TUM SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY

TECHNISCHE UNIVERSITÄT MÜNCHEN

Dissertation

Computational methods for design and analysis of population-based multiomics studies

Katharina Theresia Schmid

November 2022

HELMHOLTZ MUNICH

TUM SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY

TECHNISCHE UNIVERSITÄT MÜNCHEN

Computational methods for design and analysis of population-based multiomics studies

Katharina Theresia Schmid

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:	Prof. Dr. Stephan Günnemann
Prüfer der Dissertation:	1. TUM Junior Fellow Dr. Matthias Heinig
	2. Prof. Dr. Julien Gagneur
	3. Prof. Dr. Michael Love

Die Dissertation wurde am 16.11.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 27.11.2023 angenommen.

Acknowledgments

This thesis would not have been possible without the support of an awesome group of colleagues and friends. First of all, I have to thank my supervisor, Matthias Heinig, for his support, all the opportunities I got and the many new things I could learn. Matthias was always full of enthusiasm for science and a never-ending source of new ideas.

All research projects were the result of great collaborations with different labs. Especially thanks goes here to John Chambers, Marie Loh, Melanie Waldenberger, Christian Gieger and Rory Wilson for our collaboration in the meQTL and eQTM project. Elisabeth Binder and Fabian Theis helped not only to developed scPower, but gave valuable feedback for all my projects as my Thesis Advisory Committee. Lude Franke and his group welcomed me warmly at the UMCG in Groningen during my lab exchange and together, we set up an exciting project about co-eQTLs. Thank you, Shuang, Maryna, Dylan, Roy and all the other members at UMCG who enabled me a great research stay in those difficult times.

I am very glad that I could share all the ups and downs in the last years with the awesome members of the Heinig lab. Nothing helps more on a frustrating coding day than a (virtual) coffee break with you. Thank you, Hans, Ines and Toray, for sharing all your PhD experiences. With your help, I became a better coder and scientist and survived the bureaucratic jungle. Thank you, Barbara, for rescuing both scPower and many lab activities. Thank you, Simon, for all the conversations about science and life. Thank you, Corinna, for all your energy during our teaching sessions and the trips to the OSH round-about. Thank you, Annie, Sergey and Florin, for all the fun activities together.

Of course, thanks also to my second unofficial lab - Akshaya, Anna, Johanna, Maria (R) and the ducks - for all the distractions and all the craziness. And thanks to Maria (CT) and my new lab for giving me the time and support to finish this thesis. Thanks to all the other members of the ICB, especially the ICB office, for all the support and the great time together.

Last, but not least, I want to thank my friends and family. For Anne, who kept the close contact despite the long distance. For my parents, Richard and Gisela, and my brother Valentin, who always supported me, long before this PhD journey started. They had to endure a lot of boring scientific stories in the last years and only rarely fell asleep. And Jay, who was always at my side and cheered me up - after coding bugs and bad reviewer comments - the thesis would have never been finished without you.

Abstract

An interplay between genetic and environmental factors cause many common complex diseases, such as asthma, diabetes or depression. While more and more of these associations are identified in large human cohort studies, the molecular consequences of genetic and environmental factors that lead to the diseases are often still poorly understood. To explore the genetic consequences, association studies between genetic variants and gene expression, called expression quantitative trait loci (eQTLs), could successfully uncover part of the downstream genetic effects. The consequences of environmental influence on cellular level can be captured by measuring epigenetic factors, which regulate gene activation and are strongly influenced by environmental stimuli. Also here, association analyses of epigenetic factors with other omics levels, such as the transcriptome and the genome, help to uncover interactions and downstream effects.

Nevertheless, only a small part of the phenotypic effects of genetic and environmental factors could be explained on molecular level yet. At least part of this knowledge gap is caused by the cell type specificity of molecular QTLs. Many associations are likely not identified yet without a proper analysis in the relevant cell types. Until recently, most analyses were performed with bulk tissue datasets, which capture only the average omics levels across all measured cell types. For this reason, either specialized methods for cell type specific analysis with bulk datasets are necessary or the use of single cell datasets to increase the number of identified associations and improve their interpretability. Both strategies are explored in this thesis. It covers in total three different projects, in which current association strategies are applied and refined on both bulk and single cell datasets with different omics.

In the first project, we analyzed DNA methylation, an important epigenetic mark, and its relationship with genetics and gene expression to better understand its role in gene regulation. We identified 11,165,559 methylation quantitative trait loci (meQTLs), i.e. genetic variants which influence DNA methylation, in a large multi-ethnic bulk cohort. To capture the cell type and context specificity of meQTLs with this bulk dataset, we mapped interaction meQTLs (iQTLs), meQTLs whose effect sizes are influenced by the cell type composition of the donors or environmental factors. Following this, we investigated the influence of DNA methylation on gene expression via expression quantitative trait methylations (eQTMs), again with specific focus on the cell types. The number of eQTMs reduced drastically from 54,898,225 to 98,050 after correction for cell type composition. To better understand which genes are affected by DNA methylation in which genomic regions, we analyzed the genomic context for both types of eQTMs using a combination of machine learning approaches and identified clear differences:

eQTMs before cell type correction were connected with DNA methylation in enhancer regions, cell type corrected eQTMs with DNA methylation in promoter regions.

In the second project, we explored the use of single cell transcriptomics, a recent technical development, for cell type specific eQTL analyses. However, only few single cell cohorts currently exist. Sophisticated experimental design of future cohorts is crucial to perform powerful studies without wasting resources. For this reason, we developed of a statistical power analysis framework for multi-sample single cell transcriptomics studies called *scPower*. In contrast to other existing tools, it takes single-cell specific characteristics into account, while being very time and memory efficient. It is easily applicable to different use cases because of cell type specific priors. The efficient calculation allows optimization of experimental parameters given certain cost restrictions. Our evaluation showed that in a large range of settings, microfluidics-based single cell technologies, such as 10X Genomics and Drop-Seq, in combination with shallow sequencing of many cells gave the best power.

On top of that, single cell transcriptomics studies allow new analysis approaches compared to classical eQTL studies, which identify only the target genes affected by genetic variants. The multiple measurement points per individual enable the reconstruction of personalized gene regulatory networks and so also the identification of the upstream regulatory processes disturbed by the eQTL variants. For this, we mapped co-expression QTLs (co-eQTLs), genetic variants that change the co-expression relationship between two genes, in the third project of the thesis. We developed a new strategy to systematically detect co-eQTLs, while taking the large multiple testing burden into account caused by the huge search space. Using this, we identified a robust set of 72 co-eQTL SNPs associated with 946 gene pairs in a human single cell cohort. We interpreted the co-eQTLs with a combination of different enrichment analyses and found so new insights into several disease-associated eQTLs. For example, we identified a connection of the eQTL rs1131017-*RPS26* with T cell activation, potentially explaining its involvement in immune-related diseases.

Overall, we developed new methods and strategies around association studies, from study design over identification of associations to their interpretation. Both the analysis of DNA methylation and of single cell expression gave us new insights into the cell type specificity of associations. All developed computational methods are publicly available and will hopefully aid future users with their population analyses.

Kurzfassung

Ein Zusammenspiel aus genetischen und umweltbedingten Faktoren ist die Ursache vieler häufiger komplexer Erkrankungen, wie Asthma, Diabetes oder Depression. Obwohl zunehmend mehr dieser Zusammenhänge identifiziert werden können in großen menschlichen Kohortenstudien, sind die molekularen Ursachen der genetischen und umweltbedingten Faktoren, die zu den Krankheiten führen, oft noch kaum verstanden. Um die genetischen Konsequenzen besser zu untersuchen, konnten Assoziationsstudien zwischen genetisch Varianten und Gen-Expression, genannt expression quantiative trait loci (eQTL), einen Teil der genetischen Effekte erfolgreich aufdecken. Die Konsequenzen von Umwelteinflüssen auf zellulärer Ebene können eingefangen werden durch die Messung epigenetischer Faktoren, die die Gen-Aktivierung regulieren und stark beeinflusst werden von Umweltsignalen. Auch hier helfen Assoziations-Analysen zwischen epigenetischen Faktoren und anderen omics Leveln, z.B. dem Transcriptome oder dem Genome, dabei, Interaktionen und nachfolgende Effekte aufzudecken.

Dennoch konnte bisher nur ein kleiner Teil der phänotypischen Effekte von genetischen und umweltbedingten Faktoren auf molekularer Ebene erklärt werden. Zumindest teilweise wird diese Wissenslücke durch die Zelltyp-Spezifität von molekularen QTLs verursacht. Viele Assoziationen sind wahrscheinlich noch nicht identifiziert, ohne eine geeignete Analyse in den relevanten Zelltypen. Bis vor kurzem wurden die meisten Analysen mit bulk Gewebedaten ausgeführt, welche nur die durchschnittlichen omics Werte über alle gemessenen Zelltypen erfassen. Deswegen sind entweder spezialisierte Methoden für Zelltyp-spezifische Analyse mit bulk Daten notwendig oder die Verwendung von single cell Daten, um die Anzahl der identifizierten Assoziationen zu vergrößern und ihre Interpretierbarkeit zu verbessern. Beide Strategien werden in dieser Doktorarbeit untersucht. Die Arbeit umfasst insgesamt drei verschiedene Projekte, in denen aktuelle Assoziationsstrategien auf bulk und single cell Datensätze mit verschiedenen omics angewendet und verbessert werden.

Im ersten Projekt analysierten wir DNA Methylierung, eine wichtige epigenetische Modifikation, und ihre Interaktionen mit Genetik und Gen-Expression, um ihre Rolle in der Genregulation besser zu verstehen. Wir identifizierten 11,165,559 methylation quantitative trait loci (meQTLs), d.h. genetische Varianten, die DNA Methylierung beeinflussen, in einer großen multi-ethnischen bulk Kohorte. Um Zelltyp- und Kontext-Spezifität der meQTLs in diesem bulk Datensatz zu erfassen, identifizierten wir interaction meQTLs (iQTLs), meQTLs, deren Effektgrößen beeinflusst werden von der Zelltyp-Zusammensetzung der Spender oder Umweltfaktoren. Anschließend untersuchten wir den Einfluss von DNA Methylierung auf Gen-Expression mithilfe von expression quantiative trait methylations (eQTMs), wieder mit spezifischen Fokus auf die Zelltypen.

Die Anzahl eQTMs reduzierte sich drastisch von 54,898,225 auf 98,050 nach Korrektur für die Zelltyp-Zusammensetzung. Um besser zu verstehen, welche Gene von DNA Methylierung in welchen Regionen beeinflusst werden, untersuchten wir den genomischen Kontext für beiden Arten von eQTMs mithilfe verschiedener machine learning Ansätze und identifizierten deutliche Unterschiede: eQTMs ohne Korrektur für die Zelltypen konnten mit DNA Methylierung in Enhancer Regionen verknüpft werden, die Zelltyp-korrigierten eQTMs mit DNA Methylierung in Promoter Regionen.

Im zweiten Projekt erforschten wir die Verwendung von single cell transcriptomics, einer neuen technischen Entwicklung, für Zelltyp-spezifische eQTL Analysen. Jedoch existieren aktuell nur wenige single cell Kohorten. Ein durchdachtes experimentelles Design zukünftiger Kohorten ist entscheidend für die Durchführung aussagekräftiger Studien ohne Verschwendung von Ressourcen. Deswegen entwickelten wir eine statistische Power Analyse Software für single cell transcriptomics Studien mit mehreren Individuen, genannt scPower. Im Gegensatz zu anderen Methoden berücksichtigt sie die single cell spezifischen Eigenschaften und ist gleichzeitig sehr Laufzeit und Speicherplatz effizient. Sie ist einfach anwendbar auf unterschiedliche Szenarien unter Berücksichtigung der Zelltyp-spezifischen Expressionsverteilung. Die effiziente Berechnung erlaubt die Optimierung experimenteller Parameter unter Einschränkung auf vordefinierten Gesamtkosten. Unsere Evaluation zeigte, dass für eine große Anzahl Einstellungen, microfludics-basierte single cell Technologien, wie 10X Genomics und Drop-seq, in Kombination mit flacher Sequenzierung vieler Zellen die beste Power gab.

Zusätzlich erlauben single cell Studien neue Analysen verglichen mit klassischen eQTL Studien, die nur die Zielgene, beeinflusst von den genetischen Varianten, identifizieren. Die vielen Messpunkte pro Individuum ermöglichen die Rekonstruktion von personalisierten Gennetzwerken und so auch die Identifizierung von vorgelagerten regulatorischen Prozessen, die von den eQTL Varianten gestört werden. Dafür erfassten wir co-expression QTLs (co-eQTLs), genetischen Varianten, die das Co-expressions-Verhalten zwischen zwei Genen verändern, im dritten Projekt der Thesis. Wir entwickelten eine neue Strategie zur systematischen Erfassung von co-eQTLs, die die große statistische Einschränkung durch vielfaches Testen im großen Suchraum berücksichtigt. Unter Verwendung dieser Methode identifizierten wir eine robuste Menge von 72 co-eQTL SNPs assoziiert mit 946 Gen-Paaren in einer menschlichen single cell Kohorte. Wir interpretierte die co-eQTLs durch eine Kombination verschiedener Enrichment-Analysen und fanden so neue Erkenntnisse zu mehreren Krankheits-assoziierten eQTLs. Zum Bespiel entdeckten wir eine Verbindung des eQTLs rs1131017-*RPS26* zur T-Zell-Aktivierung, die potentiell seine Verbindung zu immunabhängige Krankheiten erklärt.

Insgesamt entwickelten wir neue Methoden und Strategien rund um Assoziations-Studien, vom Studiendesign über die Identifikation von Assoziationen bis zu ihrer Interpretation. Sowohl die Analyse von DNA Methylierung als auch von single cell Expression gab uns neue Einblicke in die Zelltyp-Spezifität der Assoziationen. Alle entwickelten Methoden sind öffentlich verfügbar und werden hoffentlich zukünftige Benutzer bei ihren Populations-Analysen unterstützen.

Preface

During my time as a PhD student at the Institue of Computational Biology, I contributed to several research projects that were published or are currently under review at scientific journals. This is the list of all publications and preprints together with the contribution of all collaborating authors to the respective project:

scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies

Katharina T. Schmid, Barbara Höllbacher, Cristiana Cruceanu, Anika Böttcher, Heiko Lickert, Elisabeth B. Binder, Fabian J. Theis and Matthias Heinig

Nature Communications (2021) [1]

Author contributions (detailed methods and results are described in the corresponding chapter): The idea for scPower was conceived by Matthias Heinig. Matthias Heinig and I developed the power analysis tool together with valuable input from Elisabeth B. Binder and Fabian J. Theis. I implemented the R package and ran all downstream analyses, including the comparison with simulation-based methods and the budget evaluation for different settings. I set up the website with support of Barbara Höllbacher. We obtained test data for the tool from Cristiana Cruceanu, Anika Böttcher, Heiko Lickert and Elisabeth B. Binder. Matthias Heinig and I wrote together the manuscript with input from all authors, especially from Barbara Höllbacher who improved the text and design of the first two overview figures.

Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function

Johann S. Hawe *, Rory Wilson *, **Katharina T. Schmid** *, Li Zhou, Lakshmi Narayanan Lakshmanan, Benjamin C. Lehne, Brigitte Kühnel, William R. Scott, Matthias Wielscher, Yik Weng Yew, Clemens Baumbach, Dominic P. Lee, Eirini Marouli, Manon Bernard, Liliane Pfeiffer, Pamela R. Matías-García, Matias I. Autio, Stephane Bourgeois, Christian Herder, Ville Karhunen, Thomas Meitinger, Holger Prokisch, Wolfgang Rathmann, Michael Roden, Sylvain Sebert, Jean Shin, Konstantin Strauch, Weihua Zhang, Wilson L. W. Tan, Stefanie M. Hauck, Juliane Merl-Pham, Harald Grallert, Eudes G. V. Barbosa, MuTHER Consortium, Thomas Illig, Annette Peters, Tomas Paus, Zdenka Pausova, Panos Deloukas, Roger S. Y. Foo, Marjo-Riitta Jarvelin, Jaspal S. Kooner **, Marie Loh **, Matthias Heinig **, Christian Gieger **, Melanie Waldenberger ** and John C. Chambers **

Nature Genetics (2022) [2]

Author contributions (detailed methods and results are described in the corresponding

chapter): Our collaboration partners from the Imperial College in London and the AME of the Helmholtz Center provided both cohorts (KORA and LOLIPOP) and additional dataset for validation. Basic meQTL analysis was performed by Benjamin Lehne, Roy Wilson and Marie Loh. Functional follow-up analysis and replication analyses were done by Johann Hawe. I contributed to the project with the global methylation iQTL analysis, the eQTL analysis and the eQTM analysis, supported by Johann Hawe and Matthias Heinig. Matthias Heinig did also the follow-up analysis of the iQTLs (replication and GWAS enrichment). The restricted iQTL analysis was performed by Roy Wilson. The project was led by John Chambers, Matthias Heinig, Christian Gieger, Melanie Waldenberger and Marie Loh. The unpublished follow-up analyses to characterize the eQTM context dependency in chapter 3 were all performed by myself under the supervision of Matthias Heinig.

Identification of genetic variants that impact gene co-expression relationships using large-scale single-cell data

Shuang Li *, **Katharina T. Schmid** *, Dylan de Vries *, Maryna Korshevniuk, Roy Oelen, Irene van Blokland, BIOS Consortium, sc-eQTLgen Consortium, Hilde E. Groot, Morris Swertz, Pim van der Harst, Harm-Jan Westra, Monique van der Wijst, Matthias Heinig ** and Lude Franke **

BioRxiv (2022) [3]

Author contributions (detailed methods and results are described in the corresponding chapter): The study was created in close collaboration together with researcher from the UMCG in Groningen. In general, Shuang Li, Dylan de Vries and me planned all analysis steps and discussed the results, with support and valuable input from Maryna Korshevniuk, Harm-Jan Westra, Monique van der Wijst, Matthias Heinig and Lude Franke. Specifically, we separated the analysis steps as follows: Dylan de Vries explored the Rho proportionality values and the Simpson's paradox. Shuang Li tested GRNBoost2, Scorpius, ran the co-eQTL mapping pipeline, the replication across cell types and the BIOS replication. Maryna Korshevniuk performed the eQTL mapping. I tested RNA velocity and the MetaCell algorithm, compared the correlation structure across datasets, cell types and donors, performed the evaluation with the CRISPR data and implemented the GO and TF enrichment analyses for the interpretation. Matthias Heinig performed the GWAS enrichment analysis. We obtained the published data from Monique van der Wijst, Roy Oelen, Irene van Blokland, the BIOS Consortium, the sc-eQTLgen Consortium, Hilde E. Groot and Pim van der Harst. The manuscript was written by Shuang Li, Dylan de Vries and me with input from all authors.

Another project I contribute to during my PhD, but that is not discussed in this thesis:

Identification of genetic effects underlying type 2 diabetes in South Asian and European populations

Marie Loh *, Weihua Zhang *, Hong Kiat Ng, **Katharina T. Schmid**, Amel Lamri,

Lin Tong, Meraj Ahmad, Jung-Jin Lee, Maggie C. Y. Ng, Lauren E. Petty, Cassandra N. Spracklen, Fumihiko Takeuchi, Md. Tariqul Islam, Farzana Jasmine, Anuradhani Kasturiratne, Muhammad Kibriya, Karen L. Mohlke, Guillaume Paré, Gauri Prasad, Mohammad Shahriar, Miao Ling Chee, H. Janaka de Silva, James C. Engert, Hertzel C. Gerstein, K. Radha Mani, Charumathi Sabanayagam, Marijana Vujkovic, Ananda R. Wickremasinghe, Tien Yin Wong, Chittaranjan S. Yajnik, Salim Yusuf, Habibul Ahsan, Dwaipayan Bharadwaj, Sonia S. Anand, Jennifer E. Below, Michael Boehnke, Donald W. Bowden, Giriraj R. Chandak, Ching-Yu Cheng, Norihiro Kato, Anubha Mahajan, Xueling Sim, Mark I. McCarthy, Andrew P. Morris, Jaspal S. Kooner **, Danish Saleheen **,John C. Chambers **

Communications Biology (2022) [4]

Contents

1. Introduction

1.1. Understanding complex diseases through omics data

Many common diseases are so-called complex traits induced by a combination of genetic variants and environmental factors. These factors individually have only a small effect on the phenotype, but cause in combination the diseases [5]. While many of these disease-associated genetic variants could be identified in large-scale association studies in the last years [6], the mechanistic insights of how the variants influence diseases are thus far often lacking [7], despite rapid technological developments in medicine and biology over the last years.

