lk}^2 \sum_{d=1}^{D} w_{kd}^2 \Big(\mathbb{E}_s[\mathbf{c}_k^2]\text{Var}_s(\mathbf{g}_l) + \text{Var}_s(\mathbf{c}_k)\mathbb{E}_s[\mathbf{g}_l^2] + \text{Var}_s(\mathbf{c}_k)\text{Var}_s(\mathbf{g}_l) \Big), \tag{4.14}$$

where $\mathbb{E}_s$ is the sample mean and the genotype and context vectors have been assumed to be independent. If $\mathbf{c}_k$ and $\mathbf{g}_l$ have been standardized to mean zero and unit variance, the above simplifies to

$$\text{Var}_{\text{Total}}(\mathbf{X}_l^{\text{int}}) = u_{lk}^2 ||\mathbf{w}_k||_2^2 \tag{4.15}$$

[1] That is, $\text{Var}_s(\mathbf{x}_d)$ corresponds to the population variance of the finite population $x_{1d}, \ldots, x_{Nd}$.

where $||\cdot||_2$ is the euclidean norm. Let R be the total number of variants with non-zero interaction effects. To match the desired total variance of the interation term, σ^2_{int}, u_{lk} can be chosen as

$$u_{lk} = \pm\sqrt{\frac{\sigma^2_{\text{int}}}{R||\mathbf{w}_k||^2_2}}. \tag{4.16}$$

4.3.2 Simulation setup

I simulated 40-dimensional cell states for 30,000 cells, by randomly assigning cells to one of 5 fictional celltypes and drawing each factor independently as

$$\mathbf{c}_k \sim \mathcal{N}(0, \eta\mathbf{E}\mathbf{E}^T + (1-\eta)\boldsymbol{\mathcal{I}}_N), \tag{4.17}$$

where $\mathbf{E}$ denotes one-hot encoded celltype assignments and $\eta \in [0, 1]$ controls the strength of intra- vs inter-celltype variation, i.e., discrete vs. continuous effects.

I simulated 100 donors and sampled persistent and interaction effects for non-overlapping subsets of 60 variants. For each causal variant, I sampled a single associated factor. I fixed the fraction of cumulative genetic variance across all variants and genes and chose variant-factor effect sizes as described in the previous section. I then sampled synthetic expression counts from a Poisson noise model. Default simulation hyperparameters are summarized in table 4.1.

As a baseline, I considered a Variational autoencoder with linear decoder and negative-binomial observation model [131, 173]. Similar to LIVI, I used an implementation that modeled size-factors as observed rather than latent variables. Notably, the baseline model did not incorporate donor labels such that inter-donor effects could be preserved in the latent space. The model was combined with post-hoc clustering in the latent space to test for genetic effects on average factor activities in each cluster (VAE + LM). LIVI parameters used in all simulations are shown in table 4.2. Power was evaluated using the minimum P-value across factors (Bonferroni-adjusted) for each variant. Models were evaluated across 5 random initializations.

4.3.3 Impact of adversarial penalty

First, I assessed the effect of the adversarial penalty on power to detect interaction effects (LIVI (int)), varying the weight of the auxiliary loss $\lambda_{adv} \in \{0, 10, 25, 50, 100, 250, 500\}$.

Parameter	value
Latent dimension	40
Cells	30,000
Latent space dimension	40
Donors	100
Genes	1000
Variants	60
Min. minor allele frequency	0.3
Max. minor allele frequency	0.5
Number of persistent effects	20
Number of interaction effects	20
Number of control variants	20
FEV by genetics	0.02
Fraction of genetic variance explained by interaction effects	0.7
Celltypes	5
η	0.5

Table 4.1. Default simulation parameters.

Parameter	value
Latent dimension	40
Encoder hidden nodes	[256, 256]
Learning rate	8e-4
λ_{adv}	300
Adversary hidden nodes	[256, 256]
Adversary learning rate	1e-4
Adversary update frequency	2

Table 4.2. Default model parameters for simulated data.

For comparison, I also tested for linear associations between simulated genetic variants and average factor activities in Leiden clusters [100, 214] of the shared latent space **c**, (LIVI + LM). Increasing λ_{adv} successfully reduced the number of significant associations identified by LIVI + LM, showing that the adversarial penalty aids in removing genetic effects from the shared representation (**Fig.**. 4.2a). Conversely, this analysis confirmed that the power to detect genetic interaction effects using LIVI (int) was improved for larger values of λ_{adv}. Reassuringly, the adversarial penalty did not affect statistical calibration of LIVI (int) under the null (**Fig.** 4.2b).

4.3.4 Power to detect discrete *vs.* continuous effects

I compared LIVI and the baseline approach (VAE + LM) when varying η, the strength of intra- vs. inter-celltype variation (**Fig.** 4.2c). The VAE + LM approach, like many other existing methods, relies on discretizing the data and performed poorly for small values of η. LIVI, on the other hand, yields consistent results both in the continuous and discrete setting.

4.3.5 Disentanglement of ground truth factors

As a second criterion, I considered the ability of LIVI to recover the true simulated latent space of cellular contexts compared with the baseline VAE. I computed the maximum average Pearson correlation between inferred and ground truth loading matrices across all possible pairings (Pearson MCC, **Fig.** 4.2d). For medium to large genetic effect sizes, LIVI significantly improved the identification of the ground truth factors.

4.4 Empirical evaluation on a real dataset of one million PBMCs

Next, we applied LIVI on a published dataset of more than one million peripheral blood mononuclear cells (PBMCs) from approx 1 thousand donors [86]. We tested a set of 2,910 common (MAF > 0.05) biallelic SNPs for effects on the LIVI donor embeddings, approximately half of which were selected *trans* eQTL variants from a large bulk RNA-seq meta-analysis study [50]. We identified 622 persistent and two interaction effects at at FDR 5%. The two interaction effects were between variants *rs8026803*, *rs2562754* and factor

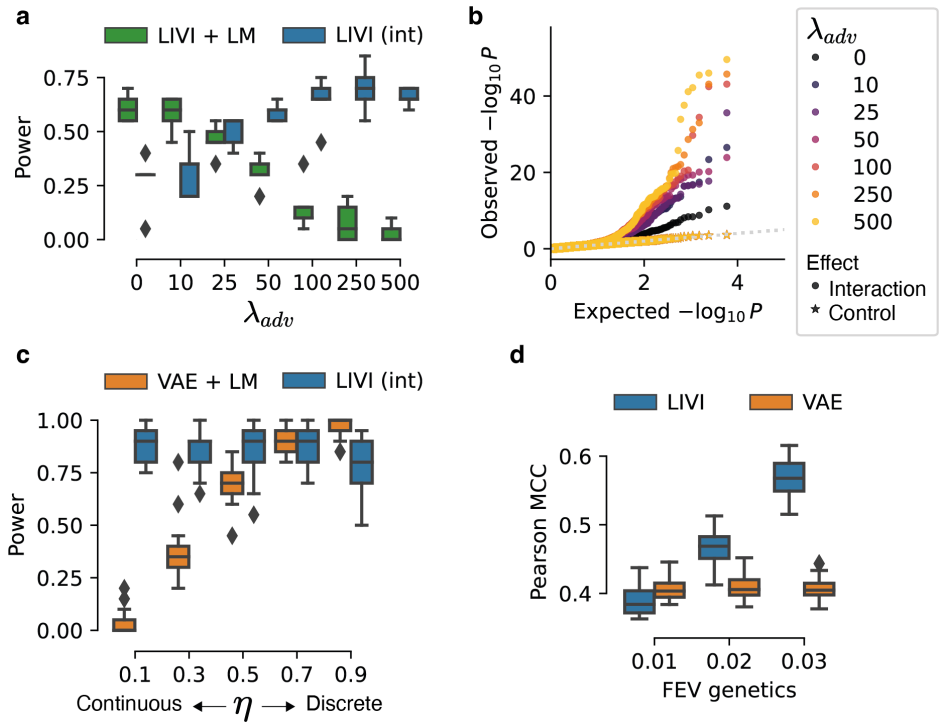

Figure 4.2. Evaluation on simulated data. (a) Power to detect interaction effects when varying the weight of the adversarial loss; LIVI interaction test (LIVI (int)) *vs.* clustering of LIVI base cell states followed by linear association testing per cluster (LIVI + LM). (b) LIVI (int) P-values when testing variants with simulated factor interactions and no effect (control). (c) Power to detect interaction effects when varying the extent of discrete *vs.* continuous simulated cell state variation; LIVI (int) *vs.* cluster-wise test based on VAE latent space. (d) Average Pearson correlation between inferred and ground truth factor loadings (maximum across all possible pairings).