On the cellular level, genetic variants as well as environmental factors typically affect the downstream molecular levels, also called "omics" levels. The genome, the collection of all genetic information in a cell, has an influence on the transcriptome, the collection of all transcripts (and their transcription level), the proteome, the collection of all proteins (and their level), and the metabolome, the collection of all metabolites, i.e. small molecules [8]. Additionally, the epigenome, the collection of epigenetic information, such as histone modifications and DNA methylation, needs to be taken into account. Overall, the different omics levels are all connected in a complex and context dependent regulatory network. Identifying these relationships is necessary for the interpretation of genetic variants. In order to analyze these large and complex omics datasets, specialized computational and statistical methods are required, which need to be continuously adapted to new technological developments such as single cell omics technologies. In the end, increasing the knowledge about disease-associated genetic variants, their molecular consequences and the interplay with the environment is crucial for a better understanding of diseases, leading ultimately to better prevention, diagnosis and treatment.

1.1.1. Genome-wide association studies

The genome is the only stable omics layer, as opposed to the epigenome (anything else by definition). This means that the genome is identical in all cells, except for somatic mutations occurring during the lifetime of an individual. The genomic information is encoded in the DNA of each cell. Only a small fraction of the genome, the genes, are transcribed into RNA (Figure 1.1). On top of that, the genome harbors many regulatory regions, such as promoters and enhancers, that are part of the regulatory machinery guiding the time point and amount of transcription [9]. For many of these non-coding

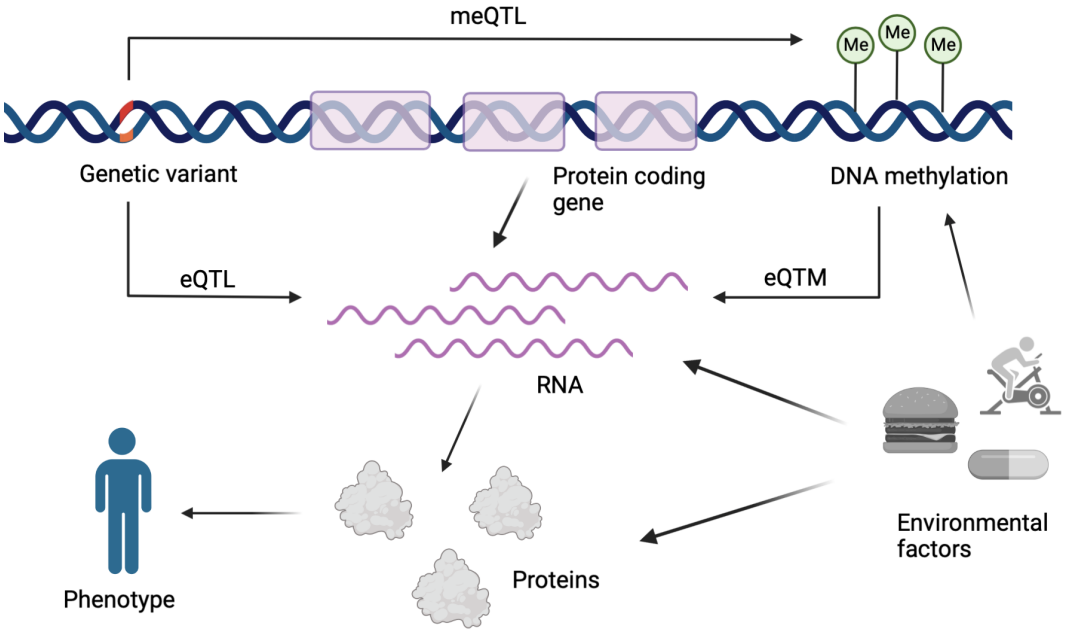


Figure 1.1.: **Molecular entities in a cell and their relationships**
Schematic representation of different biological entities, their relationships and how they can be studied in population genetics, focusing especially on aspects that are covered in this thesis. RNA levels of (protein coding) genes can be affected both by genetic factors (i.e. genetic variants) and epigenetics factors such as DNA methylation, which in turn is influenced itself by genetic factors and environmental factors, for example nutrition, medication and lifestyle. These genetic and epigenetic effects on RNA level can be propagated to the protein level (regulatory mechanisms might compensate it to some degree) and affect so the phenotype of the individual in the end. To study all these relationships, population genetic methods can be used: the identification of expression and methylation quantitative trait loci (eQTL and meQTL) measures the effect of genetic variants on gene expression and DNA methylation, respectively. The identification of expression quantitative trait methylation (eQTM) measures the effect of DNA methylation on gene expression. The figure was created with BioRender.com.

regions, especially the ones distal to genes, their exact function is not fully understood yet.

Research on genetic variants often focuses on point mutations, so-called single nucleotide polymorphisms (SNPs), which are very frequent and cost-efficient to measure in large populations with arrays. In diploid organisms such as humans, a SNP is characterized by a combination of two nucleotides AA, where the value of one chromosome (A) is called allele and the combination of both (AA) the genotype of a certain individual. Other genetic variants include short regions of insertion or deletions (indels) and larger structural changes such as copy number variants.

Almost all known diseases show associations with genetic variants. For a better understanding of these relationships, disease-associated variants can be broadly classified into two different categories based on their minor allele frequencies (MAF) and association strengths [7]. One part of the variants are rare variants with large effects, causing Mendelian diseases such as Huntington's disease [10]. Many of these variants were detected in family-based genomic studies, that focused on a small group of patients and their families [11].

Nowadays, technological improvements allow the measurement of genomic information from large cohorts. This led to genome-wide association studies (GWAS), where all measured genetic variants are tested for genotype frequency differences between cases (i.e. patients with the diseases) and controls (i.e. healthy donors) [12]. GWAS with large sample sizes identified numerous frequent variants with small effect sizes for many common diseases such as asthma, diabetes or depression [5]. But not only binary (or more general qualitative) traits, such as the disease status, can be associated with genetic variants. Furthermore, continuous traits, for example the height of an individual, can be associated with genetic variants [13], which are called quantitative trait loci. This shows also that GWAS can be used to explore arbitrary phenotypic traits besides diseases, for example educational attainment [14], and the same concepts hold true for all traits.

This distribution that variants with large effects are usually rare and variants with small effects are more frequent is explainable by selective forces, as deleterious variants causing diseases tend to be removed from a population [15]. Nevertheless, also frequent variants with small effect sizes are important for studying the disease phenotype. The omnigenic model proposes that complex diseases are driven by many variants with weak effects that accumulate to few core genes that cause the diseases [16]. The mathematical basis for this has already been laid by R.A.Fisher in 1918 who postulated that a continuous, normally distributed trait is created by a linear combination of many independent genes [17]. Therefore, many variants with weak effects can have together a similar impact as one rare variant with a strong effect.

Of note, the categorization into rare variants with strong effects and common variants with weak effects is in the end a trend and no discrete classification, and the same is true for their connection with rare Mendelian diseases and common complex diseases, respectively. Rare variants can also contribute to common diseases, and common variants can influence rare diseases. Underreporting of these associations could be impacted by

the different methods to study rare and common diseases [7].

Exact mapping of causal genetic effects is complicated by the correlation structure across neighboring variants within one population, called linkage disequilibrium (LD) [18]. This correlation structure allows capturing information of larger genomic region by measuring a selected set of representation SNPs with microarrays, but it makes biological interpretation more difficult. Combination of populations with different genetic ancestry and other fine-mapping techniques can reduce this issue and pinpoint the causal variants [19].

The number of identified GWAS variants increased tremendously in the last years, shown for example in the GWAS catalog [6] that contains currently over 400,000 variant trait associations from over 6,000 studies (status November 2022 [20]). However, interpreting the downstream effects of these genetic variants remains challenging, as the majority of GWAS variants can not be directly mapped to a gene [7]. Different studies reported that about 90% of the disease- and trait-associated variants lie in non-coding regions of the genome [21, 22]. However, enrichment of these variants in regulatory regions, for example at enhancers, suggests that they still affect RNA and/or protein levels by influencing their regulation, for example by perturbing transcription factor binding [21, 22].

1.1.2. Epigenetics

The non-coding part of the genome contains many regulatory regions, such as promoters and enhancers and more general transcription factor binding sites. The accessibility of these regions is crucial for gene regulation. In open regions, regulatory factors can bind, for example to initiate the transcription of a certain gene [23]. This accessibility of the genome is regulated by epigenetic marks, which are in contrast to the genome flexible and dynamic [24]. This way, each cell can obtain a distinct status, defined by its gene expression program, despite all cells having the same genome. Epigenetic factors are so especially important to define the cell type identity of each cell, based on cell type specific modifications which regulate the cell type specific expression.

The most frequent epigenetic modifications are post-translational modifications of histone proteins, around which the DNA is wrapped, and DNA methylation, mostly on cytosines of cytosine-guanine dinucleotides (short: CpGs) [25]. In both cases, different configurations regulate the accessibility of DNA and affect so the gene expression. However, in contrast to genetic regulation, where the genetic variant is expected to be always the causal factor, epigenetic variation can be both cause and consequence of regulatory processes [26].

Important influencing factors of epigenetic marks are environmental exposures, for example nutrition or smoking [27]. Furthermore, aging itself is strongly correlated with certain DNA methylation patterns, allowing the prediction of the individual's age based on their DNA methylation [28]. This makes epigenetic studies interesting for the exploration of environmental effects on molecular level. However, epigenetic patterns are also affected by genetic effects, so different impact factors need to be distinguished

carefully.

A related concept to GWAS are epigenome-wide association studies (EWAS), which identify epigenetic associations with diseases, currently mostly focusing on DNA methylation [27]. As the epigenome is dynamic, associations from EWAS depend on the measured cell type and condition. Although the approach is more recent, the EWAS catalog contained already over 2,600 EWAS at the beginning of 2022, highlighting the relevance of connecting epigenetics with diseases [29].

1.1.3. Quantitative trait loci analysis

Nevertheless, the interpretation of a genetic variant, in particular in the non-coding region of the genome, remains challenging, even with increasing epigenetic annotations. A direct way to quantify, whether a genetic variant has an effect on the transcription of a certain gene, is to test it for association with the gene expression level. Variants associated with a gene are called expression quantitative trait loci (eQTLs) [30, 31, 32, 33]. Following the approach of quantitative traits in GWAS, eQTL analyses test for associations between the genotype of a variant and the expression level of a gene (i.e. using expression level as the trait), as visualized exemplarily in Figure 1.2.

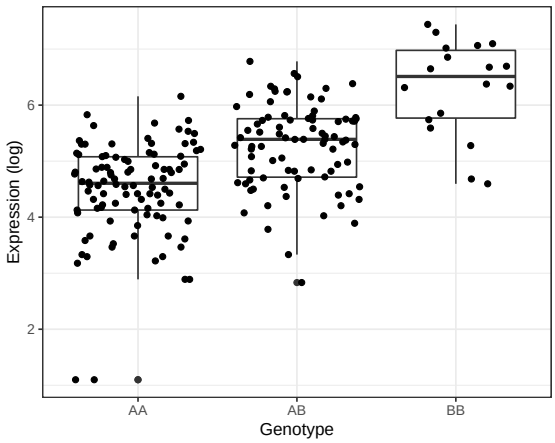


Figure 1.2.: **Expression quantitative trait locus (eQTL)**
Example visualization of an eQTL where the genotype for a specific genetic variant is significantly associated with the expression of a specific gene. Each dot represents one individual in the cohort. The example was generated using the simulation framework introduced in chapter 4, with a heritability of 40% and log-normalization of the simulated counts.

Different studies showed that many GWAS variants are also eQTL variants and vice versa [30, 31, 32, 33], proving the relevance of eQTL analyses for GWAS variant interpretation. With information from eQTL studies, causal genes of the disease can be identified, providing first a better understanding of the disease biology and leading in the long run potentially to novel diagnosis options and drug targets.

For the identification of eQTLs, the associations are commonly separated in two classes: cis eQTLs are genetic variants that are in the vicinity of the gene and trans eQTLs are variants that are farther away, potentially even on a different chromosome as the gene. The interpretation of both types differs. Cis eQTL variants tend to be in the gene body itself or the promoter region of the gene, hence, directly affecting the expression level [34]. For trans eQTLs, either the genetic variant lies in a distal enhancer region or it affects the expression of the gene indirectly by affecting the expression of a regulatory factor of the gene, for example by acting at the same time as a cis eQTL for a transcription factor of the gene [33].

In a systematic analysis of all eQTLs, all genetic variants need to be tested against all genes. This creates however a large search space and requires stringent multiple testing correction. For this reason, studies often focus only on cis eQTLs, which tend to have larger effect sizes, and so maximize their detection power [30].

A large part of the genes, the protein-coding RNAs, are further translated into proteins (Figure 1.1), in contrast to the non-coding RNAs, such as transfer RNAs, ribosomal RNAs, long non-coding RNAs and micro RNAs [35]. As proteins are the final products of gene expression, they are expected to give the best insight into phenotypic consequences of genetic variants and environmental factors [8]. In general, proteins are important regulatory and structural components of each cell, involved in a vast range of activities including transcriptional regulation, metabolism and signal transduction [36]. Applying the same association approach, protein quantitative trait loci (pQTLs) can be identified [37]. However, many current studies, including the projects discussed in this thesis, focus on the transcriptome, i.e. eQTLs, because the transcriptome can be quantified more efficiently and accurately in high-throughput experiments compared to the proteome until now. New and improved high-throughput methods for proteomics are currently developed and will potentially provide additional important insights in the future [8].

In general, the concept of quantitative trait loci can be applied to any other omics level, for example DNA methylation (meQTLs) [38, 2] (Figure 1.1), metabolite levels (mQTLs) [39] and so on. While genetic variants tend to be enriched for associations with multiple omics layers [34, 2], these other QTL types nevertheless give additional valuable information about the variants. For example, genetic variants were identified as independent pQTLs which affect protein levels, but not gene expression level [37]. This can happen when the genetic variant lies in a region important for post-transcriptional regulation, so that it affects only translation, but not transcription. Although several studies about QTLs with other omics layers exist, these associations are still underexplored compared to eQTLs.

Another underexplored type of association independent of genetics are expression quantitative trait methylation (eQTMs) [38, 2], associations between DNA methylation level and gene expression. With the exact role of DNA methylation in gene expression not fully understood, eQTMs provide the opportunity to quantify and annotate this relationship better. As both expression and DNA methylation are affected by genetics, a combined analysis of meQTLs, eQTMs and eQTLs is most promising for an exact

characterization of all dependencies.

1.1.4. Importance of cell type specific analysis

Despite the expectations for eQTL analyses, many GWAS variants remain whose functional effects could not be explained [40]. In some cases, this might be caused by power issues, as eQTL studies have usually smaller sample sizes. But also large-scale consortia, such as the Genotype-Tissue Expression (GTEx) Consortium [32] and the eQTLgen Consortium [33], could not overcome this knowledge gap completely. Another substantial part of the gap is caused by the context specificity of the QTLs, with context covering the relevant cell types, time points, stimulation and so on. Associations will only be detected if the respective omics data are measured in the relevant tissues and cell types, as all omics, except the genome, are highly cell type specific, as discussed before. For this reason, effects of a disease can only be fully understood when analyzing associations in relevant tissues and cell types.

Until today, many measurements of population genetics are "bulk" measurements. This means that all cells of a sample (usually a tissue) are measured at once and only one average value per gene/protein/CpG is taken. This leads to two potential pitfalls: firstly, associations from rare cell types within a tissue are potentially missed because they do not affect the average expression level in the whole tissue sufficiently enough. Secondly, associations might be falsely identified that are in reality not based on differences between individuals but instead caused by different cell proportions between the individuals (i.e. cell type differences instead of individual differences).

Several studies showed that eQTLs are context dependent and many potential important associations are missed in classical bulk studies. One of them is the GTEx Consortium, which measured RNA-seq in a large cohort of postmortem donors and 49 different tissues and provided so a large resource for tissue specific eQTLs [32]. Extending their analyses further, they detected a large number of cell type specific eQTLs, part of them not identifiable in the tissue specific analyses, and showed the relevance of these for interpreting GWAS results [41]. To get the necessary cell type specific information from bulk data, they applied in silico cell type deconvolution methods followed by an interaction analysis with the cell type proportion as interaction term (see Methods section 2.2.4 for further details). The same strategy with an interaction model was also successfully applied in other studies to identify cell type specific eQTLs, for example neutrophil and lymphocyte specific eQTLs in whole blood [42]. However, this kind of analysis crucially depends on the performance of the deconvolution approach and has limited power compared to standard eQTL analyses, especially for less frequent cell types. In the GTEx study, the authors therefore restricted their analyses to seven well characterized cell types. Overall, this type of analysis can give interesting insights into cell type specific eQTLs, but it is not able to systematically characterize all cell type specific eQTLs.

An alternative approach for cell type specific analyses is the use of fluorescence-activated cell sorting (FACS) sorted expression data, where cells of the chosen cell type

are selected based on marker proteins before the sequencing. By comparing eQTLs from different FACS-sorted bulk dataset, several studies identified cell type specific eQTLs, for example in the BLUEPRINT project [43] and the ImmuNexUT project [44]. However, also FACS-sorted eQTL analysis has several drawbacks. The cell types of interest need to be selected before, measured separately and knowledge of suitable markers for the FACS-sorting is required. Additionally, there is the danger of technical biases during the FACS-sorting that change the expression level.

As discussed in the beginning, the biological context for the identification of eQTLs comprises also other factors beside the relevant cell type. GTEx identified sex specific eQTLs, with stronger effect sizes in male or female, and population specific eQTL [32]. Other factors that can result on context specific eQTLs are for example the stimulation of cells or the respective time point during development [45, 46, 47].

Another important remark is that the cell type specificity - and more generally the context specificity - affects not only eQTLs, but all molecular QTL analyses, as other omics layers, such as DNA methylation and proteomics, also dependent on the cell type. The same approaches can be applied to overcome this in bulk, using either deconvolution or FACS sorting. For example, for DNA methylation, deconvolution into different blood cell types can be performed by the Houseman algorithm [48].

1.1.5. Advantages of single cell omics technologies

A third option to study cell type specific and generally context specific QTLs emerged very recently with the development of single cell omics technologies, which is a major technological breakthrough for biological research in general [49]. They allow measuring one or more omics layers separately in each cell and capture so the specific status of this cell, including for example the (sub)cell type, the cell cycle status and changes upon external stimuli. This way, single cell technologies quantify the cellular heterogeneity within a sample. They provided already novel insights into various biological processes, for example an improved reconstruction of developmental processes [50, 51] and a better characterization of tumor heterogeneity for different cancer types [52, 53, 54]. Recently, first studies started to apply single cell technologies in population studies [55, 56, 57, 58], in order to identify cell type specific eQTLs.

Transcriptomics was the first omics layer that was successfully measured at single cell resolution in 2009 [59]. Following this, several technologies for single cell RNA sequencing (scRNA-seq) were developed that vary in sensitivity and accuracy with regard to the quantification of gene counts, scalability to a large number of cells and experimental costs [60, 61]. For this reason, different approaches are most suitable for different biological questions.

The scRNA-seq technologies can be broadly classified into plate-based and microfluidics-based assays, dependent on how the cells are captured [62]. Plate-based methods, such as Smart-seq [63, 64, 65], CEL-seq [66] or MARS-seq [67], sort single cells in wells, allowing accurate estimation of a small number of cells. In contrast, microfluidics-based methods, such as Drop-seq [68], InDrop [69] or the 10X Genomics platform [70],

separate cells in droplets, fluidics circuits or nanowells. This enables high-throughput measurement of many cells at once, but with reduced accuracy compared to plate-based methods.

ScRNA-seq technologies differ additionally with respect to the coverage: in some cases, the full length of the transcript is sequenced, e.g. with Smart-seq [63, 64, 65], in other cases only the 3′ end or 5′ end of the transcript, e.g. with Drop-seq [68], InDrop [69] or 10X Genomics [70]. Measuring the full transcript allows additionally the quantification of different splicing isoform on top of estimating the expression level of the gene. In contrast, most of the methods that cover only the 3′ or 5′ end of the transcript incorporate unique molecular identifier (UMI), which help to remove biases during the PCR amplification and create more accurate count estimates [60]. However, with Smart-seq3, there now exists a method that sequences full-length transcripts and can incorporate UMIs [65].