30, which characterizes monocytes (**Fig.** 4.3). Notably, those variants have been associated to phenotypes such as monocyte count [240,241], lymphocyte-to-monocyte ratio [241], eosinophil count [240,241] and eosinophil percentage of granulocytes [242].

4.5 Discussion

Most existing studies for mapping genetic effects on quantitative traits focus on *cis* regulation. However, *cis*-eQTL play only a minor role in the overall heritability of gene expression, with most of it attributed to the cumulative impact of numerous weaker *trans*-regulatory ef-

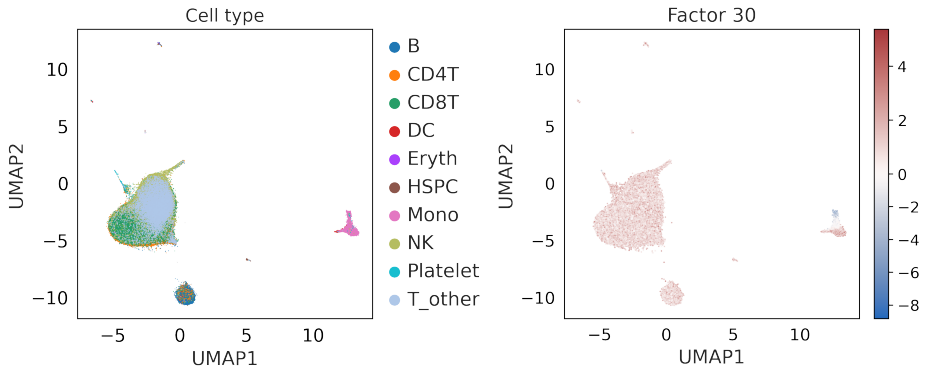

Figure 4.3. Interaction effects of variants *rs8026803*, *rs2562754* and factor 30, which characterizes monocytes (**Fig.** 4.3). Notably, those variants have been associated to phenotypes such as monocyte count [240, 241], lymphocyte-to-monocyte ratio [241], eosinophil count [240, 241] and eosinophil percentage of granulocytes [242]. Application of LIVI on real PBMC data. Uniform Manifold Approximation and Projection (UMAP) embedding of LIVI's donor-free canonical latent factors, colored by cell type (left) and factor 30 (right), which is associated with a known GWAS variants for monocyte count. Credit: Danai Vagiaki.

fects [243]. At the same time, the identification of trans-eQTL remains challenging due to the small effect sizes and the vast search space of possible variant-gene combinations. These issues can be partially addressed by mapping strategies using latent factor models, which share information across many correlated genes to improve statistical power [234–236].

Here, I described LIVI, an interpretable latent variable model for populations-scale single-cell RNA-sequencing data based on the VAE framework. A core innovations of LIVI is an interpretable and efficient approach to decompose the latent space into donor-invariant components, corresponding to canonical expression variation, as well as persistent and context-specific donor effects. A second innovation is a fast testing procedure to identify individual genetic variables or other donor covariates that are associated with the estimated persistent or context-specific latent space effects.

I evaluated the model on simulated data, finding that adversarial regularization leads to improved statistical power for downstream association tests. I showed that the model retains detection power regardless of whether cell states are continuous or discrete. Additionally, by leveraging donor labels, LIVI is able to more accurately infer the underlying data-generating

factors than a conventional VAE. In an example application to scRNA-seq of human peripheral blood mononuclear cells [86], the model identified hundreds of persistent genetic effects on expression factors, but only few context-specific effects. The latter were specific to the monocyte population and involved genetic variants that had previously been associated with monocyte count.

Chapter 5

Summary & concluding remarks

Historically, the definition of cell identity has relied on qualitative assessments of morphology, ontogeny (developmental origin), and physiological function. For example, a neuron may be identified by its unique morphology with axons and dendrites, its location in the nervous system, its embryonic origin from neural precursor cells, and its functional interactions with other neurons and cells in neural circuits. Modern sequencing-based assays such as RNA-seq have enabled the quantitative analysis of genome-wide molecular profiles underlying various cell types and states. Cells may now be identified based on specific patterns of gene expression, shaped by their genetic markup, epigenetic modifications and interactions within the cellular microenvironment. Initially, these approaches relied on the bulk analysis of tissue samples or pooled cell populations, which provides average signals of molecular traits across hundreds or millions of cells. More recently, single-cell sequencing technologies have transformed the analysis of cell identities, allowing to uncover cellular heterogeneity obscured in bulk data such as rare or unknown cell types as well as transitional or intermediate states.

Since single-cell sequencing was recognized as the 'Method of the Year 2013' by the editors of Nature Methods [244], the technology has advanced significantly, resulting in an exponential growth in the number of cells that can be profiled in a single experiment [95] and considerable reductions in sequencing cost. Improvements in droplet microfluidics, which allows for the delineation of cells in picoliter droplets, and combinatorial indexing strategies for cell barcoding have enabled the generation of datasets comprising hundreds of thousands to millions of cells [245–248]. These developments have now also made it possible to se-

quence samples at the population level, and generate data large enough study the subtle effects of genetic mutations on molecular traits at the level of single cells.

Genome-wide association studies (GWAS) have identified numerous genetic variants linked to complex diseases and traits, but understanding their functional consequences remains a challenge, especially for variants in non-coding sections of the DNA. Expression quantitative trait locus (eQTL) mapping can establish connections between genetic variants, like SNPs, and RNA levels, aiding in the identification of potential target genes and mechanisms of disease-associated variations. To be effective, however, eQTL mapping requires assaying changes in RNA expression in cell types and conditions related to the disease of interest. Most existing eQTL analyses rely on bulk sequencing data, which limits their ability to map genetic effects in specific cell types and states. The emerging field of single-cell genetics integrates single-cell sequencing data for molecular phenotyping with genotyping information, allowing for the unbiased analysis of context-specific genetic effects on molecular traits. Despite the potential of this approach, existing analysis strategies do not model genetic effects at the single-cell level, but apply methods for bulk data to pseudo-bulk aggregates of discretized single-cell profiles.

This thesis focused on the development of statistical and computational methods for modeling the relationship between genetic variation and quantitative traits at the level of individual cells. These methods combine latent variable models for cell-state inference with statistical hypothesis tests based on the linear mixed model framework. Key elements of this methodology have been reviewed in chapter 2.

In chapter 3, I proposed a mixed-effect approach to modeling context-specific effects from single-cell data. Briefly, a cell-state covariance matrix is defined based on a lower dimensional representation of the cell-state manifold, e.g., estimated using established methods for dimensionality reduction or trajectory inference. The presence or absence of context-specific effects can then be assessed using a single variance component parameter. I introduced the scDALI model, a generalized linear mixed model with Beta-Binomial likelihood for the analysis of allelic imbalance in single-cell count data. A score-based variance component test and inference scheme were derived using a penalized quasi-likelihood approximation. I validated the model on simulated data and described the application of scDALI to scATAC-seq data in *Drosophila Melanogaster* embryos to study allele-specific changes in chromatin accessibility during development. The chapter also presented a novel variant of the Variational

Autoencoder tailored to the analysis of temporally resolved scATAC-seq data, allowing to order cells along the underlying developmental trajectory.

Next, I introduced CellRegMap, a statistical model building on the same core concept as scDALI designed to model expression quantitative trait loci (eQTL) at single-cell resolution. The method uses the linear mixed model framework to decompose gene expression variation into genetic, cell state, and other factors. The model incorporates an interaction term of the cell state and genotype, allowing to capture context-specific effects. Unlike existing related methods for modeling genotype-environment interactions, CellRegMap also accounts for the possible confounding influence of sampling structure and donor kinship using additional random effect components. The model was evaluated using a semi-synthetic simulation framework, leveraging real genotyping data and expression profiles to incorporate possible sources of variation present in actual single-cell data. These experiments highlighted the importance of accounting for context-specific donor effects on gene expression in order to obtain calibrated test statistics.

The application of scDALI and CellRegMap demonstrated how the random effect framework can be used to detect genetic effects that manifest only in specific cell types and lineages or as a function of continuous processes such as cellular differentiation or development. Notably, both models identified associations that would have been masked in a (pseudo-) bulk analysis of discretized cell states. Furthermore, they revealed subtle context-specific genetic effects beyond the dominant axes of cell-state variation in the data. For example, in a recently published dataset of iPS cells differentiating towards the definite endoderm, CellRegMap identified genetic effects associated with specific stages of the cell cycle and cellular respiration that were missed in the primary analysis.

While scDALI and CellRegMap provide an extension of the state of the art in single-cell genetics, they are not without limitations and may be extended in a number of different directions. Recent work by Kumasaka et al. [249] has explored a non-linear variant of CellRegMap, based on the Gaussian process framework. Briefly, while CellRegMap models context-specific genetic effect sizes using a linear covariance function of the cell state factors, Kumasaka et al. apply a squared exponential kernel. This increase in modeling flexibility, however, further adds to the computational complexity. Kumasaka et al. consider a simplified random effect term to control for the confounding effects of donor relatedness, allowing for faster inference at the cost of a possible inflation of Type-I error rates. Fu-

ture work may focus on the computational aspect of mapping context-specific eQTL while maintaining calibrated test statistics.

scDALI and CellRegMap were developed as two independent models, leveraging either allele-specific information or population-scale data to map genetic effects. Future work may try to combine these ideas, to improve statistical power and link distal regulatory elements to gene promoters. Similar approaches have been used successfully for the analysis of bulk-sequencing data [40].

As a final contribution, I presented LIVI, a variant of the Variational autoencoder designed to disentangle donor-specific and shared gene expression variation in population-scale scRNA-seq data in chapter 4. LIVI uses an adversarial approach to remove donor-specific effects and reintroduces them in an interpretable linear interaction model. Donor effects are summarized in embedding vectors, that can be efficiently tested for associations with genetic variants or clinical covariates. In principle, this allows for the identifiation of eQTL variants that regulate groups of genes in *trans*. In particular, by testing gene expression factors rather than individual genes, this approach incurs a reduced multiple testing burden compared to standard *trans* eQTL mapping at the gene level. Using simulations I showed that adversarial regularization improves power to detect genetic effects in downstream tests as well as the identification of the true data-generating factors compared to a conventional VAE model. The model was applied to a scRNA-seq dataset of human peripheral blood mononuclear cells, uncovering hundreds of *trans*-regulatory genetic effects on expression factors. Future work may also focus on the application of LIVI to genetic perturbation screens [250–252], to model single-cell responses to gene knockouts or interferences.

Single-cell sequencing technologies continue to evolve, and the next generation of experimental assays will be cheaper [253] and able to capture longer transcripts with higher fidelity [254]. These developments will enable the creation of expansive atlases of context-specific *cis* and *trans* eQTLs in thousands of individuals from diverse populations. Meanwhile, the importance of robust statistical methods for single-cell genetics becomes even more evident, to unravel the relationship between genotype and phenotype at a genome-wide scale.

Appendix A

A.1 Generation and sequencing of *Drosophila Melanogaster* F1 embryos

The following description summarizes sample collection and library preparation as performed by Stefano Secchia, Bingqing Zhao and James P. Reddington. It is reproduced from the primary publication [185] and included here for completeness.

We generated Drosophila melanogaster F1 hybrids by crossing females from a common maternal virginizer line with males from four different inbred lines from the Drosophila melanogaster genetic reference panel [191, 255] (DGRP). Embryos were collected in 2 h windows (2–4 h, 6–8 h, and 10–12 h after egg laying) as previously described [191]. Hyperactive Tn5 transposase was purified by the EMBL Protein Expression and Purification facility as previously described [256] and stored at − 20 °C in storage buffer (25 mM Tris pH 7.5, 800 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 50% glycerol) until use. Uniquely indexed oligonucleotides from Cusanovich et al. [68] were annealed to common pMENTs oligos 95 °C 5 min, cooling to 65 °C (0.1 °C/s), 65 °C 5 min, cooling to 4 °C (0.1 °C/s)) to generate indexed transposons that were then loaded onto purified Tn5 by incubation at 23 °C with constant shaking at 350 rpm for 30 min. The loaded Tn5 transposomes were diluted 1:10 (final 0.02 mg/ml) in nuclease-free water and used immediately for tagmentation. Embryo dissociation and nuclear isolation were performed as described previously [68]. Nuclei were flash frozen in liquid nitrogen and stored at − 80 °C until use. Generation of sci-ATAC-seq libraries was performed largely as previously described [68] with minor modifications. The tagmentation reaction was performed by adding 2 μL of each of the 96 custom and uniquely

indexed Tn5 transposomes and by incubating at 55 °C for 1 h. After reverse-crosslinking, 5 µL of forward and reverse indexed primers (from Cusanovich et al. [68]), 7.5 µL KAPA HiFi DNA Polymerase ReadyMix (Roche) and 0.25 µL Bst3.0 (NEB) were added to each well. Tagmented DNA was then PCR amplified with the following cycling conditions: 72°C 5min, 98°C 30s; 98°C 10s, 63 °C 30 s, 19–22 cycles; 72 °C 1 min, hold at 10 °C. The optimal number of cycles for each library was determined beforehand by monitoring amplification on a qPCR machine for a set of test wells. Libraries were sequenced on an Illumina NextSeq 500 sequencer High Capacity 150 PE kit as previously described [68].