ScRNA-seq is interesting for cell type specific analyses - for eQTLs and in general - as it allows the unbiased characterization of all existing cell types and sub cell types in a tissue, one of the major goals of the Human Cell Atlas [71]. Here, scRNA-seq has already led to the detection of novel cell types and better understanding of existing cell types in various tissues, generating healthy references, for example for lung tissue [72] and heart tissue [73]. On top of the description of static cell types, the measurement of many cells allows ordering them along trajectories to study differentiation and other dynamic processes [50].

The research field around scRNA-seq is growing very fast, with already over 1000 scRNA-seq datasets in 2020 [74] and over 1000 related software tools in 2021 [75], covering all analysis steps such as normalization, integration, clustering, differential expression (DE) analysis and visualization. Single cell transcriptomics provides a wide area of applications, besides its use for single cell eQTLs.

Also other omics layers can be captured at single cell resolution today, for example single cell chromatin accessibility [76, 77] and single cell methylation [78]. Additionally, single cell multi-omics technologies measure different omics layers together in the same cell [79, 80, 81, 82]. While single cell transcriptomics is currently the most used technique, other single cell omics nevertheless represent very promising technologies for future studies.

1.1.6. Single cell eQTLs

Single cell eQTL (sc-eQTL) studies make use of scRNA-seq data in order to identify cell type specific eQTLs. For this, the transcriptome of each individual is measured on single cell level and this is combined with the genotype information, usually quantified in bulk. The required cell type specific expression values for each gene and individual are typically obtained with the so called pseudobulk approach. Here, each measured single cell is annotated to a cell type and the counts of all cells per cell type and donor are summed up to the three-dimensional pseudobulk matrix of genes times donors times cell types (more in methods section 2.6). Using this, separate eQTL analysis of each cell

type with classical eQTL strategies developed for bulk data is possible (more in methods section 2.2).

Different recent studies have successfully applied the pseudobulk strategy to identify cell type specific eQTLs, so far mostly measured in blood and with growing cohort sizes from 45 donors [55] up to very recently close to 1,000 donors [57]. Other single cell cohorts explored single cell eQTLs in different contexts, for example after stimulation with different pathogens [83] or in Systemic lupus erythematosus cases [58]. Importantly, among the sc-eQTLs identified in the different studies, several examples were found that further characterized known disease-associated variants [83, 57, 58]. Compared to previous bulk eQTLs studies, novel (cell type specific) eQTL associations were detected, together with the disease-relevant cell types where these associations occurred.

On top of the pseudobulk approach to explore cell type specificity, single cell transcriptomics enables additional novel strategies to characterize eQTLs more in detail. One novel direction is the identification of dynamic eQTLs that change along differentiation trajectories, for example in IPS differentiation [56], B cell maturation [57] and in T cell state transitions [84]. For these analyses, the recently proposed statistical framework CellRegMap can be used [85]. It applies linear mixed models to capture interactions between genotype and context, i.e. eQTLs with changing effect size along discrete or continuous cell states. These states are estimated via factor analysis on the scRNA-seq data and can then be mapped to biological effects such as different (sub)cell types, cell cycle and cell differentiation.

Another novel direction for eQTL analysis arises from the multiple measurements per donor, which allow the construction of individual specific gene regulatory networks [86]. Network construction from scRNA-seq is a fast-growing field with many methods. Benchmarking studies showed however that none of the current construction methods is performing best in all situations, but that the method performance depends strongly on the specific dataset and task [87]. For population genetics, the identification of genetic variants that are associated with network properties is a promising extension of classical eQTL analysis, as it provides additional insights into how genetic variants can alter regulatory mechanisms.

To identify genetic variants affecting edges in co-expression networks, the idea of co-expression QTLs (co-eQTLs) was introduced [55]. These are genetic variants that affect the co-expression of a gene pair. Biologically, this relationship can be caused by genetic variants changing the binding affinity of a transcription factor at a certain binding site and so the co-expression of the transcription factor and the respective target gene (more in chapter 5). Therefore, co-eQTLs provide additional important insights into the effects of genetic variants compared to eQTLs, not only which downstream genes are affected, but also which upstream regulatory mechanisms are distorted.

While first studies with co-eQTLs provided already interesting results for a few preselected example loci [55, 83], several open questions remained: which association measure is working best to quantify the co-expression, how to identify a robust set of significant co-eQTLs despite the multiple-testing burden caused by the huge search

space and how to interpret the identified co-eQTLs properly.

Taken together, despite some open questions, all these studies already showed the value of scRNA-seq for studying context specific eQTLs and more complex associations such as dynamic eQTLs and co-eQTLs. The number of single cell eQTLs in the different studies is still relative small compared to large bulk studies, probably due to the small sample size of the cohorts. First consortia, such as the sc-eQTLgen consortium [88], have recently been established to overcome this power issue with meta-analysis of multiple cohorts together. The design of more well-powered studies is vital for the success of single cell population studies, and appropriate power analysis methods are required for this.

Likewise, best-practice analyses workflows for single cell eQTL studies are not yet established. First efforts have been made for pseudo-bulk based analyses, with evaluating the optimal choice for normalization, aggregation of cells, covariate selection and multiple testing correction [89]. However, for other analyses, such as co-expression QTLs, this has not been done yet.

Another future direction that has yet to be explored are association studies with other single cell omics layers besides transcriptomics. Currently, only single cell eQTL studies have been performed, as other single cell omics layers are not that well established yet. In general, the same kind of analyses are possible, for example for single cell meQTLs or pQTLs, given that the corresponding cohorts are generated in the future.

1.2. Scope and structure of the thesis

1.2.1. Aim and scope of this thesis

As discussed in this chapter, population (epi)genetics provides valuable insights into genetic and epigenetic gene regulation, which support the interpretation of molecular causes of disease. While previous strategies, focusing on gene expression and bulk analyses, led to the detection of important associations, novel approaches including other omics layers and moving towards single cell measurements promise a more accurate and complete picture of all associations. To aid this process, the goal of this thesis is the extension of previous population genetic analyses towards other omics layers and single cell data. Specifically, it covers the following three main aims:

- the identification of context specific meQTLs as well as cell type specific eQTMs, followed by the implementation of machine learning models to predict specificity of these eQTMs based on genomic features
- the development of a power and design framework for single cell multi-sample transcriptomics studies in order to facilitate the generation of more and larger single cell eQTL cohorts
- the exploration of novel approaches that use single cell transcriptome to study individual specific co-expression and genetic variants affecting this co-expression,

called co-eQTLs

1.2.2. Structure of the thesis

The three aims stated above are each addressed in one project chapter (3, 4 and 5) and represent together the main scientific outcome of this doctoral thesis. Overall, the thesis is structured into six chapters. First, this introduction chapter 1 gives a general overview about population genetics and new applications of single cell transcriptomics in this field to provide the required contexts for the work.

It is followed by a general method chapter 2 with important concepts and approaches shared between the projects. These include especially linear regression models for QTL analysis, the extension to generalized linear models, multiple testing correction, power analysis and pseudobulk approaches for single cell data. Additionally, the concepts of Random Forrest models are introduced. In the beginning of the methods chapter, definitions for all statistical distributions used in the thesis are given.

The project chapters 3, 4 and 5 describe each the results of the respective analysis, together with the applied methods, part of them newly developed by us.

The first project chapter 3 describes a large cohort study about DNA methylation which explored genetic influences on DNA methylation (meQTLs) and effects of DNA methylation on gene expression (eQTMs). For this, we identified context specific meQTLs and cell type specific eQTMs. Additionally, we applied different machine learning models to characterize genomic features specific for cell type dependent and cell type independent eQTMs.

Studying cell type specificity is not only important for DNA methylation, but also for gene expression. Single cell RNA-seq data provides new possibilities to study cell type specific eQTLs and DE. To facilitate the design of such multi-sample single cell transcriptomics studies, we developed in the second project a power analysis tool for it called *scPower*, as described in chapter 4. We applied the tool to explore optimal design combinations for different use cases and scRNA-seq technologies.

In the third project chapter 5, we made use of one of the first large scRNA-seq cohorts and identified co-eQTLs, genetic variants affecting co-expression between two genes. This new approach became possible because of the multiple measurement points for each donor in scRNA-seq. As no best-practice workflow exists for co-eQTLs yet, we evaluated different co-expression metrics, developed a novel approach to gain many high confident co-eQTLs from scRNA-seq data and explored strategies to interpret the identified associations.

The thesis ends with a common discussion for the three projects in chapter 6, including current limitations and future directions.

2. Methods

This chapter provides an overview over general methods used in this thesis. First, several probability distributions, which are important in the projects, will be introduced, followed by the description of linear regression models, the classical approach for eQTL analysis, and the extension to generalized linear models. Next, different multiple testing correction strategies and the estimation of experimental power are described, both applied in all the projects. For the two single cell projects, the pseudobulk approach is an important concept, also introduced here. At last, the Random Forest algorithm, an alternative prediction method, is explained shortly.

Additional to this methods chapter, project-specific methods are explained in the three projects chapters 3, 4 and 5, always at the end of each chapter.

2.1. Probability distributions

Describing the distribution of a variable in a certain dataset is an essential part of all projects, for example to choose the appropriate statistical tests or to estimate the experimental power. For this reason, different discrete and continuous probability distributions are defined here that are used in the projects in the next chapters. The probability distribution (also called density) will always be marked with a small f, the cumulative density distribution with a F.

2.1.1. Normal distribution

The normal distribution is a frequently used continuous probability distribution, defined by two parameters, the mean μ and standard deviation σ of the dataset. We will use it to approximate the distribution of effect sizes for the eQTL model (more in chapter 4). The density function is given by

$$f_N(x, \mu, \sigma) = N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}) \tag{2.1}$$

with $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_+$.

2.1.2. Binomial distribution

The binomial distribution is a discrete distribution that will be chosen in the following projects to describe sampling processes: given an event occurs with a probability of p,

$f_{Bin}(x, n, p)$ describes the probability to observe the event x times out of n total trials. The density function is given by

$$f_{Bin}(x, n, p) = Bin(x, n, p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{2.2}$$

with $n \in \mathbb{N}$ and $p \in [0, 1]$.

2.1.3. Negative binomial distribution

The negative binomial distribution is an adaptation of the binomial distribution. We use the definition, as implemented R, which describes the probability to observe x failures before the r-th success happens, when each success occurs with probability p. The density function is given by

$$f_{NB}(x, r, p) = NB(x, r, p) = \binom{x+r-1}{r-1} (1-p)^x p^r \tag{2.3}$$

with $r > 0$ and $p \in [0, 1]$.

In the following projects, we will describe negative binomial distributions with an alternative parameterization based on the mean $\mu = \frac{r(1-p)}{p}$ and the dispersion $\phi = \frac{1}{r}$, where the dispersion is related to the variance σ^2 of the distribution as $\sigma^2 = \mu + \mu^2 * \phi$.

The negative binomial distribution is especially important, as we apply it to describe the expression distribution of a gene. Several studies showed that not only bulk RNA-seq, but also microfluidic-based single cell RNA-seq data follow a negative binomial distribution [90, 91]. For plate-based technologies such as Smart-seq, a zero-inflated negative binomial distribution can be more appropriate.

2.1.4. Gamma distribution

The gamma distribution is a generalization of the exponential distribution. We will use it to describe the distribution of gene expression mean values across all genes in a dataset (more in chapter 4). It is parameterized by rate r and shape s. The density function is given by

$$f_\Gamma(x, r, s) = \Gamma(x, r, s) = \frac{s^r x^{r-1} e^{-sx}}{\Gamma(r)} \tag{2.4}$$

for $x \geq 0$ with $r, s \in \mathbb{R}_+$ and Gamma function $\Gamma(r)$ defined as

$$\Gamma(r) = \int_0^\infty x^{r-1} exp(-x) dx \tag{2.5}$$

We will also use the alternative parameterization with mean $\mu = \frac{s}{r}$ and standard deviation $\sigma = \sqrt{\frac{s}{r^2}}$.

2.2. Linear regression models

More complex approaches are necessary to capture not only the distribution of a variable, but model relationships between different variables. One of them is a linear regression model. Linear regression models are typically used to identify eQTLs (and other QTL), based on the assumptions that genetic variants have a linear effect on expression (and other omics layers). While non-linear eQTL models exist, including generalized linear and mixed models [92, 93, 94], they are usually very computationally intensive. In contrast, linear regression models allow efficient testing of a large number of SNP-gene pairs as required for systematic genome-wide analyses, for example implemented in the tool *Matrix eQTL* [95].

In general, a linear regression model estimates linear effects between a response variable $\boldsymbol{y}$ and a set with P predictor variables $\boldsymbol{x}_1, ..., \boldsymbol{x}_P$, based on observations from N samples. The predictors are combined in a data matrix $\boldsymbol{X} = x_{i,p}$ of size $Nx(P+1)$ for sample $i \in \{1, ...N\}$ and feature $p \in \{0, ..., P\}$, with $x_{i,0} = 1$ for the intercept of the model (see below). For a sample $i \in \{1, ..., N\}$ and the response vector $\boldsymbol{y} = (y_1, ..., y_N)$, the linear regression model is defined as

$$y_i = \beta_0 + \sum_{p \in \{1,...,P\}} \beta_p * x_{i,p} + \epsilon_i \tag{2.6}$$

with the error term ϵ_i assumed to be independent and identically normal distributed as $\epsilon_i \in N(0, \sigma^2)$ with unknown variance σ^2.

In the following, the matrix-vector notation of the model is used equivalently with coefficient vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ and error term vector $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_N)$ as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.7}$$

Based on the Gaussian error assumption, the linear regression model can also be formulated as

$$P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) \sim N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \sigma^2) \tag{2.8}$$

This coefficient vector $\boldsymbol{\beta}$, also called the effect sizes, can be estimated using Maximum Likelihood Estimation (MLE) or equivalently Ordinary Least Square (OLS) [96].

In case of a classical eQTL analysis, y_i represents the expression value of one specific gene of the individual i. $x_{i,1}$ is the genotype of this individual for one specific genetic variant, which is numerically encoded: in diploid organisms, genetic variants are numeric encoded with 0, 1 and 2, based on the number of minor alleles in the genotype, i.e. the allele dosage. So 0 represents the homozygote with the major allele, 1 the heterozygote and 2 the homozygote with the minor allele. Additionally, biological and technical covariates can be added to the model as $x_{i,2}, ...x_{i,P}$, such as the age of the donor or the experimental batch of the sample. The addition of covariates allows removing spurious associations that are not caused by direct interactions between the

genetic variant and the gene. One linear regression model is constructed for each tested SNP-gene pair, resulting in a family of tests (more in section 2.4).

Besides the testing for associations (more in section 2.2.1), linear regression models can be used to adjust a dataset for certain confounding factors, such as the covariates named above. For this, the residuals r_i of the model are calculated, which represent estimates for ϵ_i:

$$r_i = y_i - (\hat{\beta}_0 + \sum_{p \in \{1,...,P\}} \hat{\beta}_p * x_{i,p}) \tag{2.9}$$

The residuals capture the variance of the response variable $\boldsymbol{y}$ that can not be explained by the predictor variables $\boldsymbol{x}_1, ..., \boldsymbol{x}_P$. For this reason, these residuals represent an adjustment of the observations $\boldsymbol{y}$ for the confounding factors $\boldsymbol{x}_1, ..., \boldsymbol{x}_P$. We employed this approach to adjust eQTMs for several confounding factors, namely a set of basic covariates, genetic variants and the cell type composition (more in chapter 3).

2.2.1. Hypothesis testing

The linear regression models for eQTLs are used to test whether a certain SNP has a significant effect on the expression of the gene. For this, the effect size β_j of the variable $\boldsymbol{x}_j$, representing the SNP, is evaluated: the null hypothesis $H_0 : \beta_j = 0$, i.e. no effect of the SNP x_j, is compared with the alternative hypothesis $H_1 : \beta_j \neq 0$ [96]. The null hypothesis can be evaluated via a t-test statistic, using the estimated effect sizes $\hat{\beta}_j$ and its standard error $se(\hat{\beta}_j)$:

$$\hat{t} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{2.10}$$

The p-value under the null model is calculated based on the t-statistics with $N - P - 1$ degrees of freedom (for N samples and P predictors): $P(|t| \geq |\hat{t}|) \sim t_{N-P-1}$. If this p-value is smaller than a chosen significance threshold α, H_0 is rejected and the association between $\boldsymbol{x}_j$ and $\boldsymbol{y}$, in our example the SNP and the gene, is called significant. This approach with the t-test evaluates the effect size of one predictor variable $\boldsymbol{x}_j$. Alternatively, more complex hypotheses, such as multiple effect sizes being 0, can be evaluated using a F-test statistic (more in [96]).

2.2.2. Performance measures

The performance of the eQTL linear regression models is evaluated by comparing the outcomes, i.e. the results of the statistical tests, with the ground truth, which needs to be available for this evaluation. In general, the hypothesis testing in the eQTL models is a specific case of a binary classification, which classifies SNP-gene pairs into significantly associated pairs and not associated pairs, and can therefore be evaluated just like any other binary classifier. For this evaluation, the test results (over all SNP-gene pairs) are separated into four classes: true positives (TP), false positives (FP), false negatives (FN)

and true negatives (TN), dependent if the null hypothesis H_0 is accepted or rejected and if the ground truth is true or false (explanation in the so-called confusion matrix of Table 2.1).

	H_0 rejected	H_0 accepted
H_0 false	True positives (TP)	False negatives (FN)
H_0 true	False positives (FP)	True negatives (TN)

Table 2.1.: **Confusion matrix for hypothesis testing**
The outcome of a statistical test can be classified into four categories: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

With this categorization, the performance of a model can be evaluated under different aspects. Commonly used performance metrics that are also used in this thesis are:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.12}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.13}$$

$$False\ positive\ rate = \frac{FP}{TN + FP} \tag{2.14}$$

$$False\ discovery\ rate = \frac{FP}{FP + TP} \tag{2.15}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.16}$$

These metrics are generally used to evaluate the performance of classifications for all kind of machine learning models. We also applied it for the Random Forest results in chapter 3.

The disadvantage of the metrics is that they dependent on the chosen threshold for the classification, in case of the linear regression models the p-value cutoff. A threshold-independent alternative evaluation is the area under the receiver operating characteristic (ROC) curve, short AUC [97]. The ROC curve plots the sensitivity of a classifier against its false positive rate (1-specificity) (Figure 2.1). This is done systematically for all possible thresholds of the classifier, so that a curve is created from (0,0) to (1,1). The closer the curve gets to the optimal point of (1,0), which represents 100% sensitivity and 0% false positive rate, the better the classifier. To quantify this, the area under the ROC curve (AUC) is calculated. AUC values range from 0 to 1, with the best performance at 1. An AUC value of 0.5 represents a random classifier (a diagonal line in the ROC curve). An AUC value clearly below 0.5 is an indication that the class labeling might be wrong, swapping positive and negative class labels would result in a better performing classifier.

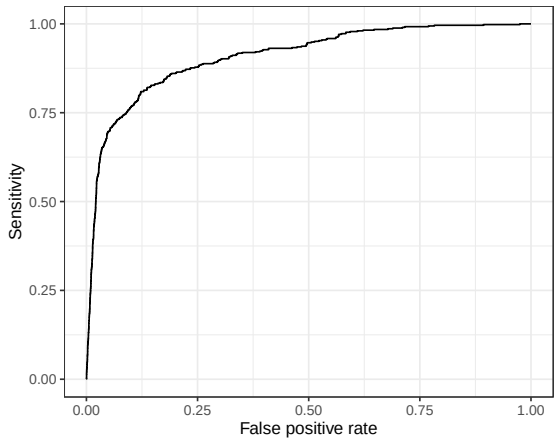


Figure 2.1.: **Receiver operating characteristic (ROC) curve**
Example visualization of a ROC curve to evaluate the prediction performance of a classifier in a threshold-independent way. The shown example depicts the performance of a logistic regression model that classifies CpG-gene pairs into significant eQTMs and non-significant pairs (taken from chapter 3). The area under the ROC curve (AUC) of the model is 91% in this example.

2.2.3. Meta analysis

Many of the first GWAS focused on individuals from European descent, which limited the generalizability with regard to individuals from other ancestry and the detection power in general [98]. For this reason, more and more current studies combine data from individuals with diverse ancestry backgrounds, including also our meQTL and eQTM study (chapter 3). Meta analysis is a frequently used approach in multi-population studies to avoid an excess of false-positives caused by the population structure [98]. The different cohorts, each comprising people with similar ancestry, are analyzed separately, i.e. the association tests are performed within the cohorts. Afterwards, the summary statistics from the different cohorts are combined to one general result. More broadly, meta analyses can be used to increase power by combining results from different cohorts. For this reason, we applied a meta analysis also in our co-eQTL project (chapter 5), where different scRNA-seq protocols were used for the different cohorts to avoid technical batch effects during the analysis.

In both studies (chapter 3 and 5), we applied a meta analysis approach that combined the effect sizes of the separate studies. P-value based meta analyses have several limitations, such as reduced interpretability (more in [99]). Meta analysis approaches using effect sizes can be divided in fixed effect and random effect models. Fixed effect models assume that the same (fixed) effect occurs in each cohort, only masked by power issues, while random effect analysis assumes different effects for different cohorts, which all follow a common distribution that can be identified in the meta analysis [99]. We chose a fixed effect model which weights the effect sizes of each study by the inverse-

variance of the study. This is beneficial, as it gives more weight to studies with lower variance, for example caused by higher sample sizes [100].

The inverse-variance weighted fixed effect model works as follows for one hypothesis test and a meta analysis of N studies: the weight w_i for each study $i \in \{1, ..., N\}$ is estimated based on its standard error $\hat{\sigma}_i$:

$$w_i = \frac{1}{\hat{\sigma}_i^2} \tag{2.17}$$

The pooled effect size estimate of the meta analysis $\hat{\beta}$ and the standard error $\hat{\sigma}$ are then estimated based on the weights w_i and the estimated effect sizes $\hat{\beta}_i$ of all studies:

$$\hat{\beta} = \frac{\sum_{i=1}^{N} w_i * \hat{\beta}_i}{\sum_{i=1}^{N} w_i} \tag{2.18}$$

$$\hat{\sigma} = \frac{1}{\sqrt{\sum_{i=1}^{N} w_i}} \tag{2.19}$$

The p-value for the meta analysis can be estimated via the Z-score $\hat{\beta}/\hat{\sigma}$ as:

$$p = 2 * F_{norm}(-|\hat{\beta}/\hat{\sigma}|) \tag{2.20}$$

2.2.4. Interaction models

Gene regulation is a very complex process, as a variety of both cellular and extrinsic factors influence the expression level of a gene. The classical eQTL model that assume a linear relationship between the expression level of a gene and the genotype of a genetic variant is an over-simplification, which allows fast identification of eQTL associations. However, it neglects that the eQTL effect size of a genetic variant is not always constant, but often influenced by extrinsic factors. This relationship is called genotype environment interaction [101]. In these cases, linear regression models with an interaction term can be used to evaluate interacting effect between two variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, with $\boldsymbol{x}_1$ representing the SNP and $\boldsymbol{x}_2$ an environmental factor for our use case [46]. For sample $i \in \{1, ..., N\}$, the interaction model is defined as:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} * x_{i,2} \tag{2.21}$$

If the hypothesis test with $H_0 : \beta_3 = 0$ and $H_1 : \beta_3 \neq 0$ results in a significant p-value, there is an interaction effect between the two variables. This means that the effect of the SNP on the gene expression can be stronger or weaker, dependent on the environmental factor. We applied this approach in two of our projects, to identify interaction meQTLs (chapter 3) and to identify co-eQTLs in bulk datasets (chapter 5).

2.3. Generalized linear models

Generalized linear models (GLMs) are an extension of classical linear regression models [102]. The addition of a link function $g(x)$ allows connecting the linear predictor $\eta = \beta_0 + \sum_{p\in\{1,...,P\}} \beta_p * x_{i,p}$ to the expected mean of a distribution $E(\boldsymbol{y}|\boldsymbol{X}) = \mu$:

$$\eta = g(\mu) \tag{2.22}$$

The error terms do not need to be normally distributed anymore, as was the case for the classic linear regression, but instead can follow an exponential family of distributions. The details behind GLMs are further introduced with two common variants of GLMs, the logistic regression and the negative binomial regression.

2.3.1. Logistic regression

The logistic regression enables the prediction of a binary outcome (0,1) instead of a continuous outcome, as is the case for the linear regression. For this, the logit function $logit(\mu) = ln(\frac{\mu}{1-\mu})$ is used as the link function, which restricts the outcome values to the range of 0 to 1 for $\mu \in \mathbb{R}$. The logit function as link function produces the following generalized linear model for a set of P predictors:

$$ln(\frac{P(y_i = 1)}{1 - P(y_i = 1)}) = ln(\frac{P(y_i = 1)}{P(y_i = 0)}) = \beta_0 + \sum_{p\in\{1,...,P\}} \beta_p * x_{i,p} \tag{2.23}$$

This can be transformed using the inverse link function, which is in this case the logistic function:

$$P(y_i = 1) = \frac{1}{1 + exp(-(\beta_0 + \sum_{p\in\{1,...,P\}} \beta_p * x_{i,p}))} \tag{2.24}$$

This way, the logistic regression predicts the probability of a sample to be in class 1. The effect sizes β_p are log odds ratios in the logistic regression model. This means that odds ratio $\frac{P(y_i=1)}{P(y_i=0)}$ is increased by $exp(\beta_p)$ for each unit increase of the predictor variable $\boldsymbol{x}_p$. We applied the logistic regression for the prediction of eQTMs in chapter 3, also because of their good interpretability of the gained effect sizes.

2.3.2. Negative binomial regression

Another variant of generalized linear regression is the negative binomial regression, which is the basis of several established differential expression (DE) analysis methods, such as *edgeR*, *DESeq* and *DESeq2* [103, 104, 105]. DE analysis is applied to identify genes that show significantly different expression levels between two (or more) pre-defined groups, for example disease patients vs healthy controls. The choice of negative binomial regression for DE analysis follows directly from the assumption that the gene expression counts are negative binomial distributed (section 2.1.3). In the following, the negative

binomial regression is further introduced using the approach of *DESeq* [104] and *DESeq2* [105], which we utilized for our own power analysis method in chapter 4.

As mentioned, the read counts of gene i and sample j are assumed to follow a negative binomial distribution with mean $q_{i,j}$ and dispersion parameter ϕ_i: $K_{i,j} \sim NB(q_{i,j}, \phi_i)$. The mean $q_{i,j}$ is affected by differences in read depth between samples. To corrected for it, a scaling constant called size factor $s_{i,j}$ is derived by the median-ratio method (default) or an alternative normalization method and the normalized mean $\mu_{i,j}$ is estimated with $q_{i,j} = s_{i,j} * \mu_{i,j}$. Next, $\mu_{i,j}$ is used as the response variable in the generalized linear model with log_2 as link function to ensure only positive outcomes for the counts:

$$log_2(\mu_{i,j}) = \beta_{i,0} + \sum_{p \in \{1,\dots,P\}} \beta_{i,p} * x_{j,p} \tag{2.25}$$

The variables $x_{j,p}$ of the model are defined in a so-called design matrix, which contains the group distribution of the samples together with a set of covariates. The derived effect size represents the log fold change between the two tested groups. Different default hypothesis testing methods are specified in *DESeq* and *DESeq2*. In *DESeq*, a likelihood ratio test is calculated with a reduced model, based on a reduced design matrix. In *DESeq2*, a Wald test is done directly with the β values and their standard errors instead.

One limitation of the negative binomial regression models is that a reliable estimation of the variance of each gene is difficult with the usually small sample size of the datasets. For this reason, both *edgeR* and *DESeq*/*DESeq2* model the variance dependent on the mean across all genes to stabilize the estimation. We will utilize the function, as implemented in *DESeq*, for our own algorithm in chapter 4. Further "tricks" that are applied in *DESeq* and *DESeq2*, for example a shrinkage for the log fold change estimation, can be found in the respective manuscripts [104, 105].

2.4. Multiple testing correction

Both DE and QTL analyses have in common that many tests are performed, as 10,000s of genes are tested in the DE case and several millions of SNP-gene pairs in the eQTL case. These high number of tests increases the probability of at least one false positive within the experiment, i.e. the probability that the null hypothesis H_0 is rejected in one case, when H_0 is true (Table 2.1) [106]. This is also called type I error. For a single test, the significance threshold α controls the probability of a false positive. For a family of N tests, the probability of at least one false positive - the family-wise error rate - increases to $1 - (1 - \alpha)^N$. In this context, family describes a group of tests, all performed on the same dataset. For 100,000 tests, already 5,000 false positives are expected under a typical uncorrected significance threshold of 0.05, which impedes the biological interpretation of the results drastically.

To overcome this so-called multiple testing problem, either the significant threshold can be adjusted or the p-values of the tests, both strategies being equivalent. Different approaches have been developed for the multiple testing correction. One major

difference between them is how strict they correct, as a very stringent multiple testing correction reduces the overall number of positive tests and potentially also discards true positive results. In the following, two of the most common methods are introduced: Bonferroni correction, which controls the family-wise error rate (FWER), and the Benjamini-Hochberg correction, which controls the false discovery rate (FDR). Both are used in this thesis and their differences are shown in chapter 3. Furthermore, an approach for FWER correction is explained that takes correlation structure between tests into account and is used for multiple testing correction in chapter 5.

2.4.1. Family-wise error rate and Bonferroni correction

Bonferroni correction of the significance threshold α' has the goal to reduce the FWER to α [106], i.e. the probability of at least one false positive result in the family of tests should be less than or equal α. For this, the adjusted threshold α is simply corrected for the total number of tests N as

$$\alpha = \frac{\alpha'}{N} \tag{2.26}$$

Alternatively, the p-values p_i of each test i from the family of tests can be corrected to

$$p_i' = max(p_i * N, 1) \tag{2.27}$$

2.4.2. False discovery rate and Benjamini-Hochberg correction

In contrast, the FDR controls the fraction of false positives among the tests (section 2.2.2). This is less strict than the FWER, as it allows a certain number of false positives as long as they do not exceed the threshold. Benjamini-Hochberg correction is an established method to correct for FDR [107]. It depends on the p-value distribution over all tests and is performed as follows: the uncorrected p-values p_i from all N tests are sorted in increasing order so that $0 \leq p_1 \leq ... \leq p_N \leq 1$. Next, the largest index k is identified so that

$$p_k \leq \frac{k}{N} * \alpha \tag{2.28}$$

The null hypothesis H_0 is rejected for $\{p_1, ..., p_k\}$ to reach an FDR rate of α.

Alternatively, the p-values can be again adjusted, which allows easier application of different significance thresholds. For this, the adjusted p-values q_i are calculated from the ordered list of p-values, starting from the largest p-value, so in decreasing order:

$$q_i = min(p_i \frac{N}{k}, q_{i-1}) \tag{2.29}$$

2.4.3. Permutation-based FWER correction

Due to the linkage disequilibrium between SNPs, tests in GWAS and eQTL studies are strongly correlated. For this reason, correcting by the total number of tests, as done in the Bonferroni approach, is too stringent for FWER correction in these cases. Because of the correlation structure between tests, the actual number of independent tests, which need to be corrected for, is smaller than the total number of tests. A more accurate FWER correction, taking the correlation into account, can be achieved by permutation-based approaches such as the Westfall-Young method [108].

Here, the outcome variable of the dataset - for eQTLs the expression - is permuted b times. Each time, p-values of the family of tests are computed and the minimal p-value is saved in a list $\{p_{min,1}, ..., p_{min,b}\}$. Afterwards, the adjusted p-value $p_{adj,i}$ for each p_i from the original tests is calculated: it is inferred from the fraction of permuted minimal p-values $p_{min,j}$ which are smaller than the raw p-value p_i, as the permuted p-values are all false positives by design.

$$p_{adj,i} = \frac{\sum_{j=1}^{b} \#(p_{min,j} \leq p_i)}{b} \tag{2.30}$$

For the eQTL use case, this permutation strategy is applied separately for each gene, based on the assumption that the statistical tests between the genes are independent and only the correlation structure between SNPs per gene needs to be considered. This results in a two-step process: first, for each gene, the permutation-based p-values across the SNPs are calculated to correct for multiple testing and to take thereby the correlation structure between SNPs into account. Second, the eQTL results need to be corrected for multiple-testing over the different genes. Here, a standard approach such as Benjamini-Hochberg correction can be applied, given the independence assumptions between the genes.

The disadvantage of this method is that it requires a large number of permutations for accurate results, which causes long run times. The approach of *FastQTL* [109] can reduce the number of permutations by extrapolating their general distribution and still generates good estimations for the adjusted p-values. For this, they utilize the assumption that the list of minimal p-values follows a beta distribution, which is well established for independent uniformly distributed random variables [110].

The original permutation algorithm is adapted as follows: first, again b permutations are run (but taking here a much smaller number) and the minimal p-values $p_{min,j}$ of each permutation are saved. Over these, a Beta-distribution is fitted:

$$p_{min,j} \sim Beta(a, b) \tag{2.31}$$

The adjusted p-values $p_{adj,i}$ for each p_i are estimated based on the cumulative density distribution of the fitted Beta distribution F_{Beta}:

$$p_{adj,i} = F_{Beta}(p_i, a, b) \tag{2.32}$$

An adaption of the FastQTL algorithm is used in chapter 5 to account for the correlation structure between the co-eQTLs.

2.5. Power analysis

Both for experimental design and for interpretation of results, it is important to consider the power of an experiment. The power is another term for sensitivity, the fraction of true positives among all true observations of an experiment, i.e. the number of tests where H_0 is correctly rejected (Table 2.1). Reaching as high power as possible, while properly controlling for the false discovery rate (section 2.4), is one of the main goals when designing new experiments. In order to achieve this, power analysis can estimate the expected power given the planned experimental parameters and certain prior assumptions about effect sizes and similar. Additionally, power analysis can also be helpful after the experiment was conducted to compare the found results with the expected outcome. We applied power analysis methods in each project chapter, with specific focus of course in chapter 4, where we developed an analytic power analysis method called *scPower*. The differences between analytic and simulated power analysis methods will be explained more in the following.

2.5.1. Analytic power analysis methods

Analytic power analysis methods estimate the power statistically given the expected distribution for the test statistics under the null model and under the alternative model, e.g. with a certain effect size [111]. For example, going back to the eQTL hypothesis testing in section 2.2.1: we assume that the test statistics follow under both models a normal distribution with the same variance $var(\beta)$, only the mean is shifted by $\hat{\beta}$ for the alternative model:

$$P(\beta|H_0) \sim N(0, var(\beta)) \tag{2.33}$$

$$P(\beta|H_1) \sim N(\hat{\beta}, var(\beta)) \tag{2.34}$$

The significance threshold α, chosen by the user, defines the cutoff c in the test statistic that separates the cases where the null hypothesis is rejected and the cases where the null hypothesis is accepted. For the eQTL example, it would be $P(\beta > c|H_0) \leq \alpha$. The area of the expected distribution (i.e. the probability mass) from the alternative model where the null hypothesis is truthfully rejected, is equal to the power: $P(\beta > c|H_1)$.

Analytic power analysis methods are very time and memory efficient to apply compared to the simulation-based methods, which are introduced next. However, an analytic formulation of the power is not possible for every statistical test. For this reason, simulation-based methods are an important alternative.

2.5.2. Simulation-based power analysis methods

Simulation-based methods simulate the expected dataset directly, again given certain assumptions about the distributions, and then exemplarily run the analysis workflow on the simulated dataset to evaluate the power. For example, different single cell DE power analysis methods [112, 113] simulated one instance of a potential single cell count matrix and add a specified number of DE genes with certain effect sizes by shifting the distribution of some genes before sampling the matrix.

Assumptions about effect sizes, number of DE genes etc., need to be made for any power analysis method, analytic or simulation-based. Typically, data from previous experiments or pilot data can be helpful here. The simulated count matrix with the dimensions of the planned experiment (cells times genes) is then processed using the planned workflow as for the real experiment. Because of the simulation, the ground truth is known, i.e. which genes are the real DE genes. Using this, the power is calculated in the end as the fraction of correctly identified DE genes (true positives) from all simulated DE genes.

As the simulation produces only one potential instance of the count matrix, accuracy of power estimation is increased by running this simulation multiple times in a row. This process is time and memory consuming, making quick evaluations and the comparison of different design options very tedious compared to analytic power analysis methods. However, simulation-based methods have far greater flexibility, for example the single cell DE power methods allow the easy comparison of different workflows, e.g. different normalization and imputation methods. Overall, both types of methods, analytic and simulation-based, give valuable information about the power and should be chosen dependent on the use case.

2.6. Pseudobulk approach for single cell data

In contrast to bulk data, multi-sample single cell data contains multiple measurement points per individual, as typically many cells are measured from each individual. While this data structure allows interesting new analyses, methods developed for bulk, such as the (generalized) linear regression models for the eQTL and DE analysis, can not be applied directly. A common strategy to overcome this is the so-called pseudobulk approach. The idea is to aggregate measurement points per individual and cell type to one value, using either the sum or the mean function [89].

Given a count matrix $x_{i,j}$ of genes times cells, the counts for each gene i are aggregated over all cells j annotated to cell type c and individual s (noted as $j \in C$ with C the set of all cells from cell type c and $j \in S$ with S the set of all cells from individual s, respectively). The result is a three-dimensional pseudobulk matrix $y_{i,c,s}$ of genes times individuals times cell types with

$$y_{i,c,s} = \sum_{j \in C \wedge j \in S} x_{i,j} \tag{2.35}$$

Two independent benchmarking studies showed that pseudobulk approach outperformed other methods for single cell DE analysis [113, 114]. It reduces the number of false positives compared to approaches that do not account for multiple measurement points per individual, and it is computationally far more efficient than mixed models, an alternative strategy to accurately deal with the data structure. For these reasons, it is also a frequently used method for single cell eQTL studies [55, 57, 58, 83].

We developed our power analysis framework for multi-sample single cell experiments on the basis of the pseudobulk approach (chapter 4). Furthermore, we applied the pseudobulk approach for single cell eQTL mapping in the co-eQTL project (chapter 5).

2.7. Random forest classification

Linear regression models are limited in the detection of more complex interactions, as they assume linear relationships between the features and the outcome variable. Alternatively, several other more sophisticated machine learning models became frequently used for various biological applications in the last years due to the access to larger datasets, for example Random Forest [115] and Artificial (Deep) Neural Networks [116]. In this thesis, we applied a Random Forest classifier as an alternative to the Logistic Regression classifier (chapter 3). This will be quickly introduced in the following paragraphs.

The Random Forest method is a collection of many classifiers, whose results are combined to one final prediction. The approach is an extension of the classic Decision Trees [115]. After a Decision Tree is built, each sample can be classified by following a path from the root to the leafs in the tree. At each inner branching point, the sample can be categorized into one of the subtrees based on a specific condition of a chosen variable (e.g. $a < 3$ and $a \geq 3$). At the end of the path, the sample gets the class label of the corresponding leaf.

During the construction of the tree, the variable and split at each node is chosen so that it reduces the heterogeneity of the training dataset in the subtrees, i.e. that as many samples as possible belong to the same class in one branch. This is performed recursively, so that the leafs contain homogeneously (or at least nearly homogeneously) one class.

The Gini index is one possible metric to measure the purity within a node and so the best split during the construction of the tree [117]. It is defined for a node N with J classes and a relative frequency p_i for each class $i \in \{1, 2, ..., J\}$ at the node as:

$$Gini(N) = 1 - \sum_{i=1}^{J} p_i^2 \tag{2.36}$$

The Gini index is higher if there is more heterogeneity in the node. The best split A of node N into m subtrees N_i with $i \in \{1, 2, ..., m\}$ is the one with the highest $\Delta Gini(N, A)$:

$$\Delta Gini(N, A) = Gini(N) - \sum_{i=1}^{m} \frac{|N_i|}{|N|} * Gini(N_i) \tag{2.37}$$

The Gini index is used in the Random Forest implementation of the R package *randomForest* [118], which we applied in our project (chapter 3). Other alternative metrics exist to determine the best split, for example Information Gain [117].

While Decision Trees are efficient to train and easy to interpret, they tend to overfit quickly. This risk is reduced in the Random Forest algorithm, which builds on multiple Decision Trees, often hundreds of trees, each trained on a random subset of the data. Typically, for each tree, a new training dataset is sampled with replacement from the original dataset and at each node, the variable set is sampled which is tested for the best split. The two randomization steps ensure that the individual trees in the Forest can be regarded as independent classifiers. In the end, the predictions of all trees are combined to one majority voting. We will apply Random Forest for classification, but Random Forest regression is also possible.

3. Studying the relationship of DNA methylation with genetics and transcriptomics

3.1. Increasing the knowledge about DNA methylation and its complex interplay with the genome and transcriptome

DNA methylation is an important epigenetic mark connected with many diseases from type 2 diabetes and obesity over asthma to cancer [119, 120, 121, 122, 123]. Interestingly, DNA methylation captures environmental influences such as smoking or diet [124, 125], which makes it a promising surrogate marker to study the effect of environmental factors on diseases. As DNA methylation is also affected by genetic variants, called methylation quantitative trait loci (meQTLs) [38], both genetic and environmental effects need to be taken into account when studying DNA methylation. In contrast to eQTL studies, meQTLs are not as extensively characterized in large cohorts yet.