A.2 Extended calibration analysis for CellRegMap

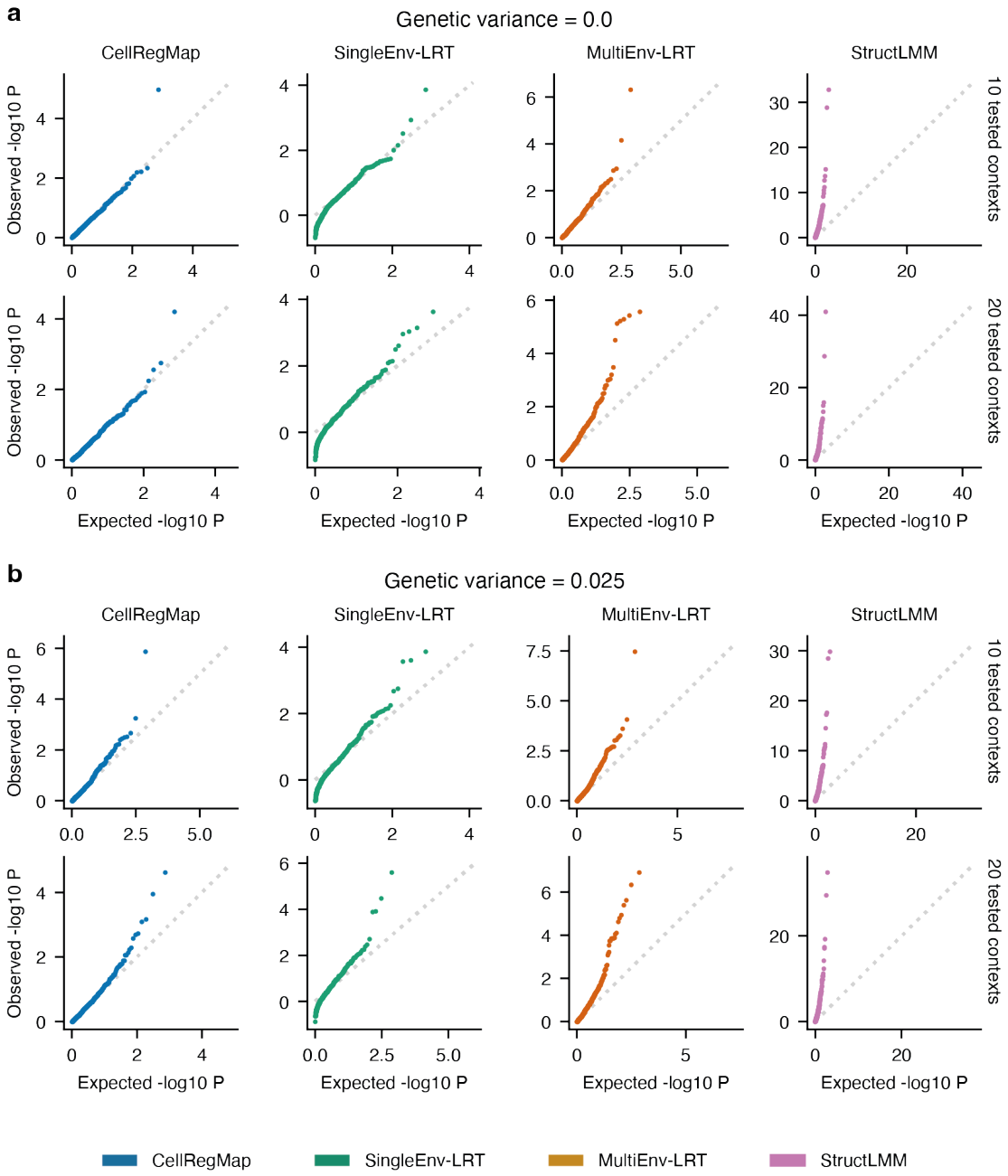

Figure A.1. Extended calibration analysis for CellRegMap. **(a,b)** QQ plots of expected versus observed negative log P-values, for different levels of persistent genetic variance (**(a)** vs. **(b)**). Rows represent different numbers of tested context variables (10 or 20). Shown are CellRegMap (blue), a fixed-effect likelihood-ratio-test for single contexts (SingleEnv-LRT; green), a multicontext fixed-effect test (MultiEnv-LRT; orange) and StructLMM (pink; [184]). The QQ plots for CellRegMap, SingleEnv-LRT and StructLMM from the first row in panel A (genetic variance = 0 and 10 contexts tested) are also shown in **Fig.** 3.25.

Bibliography

[1] Bruce Alberts. *Molecular biology of the cell*. Garland science, 2017.

[2] William S Bush and Jason H Moore. Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

[3] Robert J Klein, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, et al. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

[4] Keiko Yamazaki, Dermot McGovern, Jiannis Ragoussis, Marta Paolucci, Helen Butler, Derek Jewell, Lon Cardon, Masakazu Takazoe, Torao Tanaka, Toshiki Ichimori, et al. Single nucleotide polymorphisms in tnfsf15 confer susceptibility to crohn's disease. *Human molecular genetics*, 14(22):3499–3506, 2005.

[5] Richard H Duerr, Kent D Taylor, Steven R Brant, John D Rioux, Mark S Silverberg, Mark J Daly, A Hillary Steinhart, Clara Abraham, Miguel Regueiro, Anne Griffiths, et al. A genome-wide association study identifies il23r as an inflammatory bowel disease gene. *science*, 314(5804):1461–1463, 2006.

[6] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, 2007.

[7] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies

for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356–369, 2008.

[8] Evangelos Evangelou and John PA Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–389, 2013.

[9] Loic Yengo, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, Joel Hirschhorn, Jian Yang, Peter M Visscher, et al. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. *Human molecular genetics*, 27(20):3641–3649, 2018.

[10] 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.

[11] Yvonne G De Souza and John S Greenspan. Biobanking past, present and future: responsibilities and benefits. *AIDS (London, England)*, 27(3):303, 2013.

[12] Barbara E Stranger, Eli A Stahl, and Towfique Raj. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383, 2011.

[13] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[14] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2014.

[15] H Franklin Bunn. Pathogenesis and treatment of sickle cell disease. *New England Journal of Medicine*, 337(11):762–769, 1997.

[16] Nino Spataro, Juan Antonio Rodríguez, Arcadi Navarro, and Elena Bosch. Properties of human disease genes and the role of genes linked to mendelian disorders in complex disease aetiology. *Human molecular genetics*, 26(3):489–500, 2017.

[17] Alkes L Price, Chris CA Spencer, and Peter Donnelly. Progress and promise in understanding the genetic basis of common diseases. *Proceedings of the Royal Society B: Biological Sciences*, 282(1821):20151684, 2015.

[18] Feng Zhang and James R Lupski. Non-coding genetic variants in human disease. *Human molecular genetics*, 24(R1):R102–R110, 2015.

[19] US DOE Joint Genome Institute: Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4, RIKEN Genomic Sciences Center: Sakaki Yoshiyuki 9 Fujiyama Asao 9 Hattori Masahira 9 Yada Tetsushi 9 Toyoda Atsushi 9 Itoh Takehiko 9 Kawagoe Chiharu 9 Watanabe Hidemi 9 Totoki Yasushi 9 Taylor Todd 9, Genoscope, CNRS UMR-8030: Weissenbach Jean 10 Heilig Roland 10 Saurin William 10 Artiguenave Francois 10 Brottier Philippe 10 Bruls Thomas 10 Pelletier Eric 10 Robert Catherine 10 Wincker Patrick 10, Institute of Molecular Biotechnology: Rosenthal André 12 Platzer Matthias 12 Nyakatura Gerald 12 Taudien Stefan 12 Rump Andreas 12 Department of Genome Analysis, GTC Sequencing Center: Smith Douglas R. 11 Doucette-Stamm Lynn 11 Rubenfield Marc 11 Weinstock Keith 11 Lee Hong Mei 11 Dubois JoAnn 11, Beijing Genomics Institute/Human Genome Center: Yang Huanming 13 Yu Jun 13 Wang Jian 13 Huang Guyang 14 Gu Jun 15, et al. Initial sequencing and analysis of the human genome. *nature*, 409(6822):860–921, 2001.

[20] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, 2018.

[21] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[22] Yoav Gilad, Scott A Rifkin, and Jonathan K Pritchard. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends in genetics*, 24(8):408–415, 2008.

[23] Harm-Jan Westra and Lude Franke. From genome to function by studying eqtls. *Biochimica et Biophysica Acta (BBA)-molecular basis of Disease*, 1842(10):1896–1902, 2014.

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

[25] Helena Kilpinen, Angela Goncalves, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, Dalila Bensaddek, Francesco Paolo Casale, Oliver J Culley, et al. Common genetic variation drives molecular heterogeneity in human ipscs. *Nature*, 546(7658):370–375, 2017.

[26] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[27] GTEx Consortium Lead analysts: Aguet François 1 Brown Andrew A. 2 3 4 Castel Stephane E. 5 6 Davis Joe R. 7 8 He Yuan 9 Jo Brian 10 Mohammadi Pejman 5 6 Park YoSon 11 Parsana Princy 12 Segrè Ayellet V. 1 Strober Benjamin J. 9 Zappala Zachary 7 8, NIH program management: Addington Anjene 15 Guan Ping 16 Koester Susan 15 Little A. Roger 17 Lockhart Nicole C. 18 Moore Helen M. 16 Rao Abhi 16 Struewing Jeffery P. 19 Volpi Simona 19, Pathology: Sobin Leslie 30 Barcus Mary E. 30 Branton Philip A. 16, NIH Common Fund Nierras Concepcion R. 137, et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[28] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, et al. Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature*, 482(7385):390–394, 2012.

[29] Fabian Grubert, Judith B Zaugg, Maya Kasowski, Oana Ursu, Damek V Spacek, Alicia R Martin, Peyton Greenside, Rohith Srivas, Doug H Phanstiel, Aleksandra Pekowska, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*, 162(5):1051–1065, 2015.

[30] Kaur Alasoo, Julia Rodrigues, Subhankar Mukhopadhyay, Andrew J Knights, Alice L Mann, Kousik Kundu, HipSci Consortium, Christine Hale, Gordon Dougan, and Daniel J Gaffney. Shared genetic effects on chromatin and gene expression indi-

cate a role for enhancer priming in immune response. *Nature genetics*, 50(3):424–431, 2018.

[31] Mathieu Lemire, Syed HE Zaidi, Maria Ban, Bing Ge, Dylan Aïssi, Marine Germain, Irfahan Kassam, Mike Wang, Brent W Zanke, France Gagnon, et al. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nature communications*, 6(1):6326, 2015.

[32] Marc Jan Bonder, René Luijk, Daria V Zhernakova, Matthijs Moed, Patrick Deelen, Martijn Vermaat, Maarten Van Iterson, Freerk Van Dijk, Michiel Van Galen, Jan Bot, et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nature genetics*, 49(1):131–138, 2017.

[33] Jeremy Schwartzentruber, Stefanie Foskolou, Helena Kilpinen, Julia Rodrigues, Kaur Alasoo, Andrew J Knights, Minal Patel, Angela Goncalves, Rita Ferreira, Caroline Louise Benn, et al. Molecular and functional variation in ipsc-derived sensory neurons. *Nature genetics*, 50(1):54–61, 2018.

[34] 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.

[35] 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.

[36] Tomi Pastinen. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics*, 11(8):533–538, 2010.

[37] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, 2010.

[38] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis.

Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 464(7289):773–777, 2010.

[39] Helena Kilpinen, Sebastian M Waszak, Andreas R Gschwind, Sunil K Raghav, Robert M Witwicki, Andrea Orioli, Eugenia Migliavacca, Michaël Wiederkehr, Maria Gutierrez-Arcelus, Nikolaos I Panousis, et al. Coordinated effects of sequence variation on dna binding, chromatin structure, and transcription. *Science*, 342(6159):744–747, 2013.

[40] Natsuhiko Kumasaka, Andrew J Knights, and Daniel J Gaffney. Fine-mapping cellular qtls with rasqual and atac-seq. *Nature genetics*, 48(2):206–213, 2016.

[41] Paul R Burton, Martin D Tobin, and John L Hopper. Key concepts in genetic epidemiology. *The Lancet*, 366(9489):941–951, 2005.

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

[43] Nan M Laird and Christoph Lange. *The fundamentals of modern statistical genetics*. Springer, 2011.

[44] Qin Qin Huang, Scott C Ritchie, Marta Brozynska, and Michael Inouye. Power, false discovery rate and winner's curse in eqtl studies. *Nucleic acids research*, 46(22):e133–e133, 2018.

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

[46] Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, et al. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4):355–360, 2010.

[47] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to

account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[48] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature reviews genetics*, 11(7):459–463, 2010.

[49] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics*, 48(5):481–487, 2016.

[50] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Seyhan Yazar, et al. Large-scale cis-and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature genetics*, 53(9):1300–1310, 2021.

[51] Farhad Hormozdiari, Martijn Van De Bunt, Ayellet V Segre, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

[52] Eleonora Porcu, Sina Rüeger, Kaido Lepik, Federico A Santoni, Alexandre Reymond, and Zoltán Kutalik. Mendelian randomization integrating gwas and eqtl data reveals genetic determinants of complex and clinical traits. *Nature communications*, 10(1):3300, 2019.

[53] Matthew R Nelson, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, et al. The support of human genetic evidence for approved drug indications. *Nature genetics*, 47(8):856–860, 2015.

[54] Noah J Connally, Sumaiya Nazeen, Daniel Lee, Huwenbo Shi, John Stamatoyannopoulos, Sung Chun, Chris Cotsapas, Christopher A Cassa, and Shamil R Sunyaev. The missing link between genetic association and regulatory function. *Elife*, 11:e74970, 2022.

[55] Sung Chun, Alexandra Casparino, Nikolaos A Patsopoulos, Damien C Croteau-Chonka, Benjamin A Raby, Philip L De Jager, Shamil R Sunyaev, and Chris Cotsapas. Limited statistical evidence for shared genetic effects of eqtls and autoimmune-disease-associated loci in three major immune-cell types. *Nature genetics*, 49(4):600–605, 2017.

[56] Barbara E Stranger, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, Daphne Koller, et al. Population genomics of human gene expression. *Nature genetics*, 39(10):1217–1224, 2007.

[57] Amanda J Myers, J Raphael Gibbs, Jennifer A Webster, Kristen Rohrer, Alice Zhao, Lauren Marlowe, Mona Kaleem, Doris Leung, Leslie Bryden, Priti Nath, et al. A survey of genetic human cortical gene expression. *Nature genetics*, 39(12):1494–1499, 2007.

[58] Jingyuan Fu, Marcel GM Wolfs, Patrick Deelen, Harm-Jan Westra, Rudolf SN Fehrmann, Gerard J Te Meerman, Wim A Buurman, Sander SM Rensen, Harry JM Groen, Rinse K Weersma, et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS genetics*, 8(1):e1002431, 2012.

[59] Alexandra C Nica, Leopold Parts, Daniel Glass, James Nisbet, Amy Barrett, Magdalena Sekowska, Mary Travers, Simon Potter, Elin Grundberg, Kerrin Small, et al. The architecture of gene regulatory variation across multiple human tissues: the muther study. *PLoS genetics*, 7(2):e1002003, 2011.

[60] Antigone S Dimas, Samuel Deutsch, Barbara E Stranger, Stephen B Montgomery, Christelle Borel, Homa Attar-Cohen, Catherine Ingle, Claude Beazley, Maria Gutierrez Arcelus, Magdalena Sekowska, et al. Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science*, 325(5945):1246–1250, 2009.

[61] Barbara E Stranger, Stephen B Montgomery, Antigone S Dimas, Leopold Parts, Oliver Stegle, Catherine E Ingle, Magda Sekowska, George Davey Smith, David Evans, Maria Gutierrez-Arcelus, et al. Patterns of cis regulatory variation in diverse human populations. *PLoS genetics*, 8(4):e1002639, 2012.

[62] Eric E Schadt, Cliona Molony, Eugene Chudin, Ke Hao, Xia Yang, Pek Y Lum, Andrew Kasarskis, Bin Zhang, Susanna Wang, Christine Suver, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS biology*, 6(5):e107, 2008.