For this reason, also the cell type specificity of meQTLs and more generally the context specificity is not well studied yet. It is known that DNA methylation, similar to gene expression, is very cell type specific [48]. However, currently only bulk cohorts with DNA methylation data exist, which capture the average methylation levels across all measured cell types and impede so cell type specific analyses.

Open questions exist not only about the upstream regulatory factors of DNA methylation, but also about the downstream effects of DNA methylation variation on gene expression. DNA methylation is associated with stable silencing of expression, including processes such as X-chromosome inactivation, genomic imprinting and silencing of repetitive DNA elements [126]. But also in general, methylation of promoter regions is typically negatively correlated with gene expression [38]. However, the relationship between DNA methylation and gene expression is more diverse: associations with gene expression were found for methylation at enhancers [127, 128] and in the gene body, in the second case even with positive associations, i.e. an increase in gene body methylation is associated with higher expression levels [129]. The general rules, which types of DNA methylation influences a specific gene in which way, are not understood yet [130]. The diverse results indicate that genomic context of the CpG and maybe also the gene plays a role, for example the position of the CpG, its distance to the gene or the promoter type of the gene.

In the following, we studied both the connection of DNA methylation with genetics and its influence on gene expression with different QTL approaches applied to two

large cohorts with diverse ancestry, the KORA cohort with 3,799 European donors and the LOLIPOP cohort with 3,195 South Asian donors. Both cohorts comprised the three necessary omics layers, genotype information measured with Illumina genotyping arrays, DNA methylation measured with the 450K array and gene expression measured again with arrays. In the first step, we identified both cis and trans meQTLs in whole blood via a large meta analysis of the KORA and LOLIPOP cohorts to increase our knowledge about DNA methylation and genetics. Furthermore, the large sample size enabled us to perform the first study of interaction meQTLs, called iQTLs, to explore cell type and context specificity of meQTLs. In this analysis, the cell type composition, BMI and cigarette smoking were tested as interaction terms, which potentially influence the effect size of the meQTLs.

A QTL approach was also used to study the relationship with gene expression and DNA methylation. For this, we mapped expression quantitative trait methylations (eQTMs), i.e. DNA methylation influencing gene expression, via a meta analysis of KORA and LOLIPOP. We studied especially which effect the additional adjustment for cell type composition has on the identified eQTMs. Afterwards, to follow up on the question of the genomic context of associated CpG-gene pairs, we predicted the eQTM probability of CpG-gene pairs using a large set of genomic annotations as features. For the prediction, we tested Logistic Regression and Random Forest and identified each time the most important features that drive the eQTM probability. This way, we got more insights of which CpGs influence the expression of which genes.

Overall, the study increased our knowledge about genetic effects on DNA methylation as well as gene expression and DNA methylation and how all these relationships are affected by cell types and environmental factors.

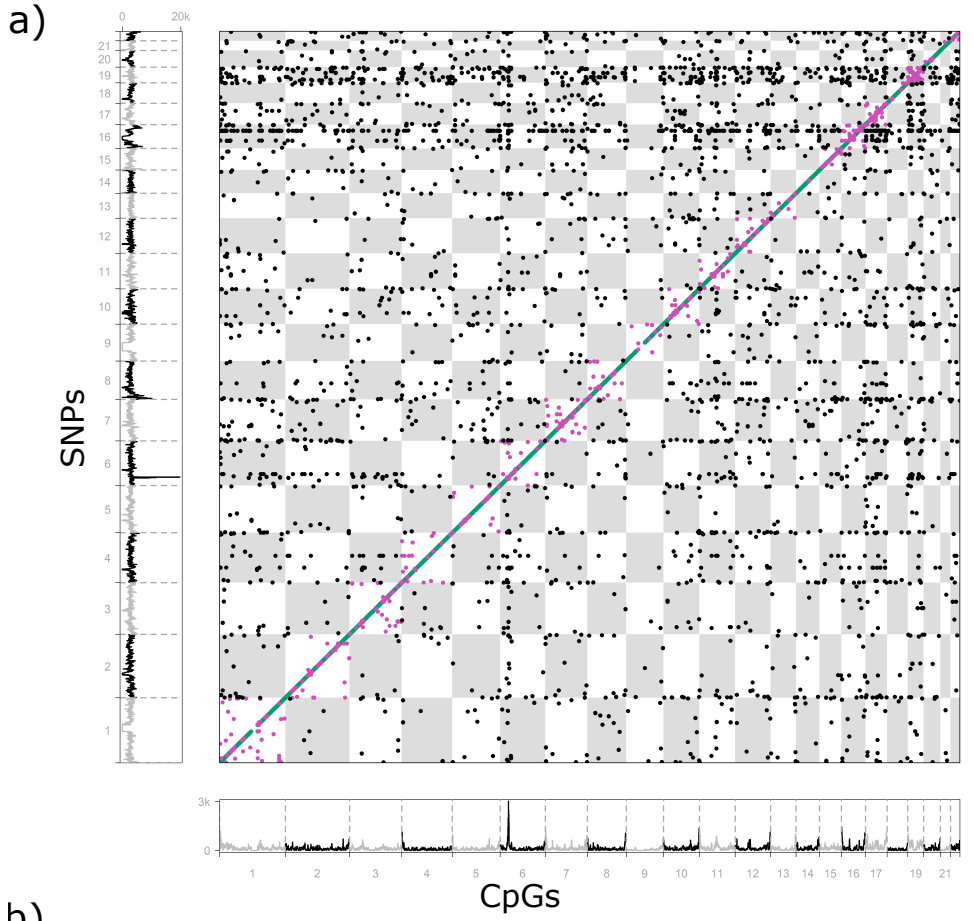
The methodology, results and figures of this project have previously been published in Hawe et al. [2] (the first part covering meQTLs and eQTM detection) or are currently prepared into a manuscript (the second part covering eQTM context analysis). All code is published in two GitHub repositories, one associated with the publication of Hawe et al. `https://github.com/heiniglab/hawe2021_meQTL_analyses`, and one for the eQTM context analysis (currently unpublished) `https://github.com/heiniglab/eqtm_prediction`.

Of note, the publication of Hawe et al. [2] contains an extensive evaluation of meQTLs. The focus in this chapter is on the parts of the analyses that I contributed to during my doctoral thesis, and does not cover the full content of the publication.

3.2. Association between genetic variants and DNA methylation

3.2.1. Genome-wide identification of meQTLs

We focused for the meQTL analysis on a set of cosmopolitan meQTLs that could be identified in both the KORA and the LOLIPOP cohort, which means that the meQTLs are present in samples of both the European and South Asian ancestries (see method section


a)
SNPs
CpGs
b)
Associated CpGs per SNP
Chromosome
CTCF
TBX6
LINC00273
SENP7
GABBR1
NFKBIE
NFKB1
CHD7
REST
ZNF676
STXBP4
ZNF607

3.6.2). With our stringent approach, we identified 11,165,559 cosmopolitan meQTLs in total (Figure 3.1 a). Most of them are cis meQTLs (n=10,346,172 with < 1Mb distance), but also several trans meQTLs (n=467,915) were identified, where the SNP and the CpG lay on different chromosomes. The trans meQTLs contained several so-called hotspots, where one SNP is associated with a high number of CpGs across the whole genome (Figure 3.1 b).

The meQTLs could be replicated in a small cohort (n=57 samples) of four different isolated white blood cell types, namely CD4+ and CD8+ T cells, monocytes and neutrophils, with replication rates of 26-37% (Supplementary Figure A.1). On top of that, many meQTLs seemed to be rather general instead of tissue specific, as 44% of meQTLs were replicable in adipose tissue (n=603 samples), 19% in subcutaneous adipocytes (n=47) and 19% in visceral adipocytes (n=47). Of note, the different sample sizes of the replication cohorts are expected to have a strong influence on the replication rates, as they affect the detection power. This can give an explanation for the higher replication rate of the larger cohort, measured in adipose tissue.

We showed the functional relevance of the meQTLs in several enrichment analyses, including enrichment for other QTLs, different genomic annotations and phenotypic traits. Additional follow-up analyses included an evaluation of differences between cis

Figure 3.1. *(preceding page)*: **Summary of results for genome-wide association and replication testing**
a Chessboard plot. Each dot represents a unique SNP–CpG pair reaching genome-wide significance in discovery ($p < 10^{-14}$) and showing both ancestry-specific and cross-ancestry replication. CpG position and background CpG density (450K array) are annotated on the x axis, and SNP position and background SNP density are annotated on the y axis. SNP–CpG pairs are color coded according to proximity of the SNP and the CpG site: cis, within 1 Mb (n = 10,346,172, green markers appearing as a diagonal line); long-range cis, distance >1 Mb but on the same chromosome (n = 351,472, purple markers); trans, SNP and CpG sites are on different chromosomes (n = 467,915, black markers). **b** Manhattan plot of trans acting SNP–CpG associations. Each marker represents the number of CpG sites associated in trans with the identified trans acting SNPs. Results are for the cosmopolitan set of SNP–CpG pairs, showing both ancestry-specific and cross-ancestry replication. SNPs with the highest number of CpG sites in trans (top 1%) are highlighted in dark blue, and the gene nearest the sentinel SNP is displayed. Figure and legend taken from [2].

and trans meQTLs and a random walk approach to identify the regulatory networks around trans meQTL hotspots. All these analyses can be found in the publication [2], but are not further elaborated in this thesis.

3.2.2. Interaction meQTLs

Another aspect of the meQTLs, which we further explored, is how the associations are affected by the cell types, as the DNA methylation itself is very cell type specific. For this, we mapped interaction QTLs (iQTLs) with the cell type proportions of CD4+ T cells, CD8+ T cells and monocytes as interaction terms (more in section 3.6.4). The cell type proportions of each individual were estimated from their DNA methylation patterns using the Houseman algorithm [48]. Likewise, DNA methylation and so potentially also the meQTLs can be influenced by environmental factors. For this reason, we additionally analyzed iQTLs with BMI and cigarette smoking as interaction term, two important environmental factors influencing DNA methylation [120, 131]. Due to the increased multiple-testing burden of the interaction models, we tested two groups of CpG-SNP pairs for each interaction term, first a targeted iQTL analysis focusing on the cosmopolitan meQTLs and then a global cis iQTL analysis for all CpG-SNP pairs within cis distance of 1 Mb.

The targeted analysis identified overall 139,552 iQTLs, which were associated with one of the five interaction terms in one cohort and could be replicated in the second cohort (Table 3.1). Most of them were associated with one of the cell types, especially with CD8+ T cells (96,455 iQTLs). The cell proportion iQTLs showed overall high replication rates across cohorts (between 25.8% and 89.4%, Figure 3.2 a). In contrast, only few iQTLs associated with BMI or smoking could be identified, and the replication rates were drastically lower (between 11.2% and 29.5%).

The global iQTL analysis, which was performed across all CpG-SNP pairs within cis distance, found with a total of 16,135 iQTLs far fewer significant associations than the targeted analysis of cosmopolitan meQTLs (Table 3.2). The reason for the fewer iQTLs is probably the far stricter significance threshold after Bonferroni correction caused by the large number of tests. The replication rate across cohorts remained high for the iQTLs associated with CD8+ and CD4+ T cells. Of note, 64% of the global iQTLs were independent of the cosmopolitan iQTLs (LD $R^2 < 0.2$). This shows that in a more well-powered iQTL analysis, more iQTLs are expected to be found than the cosmopolitan iQTL set.

The iQTLs with a cell type proportion as interaction term are an indication for cell type specific meQTLs, which are predominately present in one or a few cell types. When we checked the replication rate of these iQTLs in the corresponding isolated cell type, we obtained high replication rates between 59.9% and 70.4% for all tested cell types (Figure 3.2 b), supporting this assumption.

The importance of iQTLs for understanding phenotypic variation could be shown in GWAS enrichment analyses: the cell type associated iQTLs were enriched for several GWAS associations, among them many blood cell traits and immune traits (Figure 3.2

iQTL	Discovered (KORA)	Replicated (LOLIPOP)	Discovered (LOLIPOP)	Replicated (KORA)	Union
CD8+ T	29,128	26,052 (89.4%)	119,075	89,063 (74.8%)	96,455
CD4+ T	22,566	19,948 (88.4%)	37,638	28,225 (75%)	37,184
Monocyte	2,229	574 (25.8%)	6,504	3,557 (54.7%)	4,037
BMI	1,550	213 (13.7%)	3,647	408 (11.2%)	582
Smoking	2,496	736 (29.5%)	3,322	704 (21.2%	1,294
Total					139,552

Table 3.1.: **Number and replication rates of iQTLs identified among the cosmopolitan meQTL set.**
The number of significant iQTLs for five different iQTL phenotypes identified among the cosmopolitan meQTLs, discovered in either KORA or LOLIPOP (columns 2 & 4). Additionally, the number of replicated iQTLs in the other cohort is given (columns 3 & 5) (for iQTLs discovered in KORA, the replication was done in LOLIPOP and vice versa) and, in parentheses, the ratio of replicated iQTLs among all iQTLs. The last column 6 shows the union of replicated iQTLs from KORA and LOLIPOP. Table taken from [2].

iQTL	Discovered (KORA)	Replicated (LOLIPOP)	Discovered (LOLIPOP)	Replicated (KORA)	Union
CD8+ T	13,254	10,451 (78.9%)	8,494	5,878 (69.2%)	14,120
CD4+ T	2,010	1,160 (57.7%)	2,516	998 (39.7%)	1,714
Monocyte	7,091	224 (3.2%)	4,569	77 (1.7%)	301
BMI	3,155	96 (3.0%)	3,713	43 (1.2%)	139
Smoking	2,345	123 (5.2%)	11,821	280 (2.4%)	403
Total					16,677

Table 3.2.: **Number and replication rates of iQTLs identified in a global cis analysis of all CpG-SNP pairs.**
The number of significant iQTLs for five different iQTL phenotypes identified in a global cis analysis of all CpG-SNP pairs, discovered in either KORA or LOLIPOP (columns 2 & 4). Additionally, the number of replicated iQTLs in the other cohort is given (columns 3 & 5) (for iQTLs discovered in KORA, the replication was done in LOLIPOP and vice versa) and, in parentheses, the ratio of replicated iQTLs among all iQTLs. The last column 6 shows the union of replicated iQTLs from KORA and LOLIPOP. Table taken from [2].

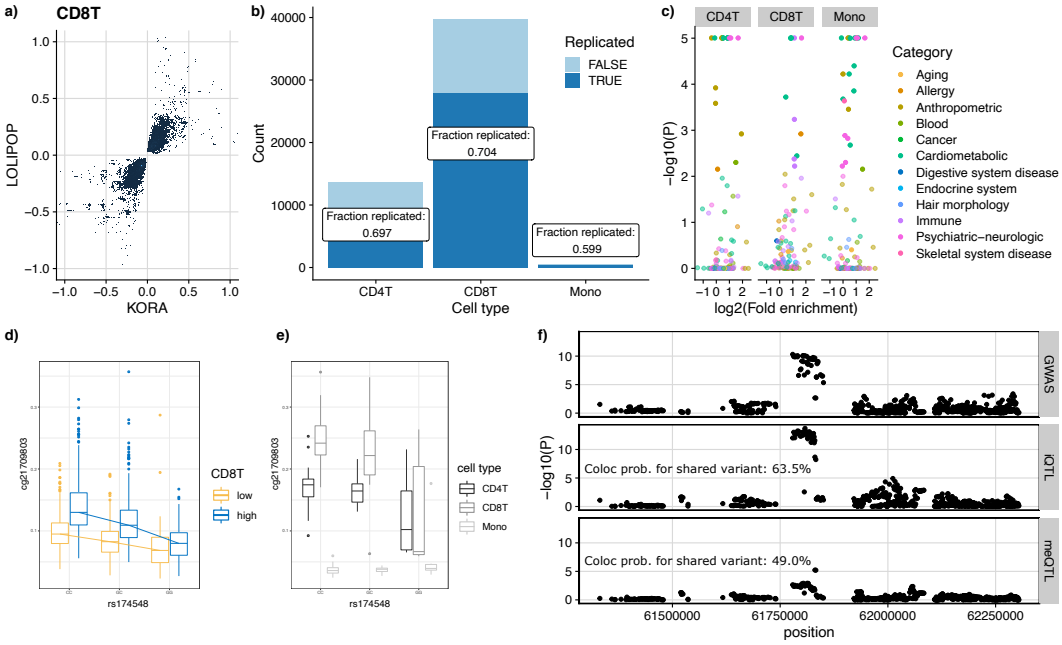


Figure 3.2.: **White cell iQTL**

a, Replication of effect sizes of significant iQTL (CD8+ T cells) between KORA and LOLIPOP cohorts. Axes show genotype–cell type interaction effect sizes, points individual associations. **b**, Replication of iQTL in isolated cells. The y-axis shows the total number of associations, the x-axis the respective cell types. Colors indicate the proportion of replicating associations (dark blue) and non-replicating associations (light blue). **c**, Enrichment of iQTL SNPs with GWAS information in diverse traits. The y-axis shows -log10 values of *QTLEnrich* P values, the x-axis the log2 fold enrichment of observed GWAS SNPs among iQTL compared to expected values. Plots are split by analyzed cell types. Points reflect individual GWAS (colors represent the phenotype category). **d**, Association plot for the rs174548-cg21709803 iQTL in KORA data, separated into individuals with 'high' and 'low' abundance (above and below the median) of CD8+ T cells. The y-axis indicates methylation residuals, the x-axis the genotypes. Box plots indicate medians (center lines) and first and third quartiles (lower and upper box limits; whisker extents, 1.5-fold of interquartile ranges). Points indicate outliers. **e**, The same association plot as in **d** but using data from isolated cells (different shades of gray). **f**, Manhattan plot of meQTL, asthma GWAS and iQTL results for the selected iQTL example show colocalization of association signals. The x-axis indicates the genomic region around rs174548, the y-axis the -log10 of association P values. Individual points represent SNPs in the locus. *Coloc* prob., colocalization probability. Figure and legend taken from [2].

c). One interesting example here is the iQTL of rs174548-cg21709803 with higher effect sizes in individuals with high CD8+ T cells proportion (Figure 3.2 d). The analysis of the isolated cell types identified the meQTL only in T cells, not in monocytes, supporting the specificity of this meQTL (Figure 3.2 e). The SNP rs174548 is associated with several metabolites in the fatty acid metabolism [132, 133], matching the position of the SNP in the gene *FADS1* (Fatty Acid Desaturase 1), but also with blood eosinophil counts [134] and inflammatory diseases such as asthma [135]. The connection between the asthma GWAS signal and the iQTL was strengthened in a colocalization analysis, which indicated a shared causal variant (coloc PP4=0.63, Figure 3.2 e). Importantly, this SNP is no cosmopolitan meQTL, the potential relationship of the meQTL rs174548-cg21709803 with asthma and its T cell specificity could only be made due to the iQTL analysis.

3.3. Association between DNA methylation and gene expression

In the previous sections, we successfully quantified the relationship between DNA methylation and genetics using meQTLs and iQTLs. Next, we included the transcriptomic information from the cohorts as an additional omics layer to get a more complete picture of the role of DNA methylation. To analyze the relationship of DNA methylation and gene expression, we performed an eQTM meta analysis on KORA and LOLIPOP and classified the detected eQTMs again as cis and trans dependent on their distance.

We ran the eQTM analysis twice with different settings to explore the effect of the cell type composition: once we corrected the expression and methylation data for basic covariates and genotype effects and once we additionally corrected both for the cell type proportions of the individuals (see section 3.6.6). With this strategy, we observed that the number of identified eQTMs was strongly driven by the cell type proportions. Without correction for cell type proportions, we found with 90,666 cis eQTMs and 54,807,559 trans eQTMs, called GTA eQTMs for "genotype adjusted" in the following (Table 3.3). The additional correction reduced the numbers to 769 cis eQTMs and 97,281 trans eQTMs, called CPA eQTMs for "genotype and cell proportion adjusted" in the following. The drastically lower number of CPA eQTMs compared to GTA eQTMs highlights that the role of DNA methylation in establishing cell type identity: GTA eQTMs, which are removed after the correction, represent differences between cell types rather than differences between individuals, which become visible due to different cell type composition in the different individuals.