[63] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[64] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181, 2014.

[65] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[66] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[67] Jonathan A Griffiths, Antonio Scialdone, and John C Marioni. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular systems biology*, 14(4):e8046, 2018.

[68] Darren A Cusanovich, James P Reddington, David A Garfield, Riza M Daza, Delasa Aghamirzaie, Raquel Marco-Ferreres, Hannah A Pliner, Lena Christiansen, Xiaojie Qiu, Frank J Steemers, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, 555(7697):538–542, 2018.

[69] Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, 2018.

[70] Nicholas Schaum, Jim Karkanias, Norma F Neff, Andrew P May, Stephen R Quake, Tony Wyss-Coray, Spyros Darmanis, Joshua Batson, Olga Botvinnik, Michelle B Chen, et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris: The tabula muris consortium. *Nature*, 562(7727):367, 2018.

[71] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *elife*, 6:e27041, 2017.

[72] Laura González-Silva, Laura Quevedo, and Ignacio Varela. Tumor functional heterogeneity unraveled by scrna-seq technologies. *Trends in cancer*, 6(1):13–19, 2020.

[73] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

[74] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

[75] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19:1–16, 2018.

[76] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[77] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, et al. Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology*, 32(10):1053–1058, 2014.

[78] Anton JM Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, 2019.

[79] Anna SE Cuomo, Aparna Nathan, Soumya Raychaudhuri, Daniel G MacArthur, and Joseph E Powell. Single-cell genomics meets human genetics. *Nature Reviews Genetics*, pages 1–15, 2023.

[80] Quin F Wills, Kenneth J Livak, Alex J Tipping, Tariq Enver, Andrew J Goldson, Darren W Sexton, and Chris Holmes. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature biotechnology*, 31(8):748–752, 2013.

[81] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89–94, 2018.

[82] Yuanhua Huang, Davis J McCarthy, and Oliver Stegle. Vireo: Bayesian demultiplexing of pooled single-cell rna-seq data without genotype reference. *Genome biology*, 20:1–12, 2019.

[83] Monique GP Van Der Wijst, Harm Brugge, Dylan H De Vries, Patrick Deelen, Morris A Swertz, LifeLines Cohort Study, BIOS Consortium, and Lude Franke. Single-cell rna sequencing identifies celltype-specific cis-eqtls and co-expression qtls. *Nature genetics*, 50(4):493–497, 2018.

[84] Anna SE Cuomo, Daniel D Seaton, Davis J McCarthy, Iker Martinez, Marc Jan Bonder, Jose Garcia-Bernardo, Shradha Amatya, Pedro Madrigal, Abigail Isaacson, Florian Buettner, et al. Single-cell rna-sequencing of differentiating ips cells reveals dynamic genetic effects on gene expression. *Nature communications*, 11(1):810, 2020.

[85] Julie Jerber, Daniel D Seaton, Anna SE Cuomo, Natsuhiko Kumasaka, James Haldane, Juliette Steer, Minal Patel, Daniel Pearce, Malin Andersson, Marc Jan Bonder, et al. Population-scale single-cell rna-seq profiling across dopaminergic neuron differentiation. *Nature genetics*, 53(3):304–312, 2021.

[86] Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas RP Taylor, Linda Clarke, et al. Single-cell eqtl mapping identifies cell type–specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, 2022.

[87] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[88] Beate Vieth, Swati Parekh, Christoph Ziegenhain, Wolfgang Enard, and Ines Hellmann. A systematic evaluation of single cell rna-seq analysis pipelines. *Nature communications*, 10(1):4667, 2019.

[89] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.

[90] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643, 2017.

[91] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163–166, 2014.

[92] Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–640, 2014.

[93] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):296, 2019.

[94] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. Mapping the mouse cell atlas by microwell-seq. *Cell*, 172(5):1091–1107, 2018.

[95] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

[96] Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.

[97] Anna SE Cuomo, Giordano Alvari, Christina B Azodi, Davis J McCarthy, and Marc Jan Bonder. Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome biology*, 22(1):1–30, 2021.

[98] Paola Benaglio, Jacklyn Newsome, Jee Yun Han, Joshua Chiou, Anthony Aylward, Sierra Corban, Michael Miller, Mei-Lin Okino, Jaspreet Kaur, Sebastian Preissl, et al. Mapping genetic effects on cell type-specific chromatin accessibility and annotating complex immune trait variants using single nucleus atac-seq in peripheral blood. *PLoS genetics*, 19(6):e1010759, 2023.

[99] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[100] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[101] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[102] Jiming Jiang and Thuan Nguyen. *Linear and generalized linear mixed models and their applications*, volume 1. Springer, 2007.

[103] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

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

[105] Adolf Buse. The likelihood ratio, wald, and lagrange multiplier tests: An expository note. *The American Statistician*, 36(3a):153–157, 1982.

[106] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482, 1943.

[107] C Radhakrishna Rao. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44, pages 50–57. Cambridge University Press, 1948.

[108] Robert F Engle. Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, 2:775–826, 1984.

[109] Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[110] Xihong Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, 1997.

[111] Daowen Zhang and Xihong Lin. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4(1):57–74, 2003.

[112] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[113] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, David C Christiani, Mark M Wurfel, and Xihong Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.

[114] Robert B Davies. Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics*, pages 323–333, 1980.

[115] Huan Liu, Yongqiang Tang, and Hao Helen Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4):853–856, 2009.

[116] Emrah Kostem and Eleazar Eskin. Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *The American Journal of Human Genetics*, 92(4):558–564, 2013.

[117] Shayle R Searle. Matrix algebra useful for statistics. Technical report, 1982.

[118] Francis J Anscombe. The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254, 1948.

[119] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243, 1964.

[120] Peter McCullagh. *Generalized linear models*. Chapman and Hall, 1989.

[121] Kendall Atkinson. *An introduction to numerical analysis*. John wiley & sons, 1991.

[122] Charles E McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170, 1997.

[123] John T Ormerod and Matt P Wand. Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 21(1):2–17, 2012.

[124] Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.

[125] Peter J Green. Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, pages 245–259, 1987.

[126] Xihong Lin and Norman E Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91(435):1007–1016, 1996.

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

[128] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500–507, 2012.

[129] Florian Buettner, Naruemon Pratanwanich, Davis J McCarthy, John C Marioni, and Oliver Stegle. f-sclvm: scalable and versatile factor analysis for single-cell rna-seq. *Genome biology*, 18:1–13, 2017.

[130] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124, 2018.

[131] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[132] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[133] Benjamin F Voight and Jonathan K Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLoS genetics*, 1(3):e32, 2005.

[134] Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, Luigi L Cavalli-Sforza, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *science*, 319(5866):1100–1104, 2008.

[135] Alkes L Price, Johannah Butler, Nick Patterson, Cristian Capelli, Vincenzo L Pascali, Francesca Scarnicci, Andres Ruiz-Linares, Leif Groop, Angelica A Saetta, Penelope Korkolopoulou, et al. Discerning the ancestry of european americans in genetic association studies. *PLoS genetics*, 4(1):e236, 2008.

[136] John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008.

[137] Ben John Hayes, Peter M Visscher, and Michael E Goddard. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research*, 91(1):47–60, 2009.

[138] Gabriel E Hoffman. Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PloS one*, 8(10):e75707, 2013.

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

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

[141] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[142] Jelle J Goeman and Aldo Solari. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978, 2014.

[143] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[144] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[145] John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012.

[146] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[147] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational*

StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers, pages 177–186. Springer, 2010.

[148] Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 2002.

[149] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[150] Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

[151] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.

[152] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[153] Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.

[154] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[155] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 835–851. Springer, 2016.

[156] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[157] Artidoro Pagnoni, Kevin Liu, and Shangyan Li. Conditional variational autoencoder for neural machine translation. *arXiv preprint arXiv:1812.04405*, 2018.

[158] Jaechang Lim, Seongok Ryu, Jin Woo Kim, and Woo Youn Kim. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of cheminformatics*, 10(1):1–9, 2018.

[159] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[160] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[161] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

[162] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

[163] Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning*, pages 50–59. PMLR, 2018.

[164] Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning hierarchical compositional visual concepts. *arXiv preprint arXiv:1707.03389*, 2017.

[165] Sjoerd Van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? *Advances in neural information processing systems*, 32, 2019.

[166] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

[167] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[168] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[169] Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune Hannes Pers, and Ole Winther. scvae: Variational auto-encoders for single-cell gene expression data. biorxiv. 2018.

[170] Jiarui Ding, Anne Condon, and Sohrab P Shah. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature communications*, 9(1):2002, 2018.

[171] Dongfang Wang and Jin Gu. Vasc: dimension reduction and visualization of single-cell rna-seq data by deep variational autoencoder. *Genomics, proteomics & bioinformatics*, 16(5):320–331, 2018.

[172] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

[173] Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell rna-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, 2020.

[174] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

[175] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[176] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

[177] Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

[178] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349, 2008.

[179] Valentine Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

[180] Xin Li, Yungil Kim, Emily K Tsang, Joe R Davis, Farhan N Damani, Colby Chiang, Gaelen T Hess, Zachary Zappala, Benjamin J Strober, Alexandra J Scott, et al. The impact of rare variation on gene expression across tissues. *Nature*, 550(7675):239–243, 2017.

[181] Nicole M Ferraro, Benjamin J Strober, Jonah Einson, Nathan S Abell, Francois Aguet, Alvaro N Barbeira, Margot Brandt, Maja Bucan, Stephane E Castel, Joe R Davis, et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science*, 369(6509):eaaz5900, 2020.

[182] GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[183] Enrico Cannavò, Nils Koelling, Dermot Harnett, David Garfield, Francesco P Casale, Lucia Ciglar, Hilary E Gustafson, Rebecca R Viales, Raquel Marco-Ferreres, Jacob F Degner, et al. Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature*, 541(7637):402–406, 2017.

[184] Rachel Moore, Francesco Paolo Casale, Marc Jan Bonder, Danilo Horta, Lude Franke, Inês Barroso, and Oliver Stegle. A linear mixed-model approach to study multivariate gene–environment interactions. *Nature genetics*, 51(1):180–186, 2019.

[185] Tobias Heinen, Stefano Secchia, James P Reddington, Bingqing Zhao, Eileen EM Furlong, and Oliver Stegle. Scdali: modeling allelic heterogeneity in single cells

reveals context-specific genetic regulation. *Genome biology*, 23(1):1–24, 2022.

[186] Anna SE Cuomo, Tobias Heinen, Danai Vagiaki, Danilo Horta, John C Marioni, and Oliver Stegle. Cellregmap: a statistical framework for mapping context-specific regulatory variants using scrna-seq. *Molecular Systems Biology*, 18(8):e10663, 2022.

[187] David A Knowles, Joe R Davis, Hilary Edgington, Anil Raj, Marie-Julie Favé, Xiaowei Zhu, James B Potash, Myrna M Weissman, Jianxin Shi, Douglas F Levinson, et al. Allele-specific expression reveals interactions between genetic variation and environment. *Nature methods*, 14(7):699–702, 2017.

[188] Bryce Van De Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061–1063, 2015.

[189] Pejman Mohammadi, Stephane E Castel, Andrew A Brown, and Tuuli Lappalainen. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome research*, 27(11):1872–1884, 2017.

[190] Jiaxin Fan, Xuran Wang, Rui Xiao, and Mingyao Li. Detecting cell-type-specific allelic expression imbalance by integrative analysis of bulk and single-cell rna sequencing data. *PLoS Genetics*, 17(3):e1009080, 2021.

[191] Swann Floc'hlay, Emily S Wong, Bingqing Zhao, Rebecca R Viales, Morgane Thomas-Chollier, Denis Thieffry, David A Garfield, and Eileen EM Furlong. Cis-acting variation is common across regulatory layers but is often buffered during embryonic development. *Genome research*, 31(2):211–224, 2021.

[192] Wei Sun. A statistical framework for eqtl mapping using rna-seq data. *Biometrics*, 68(1):1–11, 2012.

[193] Edward Tunnacliffe and Jonathan R Chubb. What is a transcriptional burst? *Trends in Genetics*, 36(4):288–297, 2020.

[194] Thomas Minka. Estimating a dirichlet distribution, 2000.

[195] Huidong Chen, Caleb Lareau, Tommaso Andreani, Michael E Vinyard, Sara P Garcia, Kendell Clement, Miguel A Andrade-Navarro, Jason D Buenrostro, and Luca Pinello.

Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome biology*, 20(1):1–25, 2019.

[196] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[197] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[198] James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR, 2015.

[199] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009.

[200] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

[201] Alexander G de G Matthews, Mark Van Der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. Gpflow: A gaussian process library using tensorflow. *J. Mach. Learn. Res.*, 18(40):1–6, 2017.

[202] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018.

[203] Pierre Duchesne and Pierre Lafaye De Micheaux. Computing the distribution of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4):858–862, 2010.

[204] Feng Yan, David R Powell, David J Curtis, and Nicholas C Wong. From reads to insight: a hitchhiker's guide to atac-seq data analysis. *Genome biology*, 21:1–16, 2020.

[205] 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.

[206] 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.

[207] Wei Chu, Zoubin Ghahramani, and Christopher KI Williams. Gaussian processes for ordinal regression. *Journal of machine learning research*, 6(7), 2005.

[208] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[209] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[210] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

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

[212] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.

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

[214] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.

[215] Vasilis Ntranos, Lynn Yi, Páll Melsted, and Lior Pachter. A discriminative learning approach to differential expression analysis for single-cell rna-seq. *Nature methods*, 16(2):163–166, 2019.

[216] 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.

[217] James P Reddington, David A Garfield, Olga M Sigalova, Aslihan Karabacak Calviello, Raquel Marco-Ferreres, Charles Girardot, Rebecca R Viales, Jacob F Degner, Uwe Ohler, and Eileen EM Furlong. Lineage-resolved enhancer and promoter usage during a time course of embryogenesis. *Developmental cell*, 55(5):648–664, 2020.

[218] Nammi Park, Jae Cheal Yoo, Jiwon Ryu, Seong-Geun Hong, Eun Mi Hwang, and Jae-Yong Park. Copine1 enhances neuronal differentiation of the hippocampal progenitor hib5 cells. *Molecules and cells*, 34:549–554, 2012.

[219] Silvia Domcke, Andrew J Hill, Riza M Daza, Junyue Cao, Diana R O'Day, Hannah A Pliner, Kimberly A Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H Milbank, et al. A human cell atlas of fetal chromatin accessibility. *Science*, 370(6518):eaba7612, 2020.

[220] 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.

[221] Drew Neavin, Quan Nguyen, Maciej S Daniszewski, Helena H Liang, Han Sheng Chiu, Yong Kiat Wee, Anne Senabouth, Samuel W Lukowski, Duncan E Crombie, Grace E Lidgerwood, et al. Single cell eqtl analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome biology*, 22(1):1–19, 2021.

[222] Jun Ding, Johann E Gudjonsson, Liming Liang, Philip E Stuart, Yun Li, Wei Chen, Michael Weichenthal, Eva Ellinghaus, Andre Franke, William Cookson, et al. Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals exten-

sive overlap in cis-eqtl signals. *The American Journal of Human Genetics*, 87(6):779–789, 2010.