For the stricter set of CPA eQTMs, we validated the replication rates between cohorts: we saw high replication both for the eQTMs identified in KORA and the eQTMs identified in LOLIPOP, especially when applying the stricter family-wise error rate (FWER) adjustment compare to false discovery rate (FDR) adjustment for multiple testing correction (Table 3.4, more about multiple testing in Method section 2.4).

	GTA eQTMs	CPA eQTMs
cis eQTMs	90,666	769
trans eQTMs	54,807,559	97,281

Table 3.3.: **Cis and trans eQTMs from meta analysis of cohorts.**
Number of cis and trans eQTMs from the meta analysis of KORA and LOLIPOP together, once without correction for cell type proportions (GTA eQTMs) and once with the correction (CPA eQTMs). Significance threshold of $p<0.05$ after FWER correction (Bonferroni). Table taken from [2].

		KORA meQTLs		LOLIPOP meQTLs	
	MT corrected	Discovered (KORA)	Replicated (LOLIPOP)	Discovered (LOLIPOP)	Replicated (KORA)
cis	FDR	1,006	684 (68%)	2,445	1,317 (54%)
trans	FDR	213,182	81,684 (38%)	766,434	266,367 (35%)
cis	FWER	301	246 (82%)	409	352 (86%)
trans	FWER	20,989	10,993 (52%)	69,497	49,943 (72%)

Table 3.4.: **Replication of eQTMs across cohorts**
Number of eQTMs, either discovered in KORA and replicated in LOLIPOP (called KORA eQTMs) or discovered in LOLIPOP and replicated in KORA (called LOLIPOP eQTMs) with FDR or FWER multiple testing correction. Focusing on CPA eQTMs here. Table adapted from [2].

3.4. Identifying the context-specific relationship of eQTMs

We extended our analysis of the GTA eQTMs and CPA eQTMs with the goal to better understand the differences between both classes and more broadly the general mechanisms behind the associations of gene expression and DNA methylation. For this reason, we explored the genomic context of both eQTM classes, for example the position of the CpG, the distance between the CpG and the gene or the promoter type of the gene (full set of tested annotations found in Table A.1). Using different machine learning (ML) models, we tested whether we can predict the eQTM probability based on the genomic context as features and which aspects were most important for the prediction.

3.4.1. Predicting the eQTM probability of a CpG-gene pair

During this evaluation, we focused on the set of cis eQTMs, with at least 1Mb between the CpG and TSS of the gene. This allowed us to generate a more well-defined data set for all downstream analyses. We used the eQTMs identified separately in both cohorts (KORA and LOLIPOP) and repeated the multiple testing correction specifically for the cis associations (adjusted eQTM found in Table 3.5). As described in the last section 3.3, we distinguished again between GTA eQTMs, which were adjusted for basic covariates and the genotypes, and CPA eQTMs, which were additionally adjusted for the cell type proportions. In the following, we will additionally compare positively associated eQTMs (effect size > 0) with negatively associated eQTMs (effect size < 0). While the proportions between negative and positive eQTMs are very similar for the GTA eQTMs, for the CPA eQTMs, a higher fraction of negative eQTMs was found. Replication rates between cohorts remained high, as already observed in the previous section 3.3. The GTA eQTMs showed overall slightly higher rates than the CPA eQTMs (Table 3.5).

Previous studies found distinct genomic features of eQTMs, such as an enrichment of eQTM CpGs in promoter and enhancer regions, as well as differences between positively and negatively associated eQTMs [38]. We can reproduce many of these findings: for example, the distance between the CpG and the TSS of the gene is on average shorter in negative eQTMs compared to positive eQTMs and non-significant pairs. Furthermore, a higher fraction of CpGs lies in enhancer regions for eQTMs compared to non-significant pairs (Supplementary Figure A.2). However, looking at individual features is insufficient for a general, comprehensive model, describing the context-sensitivity of eQTMs.

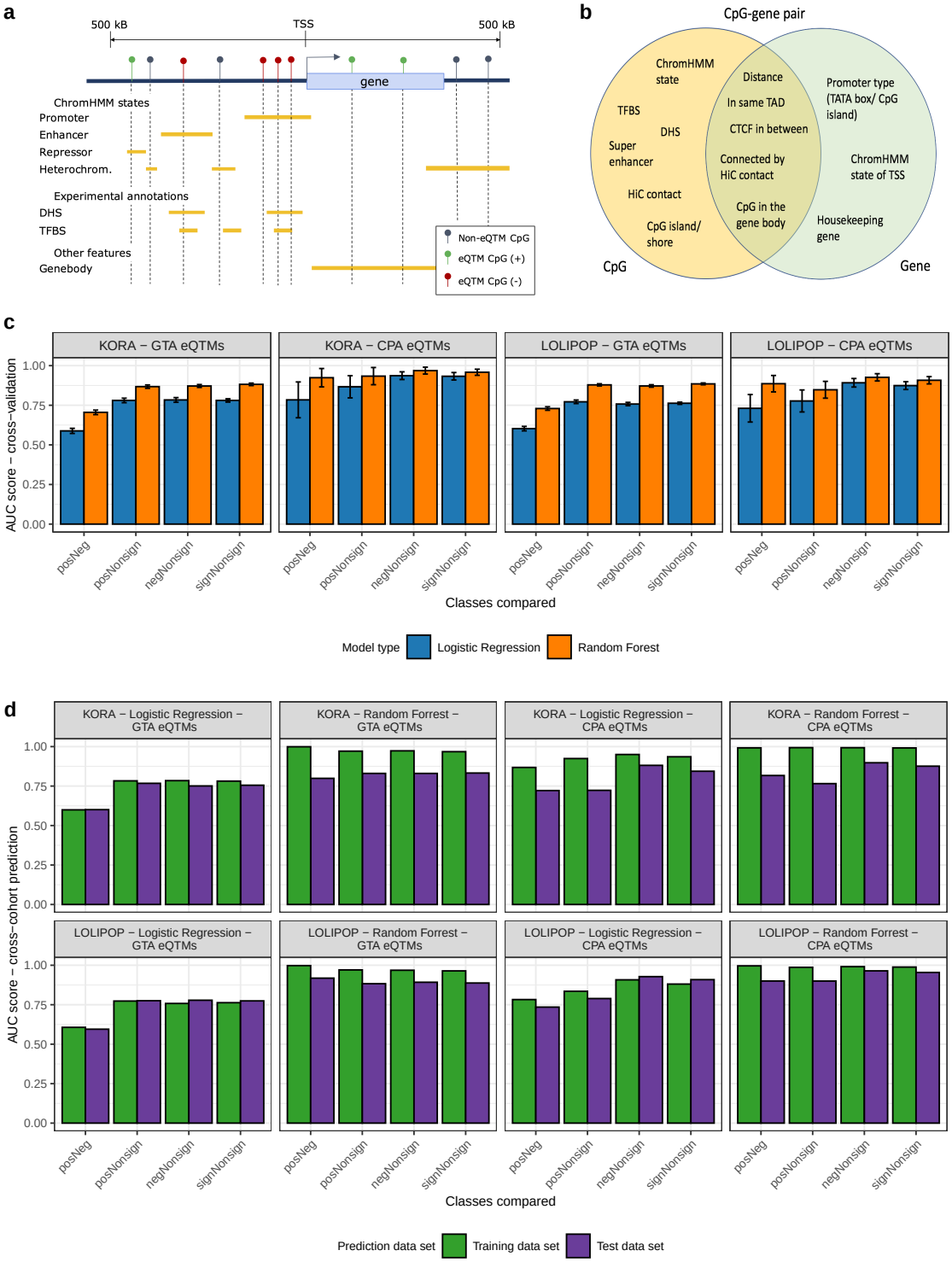
To combine and compare the different genomic annotations, we applied ML models, extending the idea of Bonder et al. [38], who built a model to predict the direction of effect for eQTMs based on histone modifications. Our models predict whether a CpG-gene pair is an eQTM or not based on a large set of genomic annotations of the pair. Two aspects can be analyzed with the model: first, a good prediction performance indicates that the eQTM probability of a CpG-gene pair is influenced by its genomic context. Second, by evaluating the feature importance of the model, we can distinguish which genomic annotations are more important for defining the genomic context of an eQTM.

	MT correction	KORA eQTMs GTA	KORA eQTMs CPA	LOLIPOP eQTMs GTA	LOLIPOP eQTMs CPA
Tested pairs		7,721,357	7,721,357	7,721,357	7,721,357
eQTMs	< 0.05	243,842 (72.2%)	1,509 (57.7%)	375,128 (59.0%)	3,389 (47.5%)
eQTMs	< 0.01	158,478 (82.4%)	914 (68.8%)	253,663 (70.0%)	1,810 (61.8%)
positive eQTMs	< 0.05	118,197 (72.0%)	421 (34.2%)	186,020 (58.2%)	1,106 (35.0%)
positive eQTMs	< 0.01	76,397 (82.0%)	205 (44.9%)	125,432 (69.2%)	463 (46.7%)
negative eQTMs	< 0.05	125,645 (72.3%)	1,088 (66.7%)	189,108 (59.8%)	2,283 (53.6%)
negative eQTMs	< 0.01	82,081 (82.7%)	709 (75.7%)	128,231 (70.8%)	1,347 (67.0%)

Table 3.5.: **Cis eQTMs per cohort, split after direction of effect.**
Identification of cis eQTMs in the KORA and LOLIPOP cohort (FDR < 0.05 or FDR < 0.01) with a cis distance of 1MB between TSS of the gene and the CpG. Percentages below the numbers show the replication rates in the other cohort (for eQTMs discovered KORA the replication in LOLIPOP and vice versa). Again, GTA and CPA eQTMs were distinguished and, additionally, positively correlated eQTMs ($\beta > 0$) and negatively correlated eQTMs ($\beta < 0$).

We selected annotations for the prediction that describe the CpG site and the gene individually as well as the pair together (Figure 3.3 a,b). A full list of all annotations can be found in Supplementary Table A.1. The position of the CpG was characterized by annotating its chromatin state and whether there was a transcription factor binding site (TFBS), a DNAse hypersensitivity site (DHS), a super-enhancer, a HiC contact or a CpG island. Each of the features is binary, as the CpG is either inside the genomic annotation or not, but the annotations differ between cell types. To apply our model on bulk data containing a mixture of different cell types, such as whole blood, we transformed the features into a weighted score based on the cell type proportions of the data set (section 3.6.7). For the gene, we annotated the promoter type, the chromatin state of its TSS and whether it is a housekeeping gene. The CpG-gene pair together was described by the distance between the CpG and the TSS of the gene, the relative position of the CpG to the gene (before, inside or after the gene) and whether they are within the same topological associated domain (TAD).

For our prediction models, we considered all CpG-gene pairs which we tested for association in cis, i.e. all pairs with a maximum distance of 1Mb and measurements for


a
500 kB
TSS
500 kB
gene
ChromHMM states
Promoter
Enhancer
Repressor
Heterochrom.
Experimental annotations
DHS
TFBS
Other features
Genebody
Non-eQTM CpG
eQTM CpG (+)
eQTM CpG (-)
b
CpG-gene pair
ChromHMM state
TFBS
DHS
Super enhancer
HiC contact
CpG island/ shore
Distance
In same TAD
CTCF in between
Connected by HiC contact
CpG in the gene body
Promoter type (TATA box/ CpG island)
ChromHMM state of TSS
Housekeeping gene
CpG
Gene
c
KORA − GTA eQTMs
KORA − CPA eQTMs
LOLIPOP − GTA eQTMs
LOLIPOP − CPA eQTMs
AUC score − cross−validation
1.00
0.75
0.50
0.25
0.00
posNeg
posNonsign
negNonsign
signNonsign
Classes compared
Model type
Logistic Regression
Random Forest
d
KORA − Logistic Regression − GTA eQTMs
KORA − Random Forrest − GTA eQTMs
KORA − Logistic Regression − CPA eQTMs
KORA − Random Forrest − CPA eQTMs
LOLIPOP − Logistic Regression − GTA eQTMs
LOLIPOP − Random Forrest − GTA eQTMs
LOLIPOP − Logistic Regression − CPA eQTMs
LOLIPOP − Random Forrest − CPA eQTMs
AUC score − cross−cohort prediction
Classes compared
Prediction data set
Training data set
Test data set

both expression and methylation. We explore in our study both the difference between eQTMs and non-significant pairs and on a more fine-grained level the difference between positive and negative eQTMs. For this reason, we modeled four different two-class comparison problems: we compared all significant eQTMs against the non-significant pairs, then the positive and negative eQTMs separately against the non-significant pairs and as a fourth option the positive eQTMs against the negative eQTMs. We applied Logistic Regression (LR) as a baseline model for the classification and Random Forest (RF) as a more sophisticated model, capturing also non-linear relationships.

For the prediction, we generated a data set of high confidence eQTMs with strict FDR correction and filtered for high variance of methylation levels. The motivation for the second filter was that only for CpGs with sufficient variation between samples, the class can be assigned reliably. CpGs with (nearly) constant methylation could still correlate with gene expression, but this is not visible in our dataset. For the threshold selection, we trained LR and RF models with different thresholds for the prediction of eQTMs vs non-significant pairs in a 10-fold cross-validation with the KORA dataset. When evaluating their performance, we saw a clear improvement with stricter filtering (Supplementary Figure A.3). We selected the thresholds separately for the GTA and the CPA eQTMs, so that the performance of the model increased but the set of significant eQTMs was not reduced too drastically. We set the FDR threshold to 0.01 for both eQTM sets. The variance threshold was chosen to keep only the 6.25% most variable CpGs for the GTA eQTMs and the 25% most variable CpGs for the CPA eQTMs. For these and all following models, we chose a set-up with balanced classes by subsampling the non-significant eQTMs, as otherwise the sensitivity of the model dropped drastically.

With these filtering criteria, all models performed well both cross-validations within each cohort and in predictions across the cohorts (Figure 3.3 c,d). The cross-validation results showed that the differentiation between non-significant CpG-gene pairs and

Figure 3.3. *(preceding page)*: **Predicting eQTMs with different ML models.**
a Schematic description how the different eQTM classes are annotated with example features. **b** Full set of all features used for the eQTM prediction. **c** Performance of the the Logistic Regression and the Random Forrest models on a 10-fold cross validation within the KORA and LOLIPOP cohort. Four different two-class comparisons were performed each time (posNeg: positive eQTMs vs negative eQTMs, posNonsign: positive eQTMs vs non-significant pairs, negNonsign: negative eQTMs vs non-significant pairs and signNonsign: all eQTMs against the non-significant), separately for the GTA eQTMs and the CPA eQTMs. **d** Cross-cohort performance between KORA and LOLIPOP, using the same settings as in **c**.

significant eQTMs (or a subclass of it) was easier for both ML models than the comparison of positive eQTMs against negative eQTMs (Figure 3.3 c), visible in higher AUC scores. The best average performance in the cross-validation was reached for the RF model with CPA eQTMs from KORA with an AUC of 95.8% for all eQTMs vs non-significant pairs and an AUC of 96.9% for negative eQTMs vs non-significant pairs. In contrast, the worst performance was reached for the LR model with GTA eQTMs from KORA with an AUC of 58.8%. However, with Random Forest and the CPA eQTMs, the prediction of positive vs negative eQTMs reached AUC values around 90%. In general, the Random Forest classifiers performed better than the Logistic Regression models, which was expected from the more sophisticated RF models. Furthermore, models with CPA eQTMs performed better than the models with GTA eQTMs. The cross-validation results between KORA and LOLIPOP were very close in all tested scenarios.

All models generalized across cohorts, as the performance for cross-cohorts prediction remained high (Figure 3.3 d). For this, the models were trained on one cohort and then used for prediction in the second cohort, which represents an independent test set. Similar trends were visible as in the within cohort validation, e.g. that the models with CPA eQTMs performed better than the ones with the GTA eQTMs (AUC values on test set of 72.2% - 96.4% for the CPA eQTMs and 59.5%-91.8% for the GTA eQTMs) and that the RF models performed better than the LR models (AUC values on test set of 76.6% - 96.4% for the RF models and 59.5%-92.8% for the LR models). Interestingly, the LOLIPOP models performed better on the independent test set (the KORA cohort) than vice versa. The reason might be the higher number of eQTMs in LOLIPOP compared to KORA, which could improve the feature learning of the model. Of note, the performance between predictions on the training cohorts and the test cohorts were very similar for the LR models, while the prediction on the training sets performed clearly better for the RF models. This could indicate a slight overfitting of the RF models.

3.4.2. Identification of important genomic features for prediction

The good model performances, both in the cross-validations and in the cross-cohort predictions, strongly supports our hypothesis that eQTMs are context-specific. Following this, we further utilized the models to identify which of the genomic annotations are the most predictive features by evaluating the feature importance of each model. A high importance score of a feature is an indication that it plays a role in the genomic context of the eQTMs, i.e. that CpG-gene pairs with this feature are more likely associated with each other.

The feature importance for Logistic Regression was evaluated using the log odds ratios and p-values for each feature of the model (see Method section 2.3.1). The feature importance for Random Forest was measured with the Mean Decrease in accuracy (MDA), which shows how much the accuracy of the out-of-bag cross validated predictions is decreased when this specific variable is permuted. In contrast to the log odds ratios, the MDA does not show the direction of the effect, i.e. if an increase of the feature value increases or decreases the eQTM probability. For Random Forest, we

generated models with all features together (as in the previous section 3.4.1), for Logistic Regression, we trained bivariate models with each feature and the distance to evaluate each feature independent of its correlation with other features.

An important aspect of our features to keep in mind is that they are correlated (Supplementary Figure A.4). Therefore, we additionally applied feature selection to validate how much redundancy currently exists in our model and whether a reduced feature set can reach similar performance as the full model. For the feature selection, we combined different methods for Logistic Regression and Random Forest and selected all features chosen in at least 90% of the runs (details in section 3.6.10). Afterwards, we combined both the results of the feature importance and the feature selection to interpret the features (Figures 3.4 and 3.5).

The GTA model to predict eQTMs versus non-significant CpG-gene pairs reached nearly the same performance with the reduced feature set, chosen in the feature selection based on the KORA eQTMs, compared to a model with all features. The mean AUC performance of the within cohort cross-validation was for the reduced model 98% of the full model, both for RF and LR models and for both cohorts. The 13 selected features described the position of the CpG using several ChromHMM states, super-enhancer annotations, TFBS and CpG islands and whether the CpG-gene pair was within the same TAD (Figure 3.4 a). The feature selection using the LOLIPOP instead of the KORA dataset showed very similar results, with 10 of the 13 KORA features being selected (only CpG.ChromHMM.12EnhBiv, CpG.ChromHMM.2TssAFlnk.far, pair.in.same.TAD fell below the selection cutoff). In general, the feature interpretation of both cohorts brought nearly the same results, also visible in the high concordance of the feature importance scores (Figure 3.4 b,c).

Looking at the importance scores for the GTA model and the comparison of eQTMs vs non-significant pairs, the effect of the selected features can be interpreted more (Figure 3.4 a). The different enhancer annotations - the ChromHMM states enhancer and genic enhancer as well as the super-enhancer annotations - played overall an important role with high importance score in both the Random Forest and the Logistic Regression models. The positive log odds ratios show that CpGs lying in these regions have a higher probability of being an eQTM. This finding is supported by previous work that connected DNA methylation with enhancers [127, 128]. Furthermore, the chromatin structure is predictive for eQTMs, represented by pairs located in the same TAD and CpG positioned at a chromatin interaction point, both also with a positive log odds ratio. In contrast, CpGs positioned in Polycomb repressed genomic regions are also important features, but with a negative log odds ratio. Interestingly, CpG islands show a strong significant negative log odds ratios, meaning that CpGs in these regions are less likely to be part of eQTMs.