[223] Enrico Petretto, Leonardo Bottolo, Sarah R Langley, Matthias Heinig, Chris McDermott-Roe, Rizwan Sarwar, Michal Pravenec, Norbert Hübner, Timothy J Aitman, Stuart A Cook, et al. New insights into the genetic control of gene expression using a bayesian multi-tissue approach. *PLoS computational biology*, 6(4):e1000737, 2010.

[224] Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eqtl analysis in multiple tissues. *PLoS genetics*, 9(5):e1003486, 2013.

[225] Jae Hoon Sul, Buhm Han, Chun Ye, Ted Choi, and Eleazar Eskin. Effectively identifying eqtls from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS genetics*, 9(6):e1003491, 2013.

[226] Antonio Fabio Di Narzo, Haoxiang Cheng, Jianwei Lu, and Ke Hao. Meta-eqtl: a tool set for flexible eqtl meta-analysis. *BMC bioinformatics*, 15(1):1–5, 2014.

[227] Gen Li, Andrey A Shabalin, Ivan Rusyn, Fred A Wright, and Andrew B Nobel. An empirical bayes approach for multiple tissue eqtl analysis. *Biostatistics*, 19(3):391–406, 2018.

[228] Sarah M Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature genetics*, 51(1):187–195, 2019.

[229] Daria V Zhernakova, Patrick Deelen, Martijn Vermaat, Maarten Van Iterson, Michiel Van Galen, Wibowo Arindrarto, Peter Van't Hof, Hailiang Mei, Freerk Van Dijk, Harm-Jan Westra, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nature genetics*, 49(1):139–145, 2017.

[230] David Heckerman, Deepti Gurdasani, Carl Kadie, Cristina Pomilla, Tommy Carstensen, Hilary Martin, Kenneth Ekoru, Rebecca N Nsubuga, Gerald Ssenyomo, Anatoli Kamali, et al. Linear mixed model for heritability estimation that explicitly

addresses environmental variation. *Proceedings of the National Academy of Sciences*, 113(27):7377–7382, 2016.

[231] Armin P Schoech, Daniel M Jordan, Po-Ru Loh, Steven Gazal, Luke J O'Connor, Daniel J Balick, Pier F Palamara, Hilary K Finucane, Shamil R Sunyaev, and Alkes L Price. Quantification of frequency-dependent genetic architectures in 25 uk biobank traits reveals action of negative selection. *Nature communications*, 10(1):790, 2019.

[232] Aaron T L Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, 17(1):1–14, 2016.

[233] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):284, 2018.

[234] Victoria Hore, Ana Vinuela, Alfonso Buil, Julian Knight, Mark I McCarthy, Kerrin Small, and Jonathan Marchini. Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094–1100, 2016.

[235] Ariel DH Gewirtz, F William Townes, and Barbara E Engelhardt. Telescoping bimodal latent dirichlet allocation to identify expression qtls across tissues. *Life Science Alliance*, 5(12), 2022.

[236] Benjamin J Strober, Karl Tayeb, Joshua Popp, Guanghao Qi, Mary Grace Gordon, Richard Perez, Chun Jimmie Ye, and Alexis Battle. Uncovering context-specific genetic-regulation of gene expression from single-cell rna-sequencing using latent-factor models. *bioRxiv*, pages 2022–12, 2022.

[237] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[238] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L Ibarra, F Alexander Wolf, Nafissa Yakubova, Fabian J Theis, and David Lopez-Paz. Compositional perturbation autoencoder for single-cell response modeling. *BioRxiv*, 2021.

[239] Pierre Boyeau, Justin Hong, Adam Gayoso, Michael Jordan, Elham Azizi, and Nir Yosef. Deep generative modeling for quantifying sample-level heterogeneity in single-cell omics. *BioRxiv*, pages 2022–10, 2022.

[240] Dragana Vuckovic, Erik L Bao, Parsa Akbari, Caleb A Lareau, Abdou Mousas, Tao Jiang, Ming-Huei Chen, Laura M Raffield, Manuel Tardaguila, Jennifer E Huffman, et al. The polygenic and monogenic basis of blood traits and diseases. *Cell*, 182(5):1214–1231, 2020.

[241] Linda Kachuri, Soyoung Jeon, Andrew T DeWan, Catherine Metayer, Xiaomei Ma, John S Witte, Charleston WK Chiang, Joseph L Wiemels, and Adam J de Smith. Genetic determinants of blood-cell traits influence susceptibility to childhood acute lymphoblastic leukemia. *The American Journal of Human Genetics*, 108(10):1823–1835, 2021.

[242] William J Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A Kostadima, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429, 2016.

[243] Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS genetics*, 7(2):e1001317, 2011.

[244] Method of the year 2013. *Nature Methods*, 11(1):1–1, December 2013.

[245] Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nature protocols*, 12(1):44–73, 2017.

[246] Sarah A Vitak, Kristof A Torkenczy, Jimi L Rosenkrantz, Andrew J Fields, Lena Christiansen, Melissa H Wong, Lucia Carbone, Frank J Steemers, and Andrew Adey. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature methods*, 14(3):302–308, 2017.

[247] Tao Luo, Lei Fan, Rong Zhu, and Dong Sun. Microfluidic single-cell manipulation and analysis: Methods and applications. *Micromachines*, 10(2):104, 2019.

[248] Dan Gao, Feng Jin, Min Zhou, and Yuyang Jiang. Recent advances in single cell manipulation and biochemical analysis on microfluidics. *Analyst*, 144(3):766–781, 2019.

[249] Natsuhiko Kumasaka, Raghd Rostom, Ni Huang, Krzysztof Polanski, Kerstin B Meyer, Sharad Patel, Rachel Boyd, Celine Gomez, Sam N Barnett, Nikolaos I Panousis, et al. Mapping interindividual dynamics of innate immune response at single-cell resolution. *Nature Genetics*, 55(6):1066–1075, 2023.

[250] Paul Datlinger, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. Pooled crispr screening with single-cell transcriptome readout. *Nature methods*, 14(3):297–301, 2017.

[251] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.

[252] Joseph M Replogle, Thomas M Norman, Albert Xu, Jeffrey A Hussmann, Jin Chen, J Zachery Cogan, Elliott J Meer, Jessica M Terry, Daniel P Riordan, Niranjan Srinivas, et al. Combinatorial single-cell crispr screens by direct guide rna capture and targeted sequencing. *Nature biotechnology*, 38(8):954–961, 2020.

[253] Sean K Simmons, Gila Lithwick-Yanai, Xian Adiconis, Florian Oberstrass, Nika Iremadze, Kathryn Geiger-Schuller, Pratiksha I Thakore, Chris J Frangieh, Omer Barad, Gilad Almogy, et al. Mostly natural sequencing-by-synthesis for scrna-seq using ultima sequencing. *Nature Biotechnology*, 41(2):204–211, 2023.

[254] Martin Philpott, Jonathan Watson, Anjan Thakurta, Tom Brown Jr, Tom Brown Sr, Udo Oppermann, and Adam P Cribbs. Nanopore sequencing of single-cell transcriptomes with sccolor-seq. *Nature biotechnology*, 39(12):1517–1520, 2021.

[255] Trudy FC Mackay, Stephen Richards, Eric A Stone, Antonio Barbadilla, Julien F Ayroles, Dianhui Zhu, Sònia Casillas, Yi Han, Michael M Magwire, Julie M Cridland, et al. The drosophila melanogaster genetic reference panel. *Nature*, 482(7384):173–178, 2012.

[256] Matthew J Rossi, William KM Lai, and B Franklin Pugh. Simplified chip-exo assays. *Nature communications*, 9(1):2842, 2018.