The important features of the CPA model changed completely compared to the GTA model, but continued to replicate very well between the cohorts (Figure 3.5). Fewer features were selected in the feature selection. Matching this, there were fewer significant features in the LR models and lower MDA scores in the RF models. Only 4 features

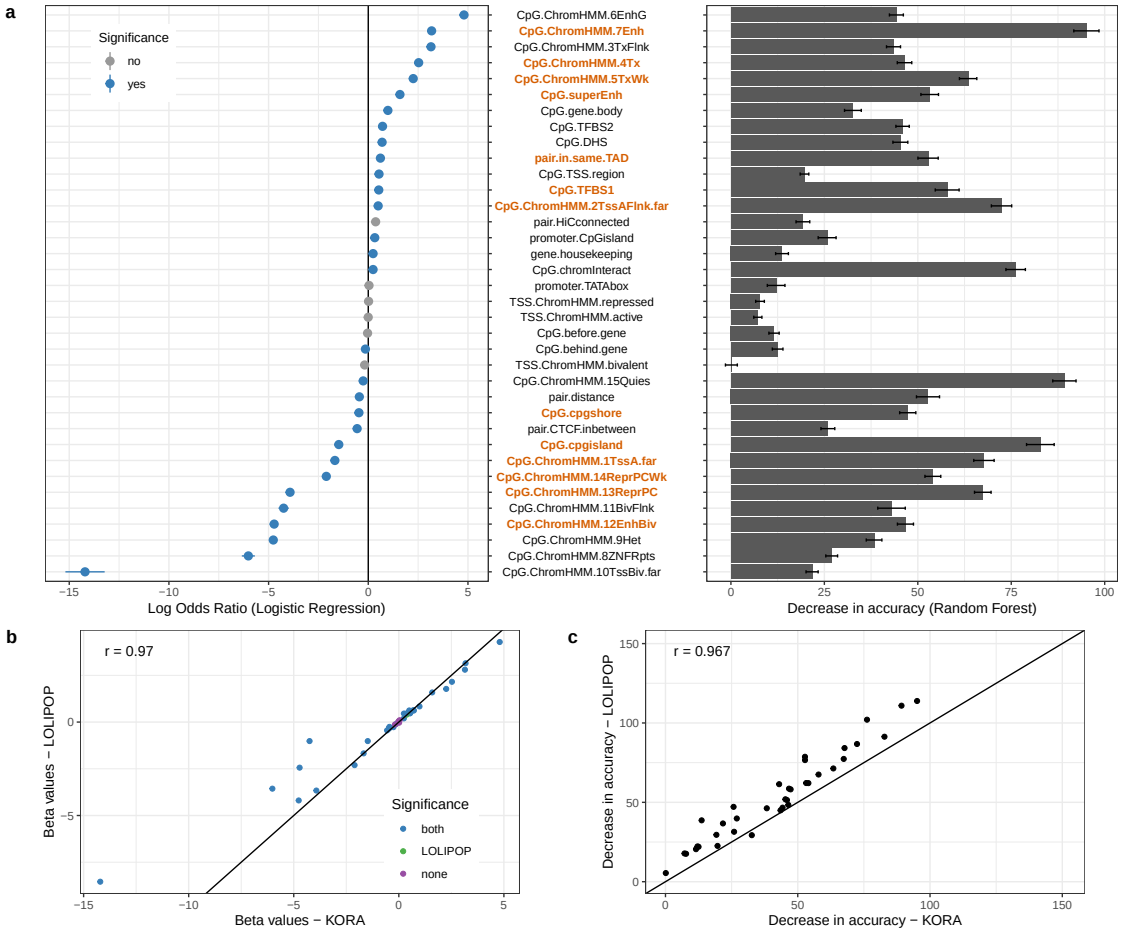
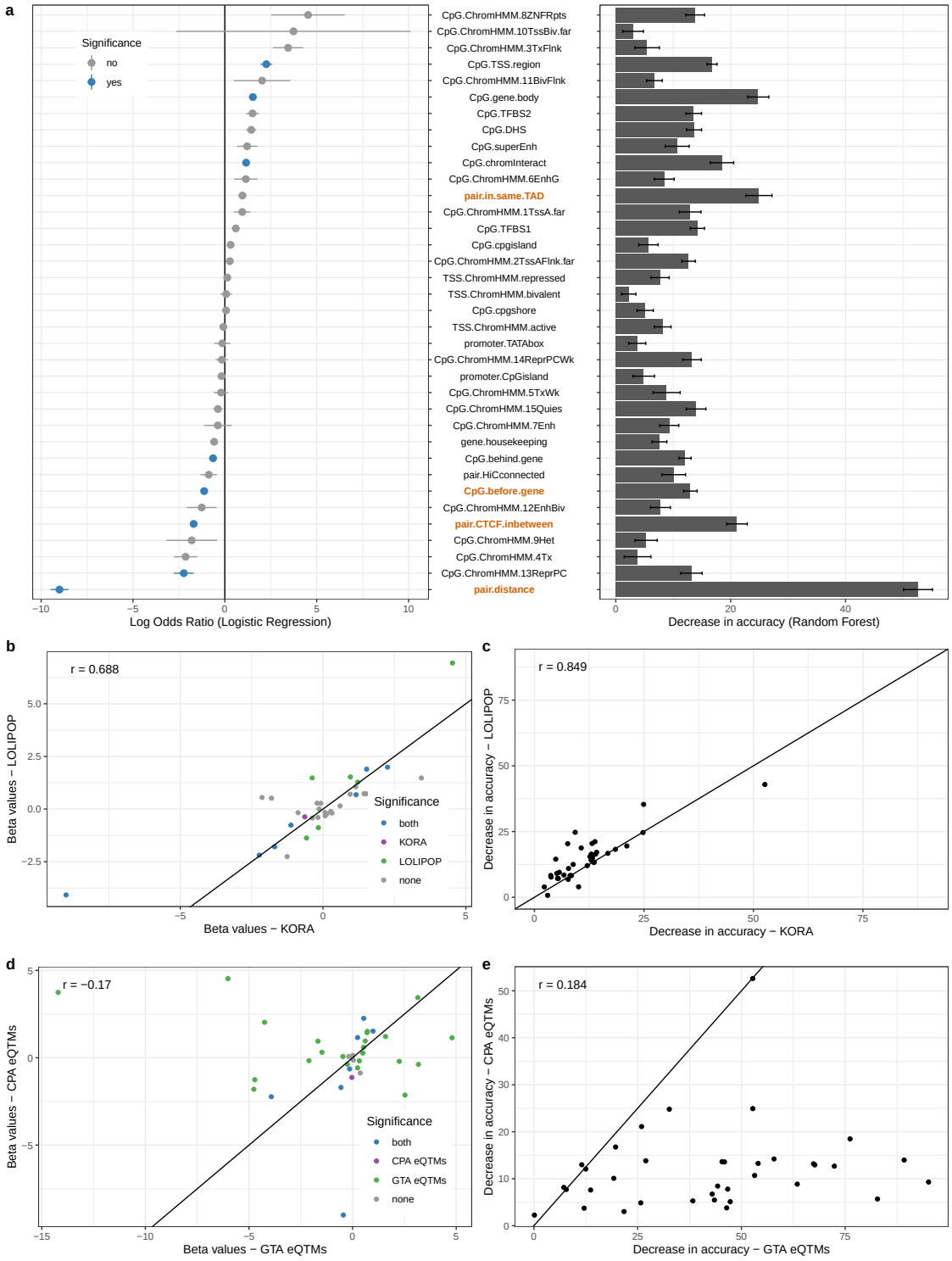


Figure 3.4.: **Feature importance of the model eQTMs vs non-significant pairs for the GTA eQTMs**

a The feature importance of the KORA models with GTA eQTMs. For the Logistic Regression, bivariate models were trained for each feature plus the distance, for Random Forest, multivariate models. Feature names marked in red were chosen in the feature selection for the reduced model (selected in at least 90% of the cases). Only features occurring in at least 1% of the CpG-gene pairs were shown here to increase visibility, a full version with all features can be found in Supplementary Figure A.5. **b,c** Replication of the feature importance in the LOLIPOP model for Logistic Regression (**b**) and Random Forest (**c**). r values (top left of each plot) are the Pearson correlations between the scores from LOLIPOP and KORA.


a
Significance
no
yes
CpG.ChromHMM.8ZNFRpts
CpG.ChromHMM.10TssBiv.far
CpG.ChromHMM.3TxFlnk
CpG.TSS.region
CpG.ChromHMM.11BivFlnk
CpG.gene.body
CpG.TFBS2
CpG.DHS
CpG.superEnh
CpG.chromInteract
CpG.ChromHMM.6EnhG
pair.in.same.TAD
CpG.ChromHMM.1TssA.far
CpG.TFBS1
CpG.cpgisland
CpG.ChromHMM.2TssAFlnk.far
TSS.ChromHMM.repressed
TSS.ChromHMM.bivalent
CpG.cpgshore
TSS.ChromHMM.active
promoter.TATAbox
CpG.ChromHMM.14ReprPCWk
promoter.CpGisland
CpG.ChromHMM.5TxWk
CpG.ChromHMM.15Quies
CpG.ChromHMM.7Enh
gene.housekeeping
CpG.behind.gene
pair.HiCconnected
CpG.before.gene
CpG.ChromHMM.12EnhBiv
pair.CTCF.inbetween
CpG.ChromHMM.9Het
CpG.ChromHMM.4Tx
CpG.ChromHMM.13ReprPC
pair.distance
−10
−5
0
5
10
Log Odds Ratio (Logistic Regression)
0
20
40
Decrease in accuracy (Random Forest)
b
r = 0.688
Beta values − LOLIPOP
Beta values − KORA
Significance
both
KORA
LOLIPOP
none
c
r = 0.849
Decrease in accuracy − LOLIPOP
Decrease in accuracy − KORA
d
r = −0.17
Beta values − CPA eQTMs
Beta values − GTA eQTMs
Significance
both
CPA eQTMs
GTA eQTMs
none
e
r = 0.184
Decrease in accuracy − CPA eQTMs
Decrease in accuracy − GTA eQTMs

exceeded the selection threshold (selected in at least 90% of the runs) compared to 13 features before, when analyzing the GTA eQTMs. From the selected features, only the pair.in.same.TAD was selected in both the GTA and the CPA model. The CpG annotations with ChromHMM states and CpG islands/shores were less important in the CPA model. Instead, more structural features were selected, namely CTCF binding sites between the pair, the distance of the pair and whether the CpG is before the gene. The performance of the models with the reduced feature set reached still at least 93% of the performance from the full model with all features.

The feature importance for the positive versus negative eQTM models needs to be interpreted more carefully, as the models did not perform so well. This indicates that the models are probably not capturing the biological context that well, compared to the models for eQTMs vs non-significant pairs. One reason for this might be the feature set, as all features represent broad categories and, for example, TF binding sites could be both associated with activating and repressing TFs. Nevertheless, we looked at the feature importance scores both for the GTA model and the CPA models (Supplementary Figures A.7 and A.8). In contrast to the eQTM vs non-eQTM predictions, the GTA and CPA models showed more similar feature importance scores, when predicting positive vs negative eQTMs. For both models, the distance between the CpG-gene pair got the highest importance score for the RF models. The positive log odds ratio of this feature shows that pairs further apart from each other are more likely to be positively correlated. Other important features were that the CpG is positioned in a TSS state from

Figure 3.5. *(preceding page)*: **Feature importance of the model eQTMs vs non-significant pairs for the CPA eQTMs**
a The feature importance of the KORA models with CPA eQTMs. For the Logistic Regression, bivariate models were trained for each feature plus the distance, for Random Forest, multivariate models. Feature names marked in red were chosen in the feature selection for the reduced model (selected in at least 90% of the cases). Only features occurring in at least 1% of the CpG-gene pairs were shown here to increase visibility. A full version with all features can be found in Supplementary Figure A.6. **b,c** Replication of the feature importance in the LOLIPOP model for Logistic Regression (**b**) and Random Forest (**c**). r values (top left of each plot) are the Pearson correlations between the scores from LOLIPOP and KORA. **d,e** Comparison of the feature importance between the GTA and the CPA model for Logistic Regression (**d**) and Random Forest (**e**). r values (top left of each plot) are the Pearson correlations between the scores from the GTA and CPA models.

the ChromHMM model (active TSS, TSS flanking and bivalent TSS), which increases the probability of negative correlation (matching the distance variable) and that the pair is in the same TAD / no CTCF binding site is in between, which increases again the probability of negative correlation.

3.4.3. Replication across tissues

After we evaluated our eQTM models in different blood datasets, we evaluated how generalizable our eQTM prediction models are across tissues. For this, we explored whether the prediction of eQTMs also performs well in other tissues, which do no match the training dataset. We applied a model trained on the whole blood CPA eQTMs from LOLIPOP to predict eQTMs in a data set from skeletal muscle [136]. When we collected the necessary muscle-specific annotations, we restricted the feature set slightly, as not all features available for whole blood could be found for skeletal muscle (see section 3.6.11).

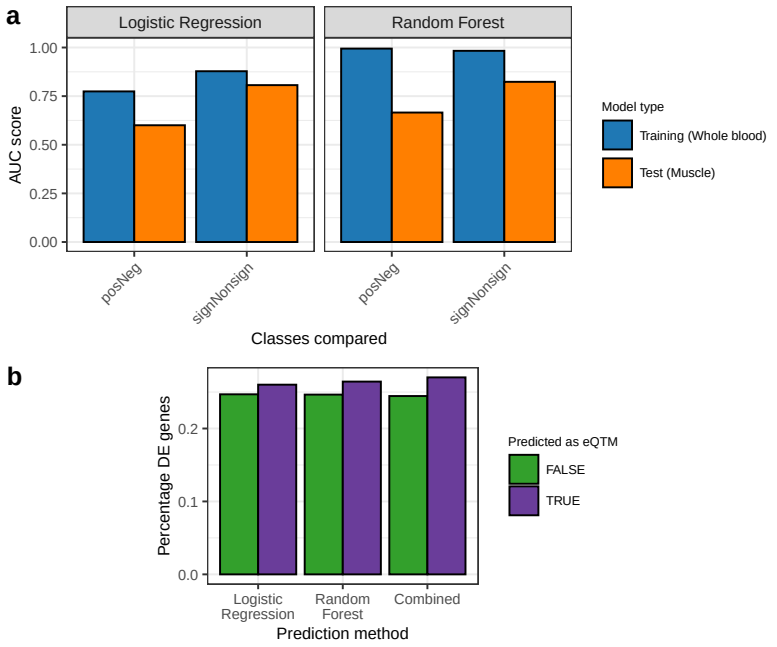


Figure 3.6.: **Prediction performance cross-tissues and for prioritizing EWAS genes** **a** Prediction performance of the eQTM model trained on the whole blood dataset (LOLIPOP CPA eQTMs) for the prediction of muscle eQTMs. **b** Enrichment of DE genes after muscle training among genes, predicted to be associated with EWAS CpGs from muscle training. Three different gene sets were considered, either predicted genes from the Logistic Regression and Random Forest eQTM model alone or the combined set of genes that was predicted by both models.

The cross-tissue prediction reached an AUC value of 80.7% for the LR model and of 82.4% for the RF model when predicting eQTMs against non-significant pairs (Figure 3.6 a). The good prediction performance of the skeletal muscle eQTMs proves that the model captures general characteristics of eQTMs, which are not only specific for whole blood. Importantly, the overlap between the whole blood eQTMs and the skeletal muscle eQTMs is very small (only 19 eQTMs from 25,867 muscle eQTMs are also in the filtered LOLIPOP CPA eQTM set). This highlights that the model can identify new tissue type specific eQTMs based on the corresponding annotations, even when it was trained on a different tissue. The performance of the cross-tissue prediction for positive vs negative eQTMs was lower, with AUC values of 60.0% for the Logistic Regression and 66.7% for Random Forest (Figure 3.6 a). This is in line with the general lower performance of these models.

3.4.4. Application of eQTM model for the identification EWAS hits

Additional to the biological interpretation of eQTMs, our model gives us the ability to predict candidate genes from Epigenome-wide association studies (EWAS). EWAS associate CpGs with phenotypic traits, such as diseases. Connecting the EWAS CpGs with genes is an important step in the interpretation of EWAS results. With our model, this is possible solely based on EWAS summary statistics to select the significantly associated CpGs and cell type specific genomic annotations as input for our model.

We showcased this application using an EWAS study that measured the effect of endurance exercise training on DNA methylation in skeletal muscles [137]. As the study additionally measured differential expressed genes between trained and untrained muscle, we have a ground truth of genes that we expect to be connected with the EWAS CpGs. We applied our eQTM model with skeletal muscle annotations and combined the prediction of Logistic Regression and Random Forest, taking only genes predicted in both models. This resulted in significant enrichment of the predicted eQTM genes among the DE genes with an odds ratio of 1.14 ($p=0.0019$) (Figure 3.6 b). The predictions of the Logistic Regression and the Random Forest models alone showed also positive odds ratios, but only the Random Forest odds ratio was significant ($p=0.026$). The predicted eQTM genes were enriched for the GO terms of "actin binding" and "transmembrane signaling receptor activity" (FDR<0.05), matching the biological expectations after muscle training. The connection with actin was also found in the original study [137].

This shows that our model can help to interpret EWAS results by connecting CpGs to genes, which can then be mapped to specific biological pathways and functions. Required annotations for the model are public available, leading to a quick evaluation of the results without the need to generate further experimental data.

3.5. Project summary and outlook

In this project, we showed how to analyze DNA methylation cohort data to better understand the interactions between DNA methylation and genetic variants (meQTLs) and DNA methylation and gene expression (eQTMs), both providing new helpful insights into regulatory processes. The comprehensive catalog of meQTL and eQTM associations generated from our large multi-ethnic cohort is a useful resource for all further studies and available online at `http://qtldb.helmholtz-muenchen.de`.

In order to identify cell type and context specific effects in a bulk DNA methylation cohort, we applied an interaction model for meQTLs with cell type proportions and environmental factors as interaction terms. We proved so the existence of several cell type and context dependent iQTLs and showed in the following also the relevance of these iQTLs for interpreting disease variants. The small set of iQTLs in the genome-wide cis analysis compared to the targeted analysis showed that our analysis was potentially still underpowered. This can be overcome in the future by larger sample sizes or the use of single cell data, as explained in the following two chapters 4 and 5.

The strong influence of the cell type proportions was even more visible during the detection of the eQTMs, as the set of genotype and cell proportion adjusted (CPA) eQTMs was drastically smaller than the genotype adjusted (GTA) eQTM set. The eQTM prediction based on genomic features proved that GTA eQTMs and CPA eQTMs both have a specific genomic context that drives the associations, which differs between both classes. Importantly, these features are generalizable across cohorts and tissues. For example, the GTA eQTMs were connected with enhancer methylation, while CPA eQTMs were connected with promoter methylations. As GTA eQTMs capture mostly the methylation differences between cell types, in contrast to donor differences for CPA eQTMs, this highlights a connection between enhancer methylation and cell type identity. In general, the eQTM models give valuable insights in the biological background behind eQTMs and can easily be extended for new features that could be evaluated as part of the genomic context, for example CpGs located at the position of non-coding RNAs or splicing sites. As an additional second use case, the eQTM prediction model can be applied to prioritize candidate genes associated with EWAS CpGs and facilitate so the downstream interpretation of EWAS results.

3.6. Materials and additional methods

3.6.1. KORA and LOLIPOP study

The major analyses in this project were performed on a combination of two different cohorts with whole blood samples, called KORA and LOLIPOP. KORA (short for: Cooperative Health Research in the Region of Augsburg) contains data from individuals with German nationality, who live in the region of Augsburg [138]. We selected a subset of the study with data from follow-up examinations between 2004 and 2008, called KORA F3 and F4 [139], where methylation and genotype data were available. This

subset contained 485 individuals from KORA F3 and 1,731 individuals from KORA F4 (no overlap between both). The KORA F4 cohort was used for meQTL discovery, the F3 cohort for meQTL replication (more in the next section 3.6.2). DNA methylation was measured with Illumina Infinium HumanMethylation450K BeadChips, genotypes with the Affymetrix Axiom platform and gene expression using the Illumina HumanHT-12 v3 BeadChip array.

The second cohort, LOLIPOP (short for: London Life Sciences Prospective Population Study) contained individuals of Indian Asian and European descend recruited in West London between 2003 and 2008 [119]. DNA methylation was measured for 1,851 donors in the discovery phase and 1,354 donors in the follow-up phase with the Illumina HumanMethylation450K array. This split was used for meQTL discovery and replication (more in the next section 3.6.2). Their genotypes were obtained from a combination of Illumina genotyping arrays (HumanHap3000, Human-Hap610, OmniExpress and OmniExomeExpress). Gene expression values were measured for a subset of 693 individuals of South Asian descent and 159 individuals of European descent with the Illumina HT-12 v4 BeadChip.

Detailed information about the processing of each omics dataset can be found in [2], including standard quality control, filtering, normalization and genotype imputation.

3.6.2. Identification of cosmopolitan meQTLs

Before the analysis, we removed CpGs, where the probe-sequences contained SNPs (MAF>1%) or were cross-hybridizing. The data of each cohort was normalized separately. MeQTLs were estimated separately for the discovery dataset of the KORA and LOLIPOP cohorts (section 3.6.1) using the linear regression models of CpG_residuals $\sim$ SNP_genotype. For this, the CpG_residuals of each CpG are obtained from linear regressions of the percentage methylation against technical covariates (the first principal components of the control-probe intensities) and biological covariates (age, gender and cell type proportion estimates) (exact details in Supplement of [2]). The significance threshold for meQTLs was set to P<0.05 after Bonferroni correction (raw P < 10^{-14}). Following this, we performed ancestry specific replication testing, separately for KORA and LOLIPOP, with the respective replication data from KORA and LOLIPOP (section 3.6.1). A combined inverse-variance meta analysis of discovery and replication data was run (R package *meta*) (more about meta analyses in Method section 2.2.3). MeQTLs were defined as replicated if the direction of effect between the discovery and replication analysis was consistent, the replication raw P-value < 0.05 and the meta analysis P-value < $10^{-}14$. As a last step in the rigid testing, the meQTLs replicated in the ancestry specific analysis were tested across ancestries using the same strategy as for the ancestry specific analysis. Only the set of meQTLs that could be replicated across ancestries was kept as the final set of cosmopolitan meQTLs.

3.6.3. Replication of meQTLs in isolated blood cell types, adipocytes and adipose tissue

Three additional datasets were generated to compare meQTLs across cell types and tissues. In the first dataset, isolated blood cell types were measured, namely CD4+ T cells, CD8+ T cells, monocytes and neutrophils. For this, blood of 60 individuals was measured, 30 of them obese with BMI > 35 and 30 of them healthy (BMI < 25). The second dataset contained subcutaneous and visceral adipose tissue samples from 48 individuals undergoing surgery, 24 of them morbidly obese with BMI > 40 and 24 healthy (BMI < 30). For both datasets, more information on the data generation and preprocessing is documented in [2]. For the meQTL testing, the linear regression model were applied with covariates for age, sex, ancestry and obesity case-control status.

The third dataset contained 603 adipose tissues samples from the MuTHER study (more information about the dataset again in [2]). Due to the relatedness of individuals, linear mixed models were used for meQTL identification using the GEMMA software [140], which estimated a kinship matrix between samples. As additional covariates, age, sex, the first 20 methylation principal components and the first 5 genetic principal components were added.

3.6.4. Identification of interaction meQTLs (iQTLs)

We tested interactions between the SNP and different phenotypes using linear regression models with an interaction term (see Methods section 2.2.4). To reduce the multiple testing burden, we chose first only CpG-SNP pairs from the set of cosmopolitan meQTLs. For the phenotypes, we selected cell type proportions of CD4+ T cells, CD8+ T cells and monocytes, estimated using the Houseman algorithm [48], and two environmental factors, cigarette smoking (binary yes/no) and BMI. The linear regression model with interaction term was built then as $CpG \sim SNP : phenotype + SNP + phenotype + covars$ with *covars* representing standard covariates used in each model: age, sex, smoking status, BMI, cell type proportions and 20 control probe principal components. iQTLs were defined significant, when the p-value of the interaction term was smaller than 0.05 after Bonferroni correction (raw p-value < $4.5 * 10^{-9}$). iQTLs were identified separately in KORA F4 cohort and LOLIPOP and tested for replication in the other cohort (P<0.05 and same direction of effect). Only replicated iQTLs were used for downstream analyses.

In a second step, we extended the iQTL analysis genome-wide for all CpG-SNP pairs in cis using *tensorQTL* (v1.0.3) [141] with age, sex, BMI and cell type proportions as covariates. Otherwise, the setup was exactly the same, the p-values thresholds after Bonferroni correction decreased to $2.0 * 10^{-11}$ for LOLIPOP and $8.8 * 10^{-11}$ for KORA.

Independence between global iQTLs and cosmopolitan meQTLs was tested by evaluating the linkage disequilibrium of the global iQTLs SNPs in the KORA cohort. All iQTL SNPs with $R^2 < 0.2$ for all meQTL SNPs were defined as independent. As CpGs are not considered here, this is potentially underestimating the real number of independent iQTLs.

3.6.5. Enrichment analysis of iQTLs among GWAS traits

iQTL SNPs were tested for GWAS enrichment using the tool *QTLEnrich*, which provides also summary statistics from 114 GWAS [32]. Enrichment analysis was performed for the global cis iQTLs and separately for each interaction phenotype, focusing on the three cell type proportions, which showed more and better replicable results. Sampling based empirical p-values were generated using an adaptive resampling scheme, which defined null variants as variants which are no iQTLs, i.e. never associated with any interaction phenotype. The significance threshold was set to FDR < 0.05 after Benjamini-Hochberg correction.

3.6.6. Identification of eQTMs

The eQTM analysis of associations could only be performed on a subset of the dataset (KORA: N=853, LOLIPOP: N=693), where also expression data was available. Gene expression was corrected for biological and technical covariates of sex, age, RNA integrity number, RNA-amplification plate (KORA), RNA-conversion batch (LOLIPOP), sample storage time (KORA) and RNA-extraction batch (LOLIPOP) via linear regression of gene expression levels against all the covariates (more in chapter 2.2). Additionally, both gene expression and DNA methylation are adjusted for genetic influences, again by linear regression (expression $\sim$ eQTL SNPs and methylation $\sim$ meQTL SNPs). meQTL SNPs were taken from this project itself (the set of sentinel meQTL SNPs), eQTL SNPs from a previous publication [142]. The resulting expression and methylation residuals were tested for associations using *Matrix eQTL* [95] (expression residuals $\sim$ methylation residuals), separately in KORA and LOLIPOP, followed by inverse-variance meta analysis to combine both. We repeated the eQTM association analysis with additional adjustment of both expression and DNA methylation for cell type proportions, estimated using the Houseman algorithm [48]. The two different eQTM sets were called GTA eQTMs (genotype adjusted eQTMs) and CPA eQTMs (genotype and cell type proportion adjusted eQTMs).

3.6.7. Annotating eQTMs with genomic features for the prediction

The features of the ML models described the CpG, the gene and the CpG-gene pair in a way that information about the eQTM probability of the pair can be derived. Annotations were taken from different public resources (full list in Supplementary Table A.1). In total, 39 annotations were collected to describe the CpG, the gene and the CpG-gene pair together, all based on genome version hg19 and scaled between 0 and 1.

All annotations used to describe the CpG are cell type specific. We selected the annotations of the blood cell types separately and combined the binary annotations (1=overlap and 0=no overlap) to a weighted average, weighted by the average cell type proportion in each cohort. These cell type proportions were estimated based on the Houseman algorithm [48]. For example, an enhancer score of 0.2 means that 20% of

the cells have an enhancer annotated at the position of this CpG, on average across the population.

3.6.8. Implementation and evaluation of different ML classifiers

To characterize the eQTMs, Logistic Regression and Random Forest models were trained to predict four different two-class classification tasks: distinguishing significant eQTMs from non-significant pairs, positive eQTMs from non-significant pairs, negative eQTMs from non-significant pairs and positive eQTMs from negative eQTMs (more details about the ML models in Method sections 2.3.1 and 2.7). To generate a high confidence annotation set for the prediction, a strict FDR cutoff of 0.01 was used and only CpGs with a high variance (top 6.25% most variable genes for the GTA model and top 25% most variable genes for the CPA model). All thresholds were chosen after a comparison of thresholds with a 10-fold cross-validation in the KORA dataset for the eQTM vs non-significant pair comparison. Thresholds were selected so that the performance of the model was increased, but the set of significant eQTMs not reduced too drastically. Except the different variance filtering threshold, there were no differences in building the model for the GTA eQTMs and the CPA eQTMs.

For the Logistic Regression, a probability cutoff of 0.5 was chosen for the classification (not necessary for AUC calculations). For the Random Forest model, the R package *randomForest* was used [118], with a subsample size of 62.5% times the complete data size, one third of all variables as the number of randomly sampled variables and 500 trees for each model.

Due to the large imbalance of the data sets, with a huge majority of non-significant pairs, the non-eQTM set was subsampled before training. This was performed 5 times combined each time with the 10-fold cross-validation of the dataset to show the stability of the performance for different negative sets. AUC, accuracy, sensitivity and specificity were used to evaluate the performance in the cross-validation (more in Methods section 2.2.2).

Replication between KORA and LOLIPOP was performed by training the models on one data set and predicting eQTMs on the same cohort (training data set) and on the other cohort (test data set). The dataset to train the model was subsampled to contain balanced classes, but not the dataset of the prediction.

3.6.9. Evaluation of feature importance

Random Forest and Logistic Regression were specifically chosen as two methods, where a good interpretation of the features is possible. For the LR models, the effect sizes are directly given by the model and represent the log odds ratios. Both the significance of a feature and the direction of an effect can be obtained from it. For the RF model, we used the Mean Decrease in Accuracy (MDA), as implemented in the *randomForest* package.

3.6.10. Feature selection

Due to the correlation structure between our features, we tested how many of the features are really necessary to build a similar well performing model compared to the full model with all features. For this, we applied three different selection algorithms: Lasso and step-wise regression for the LR models and Recursive Feature Elimination (RFE) for the RF models.

Lasso adds a penalization term using the L1 norm to the log-likelihood function of the Logistic Regression, so that the number of β values larger than 0 gets reduced during the parameter estimation [143]. We used the implementation of the R package *glmnet* [144], which identifies the optimal weight λ of the penalization term with a cross-validation. We chose the model with λ_{1se} whose performance is only one standard deviation worse than the optimal performance, but that enforces more regularization than the optimal model. For the feature selection, we counted all features with $\beta > 0$ as selected.

We tested the stepwise regression as an alternative feature selection strategy, specifically the backwards stepwise regression. First, a model with all features is trained and evaluated using the Akaike Information Criteria (AIC). The AIC rates the quality of a model dependent on its maximum likelihood combined with the number of used variables, so that a model with fewer features is rated better. Next, reduced candidate models are trained, each with one of the features removed, and their AIC is compared with the AIC of the full model. The model with the highest AIC is kept. If it was one of the candidate models, the removal of one feature is continued recursively. We used the implementation based on the R package *MASS* [145].

The third feature selection method, RFE, follows a very similar strategy than the backwards regression, but specifically for Random Forest. Least important variables are recursively removed from the dataset and the model quality is rated using the Out-of-bag prediction error. We use the R package *varSelRF* for it [146].

We combined the results of all three feature selection algorithms and chose features as top features which were selected at least 90% of times across the cross-validation runs and algorithms. With this set, we repeated the within cohort cross-validation for both KORA and LOLIPOP to see which performance can be reached with only the top features.

3.6.11. Prediction of muscle eQTMs using the blood eQTM model

To evaluate the generalizability of our model across tissues, we applied our eQTM model, trained on the whole blood CPA eQTMs from LOLIPOP to predict public available eQTMs measured in skeletal muscle [136]. We obtained the same features for muscle, as we did for the whole blood model, however, not all features were available. We collected 33 features in muscle from all 39 whole blood features, loosing mostly information about chromatin structure (TADs and HiC contacts). We decided to not add the TF annotations, as a far smaller set of TFs was measured for muscle compared to blood. For this reason, we retrained the LOLIPOP model first with this reduced feature set.

Afterwards, we ran the same performance evaluation as before.

3.6.12. EWAS interpretation with our eQTM model

As an additional application, the eQTM prediction model can support the interpretation of EWAS, because it identifies the genes associated with EWAS CpGs. The associated genes can then further be mapped to biological pathways using GO analysis or similar approaches. We showcased this application on an example EWAS study from Lindholm et al. [137], which studied methylation changes in muscle after regular exercise training for 3 months. The dataset contained not only the EWAS CpGs, but also the corresponding genes differential expressed after the exercise. First, we applied the eQTM model trained on LOLIPOP whole blood data together with the muscle annotations to predict the associated genes for the EWAS CpGs (same approach as in the last section 3.6.11). For the prediction, we used the LR and RF models separately and explored, as a third set, the intersection of predicted genes between both models. We validated all three sets with enrichment analyses of the predicted genes among the DE genes, using Fisher's exact test and all genes in cis distance to an EWAS CpG as background. To further explore the predicted EWAS genes, we ran GO enrichment analysis using the R package clusterProfiler [53] with the same background set and an FDR cutoff of 0.05.

4. Experimental design of multi-sample single cell transcriptomics

4.1. Importance of power analysis for experimental design

The recent development of single cell transcriptomics has opened new avenues for population genetics, as it facilitates the analysis of cell type specific eQTLs and enables new types of QTL analyses (more in chapter 1.1.6). One of the limitations so far is the lack of large cohorts with single cell transcriptomics data. For this reason, the generation of such datasets is currently of high priority in the field, for example shown in the efforts of the sc-eQTLgen consortium [88]. Single cell experiments are still very expensive compared to bulk sequencing, which makes a proper experimental design of those studies even more important. The generation of well-powered cohort datasets is crucial for the success of all downstream analyses.

eQTL studies can be seen as a special case of DE analyses, where the group distribution is defined by the genotypes of the individuals with regard to a certain genetic variant. This means that the population is grouped differently for each tested genetic variant. Because of this relationship between eQTL and DE studies, the corresponding power analysis tools for the study design are very similar and can even cover both types of studies in one (given a few small adaptations). However, for single cell multi-sample transcriptomics studies in general, efficient power analysis tools are missing. Either the tools were developed for bulk [147, 148, 149, 150] and lack the necessary adaptions for single cell characteristics, such as the increased sparsity dependent on the number of cells, or they are based on simulations [112, 113, 151], which makes them very computational demanding, especially for the design of larger cohorts.

To fill this gap, we developed *scPower*, the first analytic power analysis framework for single cell multi-sample transcriptomics data (Figure 4.1). It estimates the power for a single cell DE or eQTL experiment based on the three main parameters - the number of samples, the number of cells per samples and the read depth - in combination with a set of priors, including the expected effect sizes and cell type specific expression priors. The tool overcomes previous limitations: first, it takes single cell specific characteristics into account by modelling the single cell specific expression probability of the genes, dependent on the experimental parameters and prior information. Second, it is an analytic tool, making it orders of magnitude faster and more memory-efficient compared to simulation-based approaches. This way, it allows easily the design of large cohorts, which are not feasible to simulate on a standard laptop. As a third important aspect, we provided both an R package `https://github.com/heiniglab/scPower` and a website

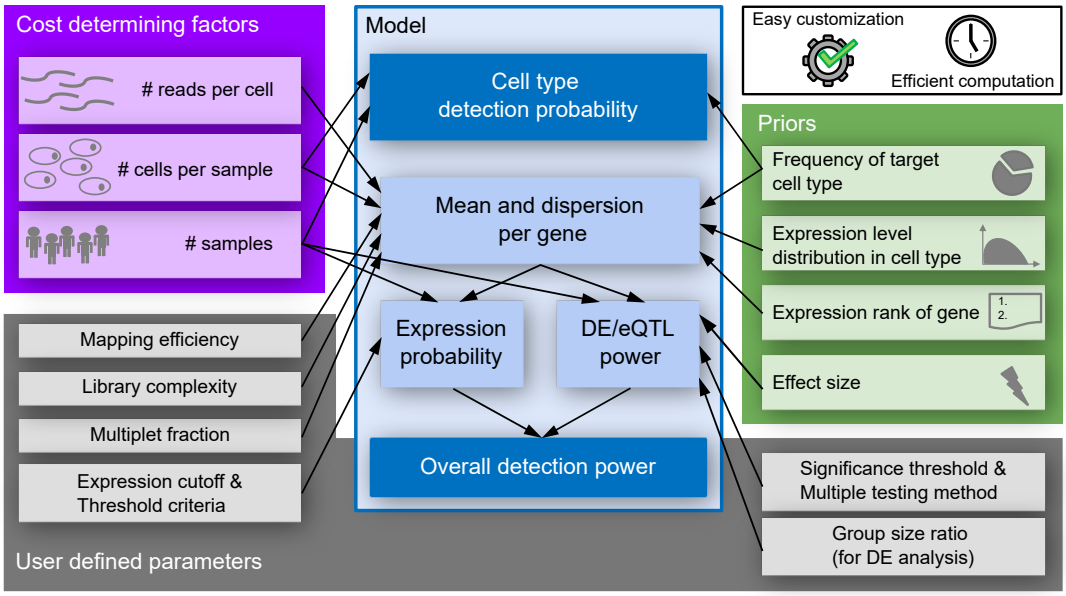


Figure 4.1.: **Schematic overview of the power and design framework *scPower***
The main experimental parameters (purple) - the number of samples, the number of cells per sample and the number of reads per cell - determine the cost of the experiment and the overall detection power (blue). Additionally, the power depends on prior knowledge or assumptions (green) and further user-defined parameters such as the significance threshold and the expression cutoff (gray). In contrast to previous power analyses tools, it enables efficient computation of power and easy adaptation to new use-cases through. The figure includes the most important parameters and the main parts of the model. Figure and legend taken from [1].

`http://scpower.helmholtz-muenchen.de` to improve the usability of our tool, also for interested users with less programming experiences.

The methodology, results and figures of this project have previously been published in Schmid et al. [1]. All code, including a description on how the figures were generated, is available in the GitHub repository associated with the publication `https://github.com/heiniglab/scPower`.

4.2. Description of the analytic power analysis framework

4.2.1. Statistical model behind DE and eQTL analyses for *scPower*

Our power analysis framework is specifically designed for cell type specific single cell eQTL and DE studies that use the pseudobulk approach (see Methods section 2.6). For multi-sample DE studies, two benchmarking studies showed that the pseudobulk

approach in combination with established DE methods, originally developed for bulk (e.g. edgeR [103] or limma-voom [152]), performed the best [113, 114]. For single cell eQTL studies, applying pseudobulk is a widely-used strategy [55, 83, 57] and also suggested in a best-practice single cell eQTL publication [89]. Importantly, the pseudobulk approach differs from classical bulk RNA-seq, as the number of detected genes depends on the number of aggregated cells. For this reason, we added the number of cells per sample as a relevant additional parameter in our power analysis model compared to bulk power analyses (more in subsection 4.2.3). Of note, every power analysis tool is crucially dependent on the statistical testing procedure. Therefore, *scPower* can only be applied for single cell eQTL and DE studies that use pseudobulk aggregation.

For the DE power, the framework's default implementation assumes the most common setup with a two-group comparison. Additionally, the power for more complex designs can be estimated with *scPower*, as long as they are described by general linear models. An example, showing the comparison of three groups, is presented in the package vignette "Extending scPower to complex designs".

The starting point for the multi-sample experiments is a count matrix of genes times cells. The counts can be either based on UMIs or reads, dependent on the scRNA-seq technology. An evaluation of technological details and differences is given in section 4.5, but the general model described in the first part is always the same. Our power evaluation does not cover the preprocessing, it starts with a fully annotated count matrix, where each cell is assigned to one cell type and one donor (see subsection 4.7.2 for how we processed our example dataset). Other studies have already covered how experimental parameters affect accurate cell type annotations, which is a general question for single cell design and not specific for multi-sample experiments [153]. On top of that, different experimental covariates, for example age, sex and experimental batch, describe each donor and are usually added into the models (see Method sections 2.2).

4.2.2. Modelling the power to detect a certain number of cells for the cell type

The main focus of *scPower* is the estimation of the so-called overall detection power of an experiment. It describes the probability of detecting a set of DE or eQTL genes in a specified cell type given the annotated count matrix and the experimental parameters. Two of the parameters, the number of samples and the read depth, can be defined directly in the experimental setup, the third parameter however, the number of cells per sample and cell type, depends on the power to observe the cell type of interest. For this reason, *scPower* covers the cell type detection probability additional to the overall detection power (Figure 4.1).

The cell type detection probability describes the probability to detect at least $n_{c,s}$ cells per individual from the cell type of interest c in each individual for a cohort of size n_s. It depends on the total number of cells per individual n_c - independent of the cell types - and the frequency of the cell type of interested f_c. The model itself is an

adaption from the website "How Many Cells" of the Satija lab [154], which did not cover the multi-sample setting. Following this approach, we model the cell type detection probability P_{CT} based on a cumulative negative binomial distribution F_{NB} parameterized by n_c, $n_{c,s}$ and f_c and take the result to the power of n_s to model the probability that we detected $n_{c,s}$ cells in each individual of the cohort:

$$P_{CT} = F_{NB}(n_c - n_{c,s}, n_{c,s}, f_c)^{n_s} \tag{4.1}$$

Applying this model, we calculated the minimal number of cells n_c that is required to reach a cell type detection probability of 95% for different parameter settings (Figure 4.2). We explored the probabilities for different cell types that belong to the peripheral mononuclear blood cells (PBMCs), matching our model dataset in the following sections. The required cell type frequencies were taken from literature [155], estimating them to be twice as high in PBMCs as in whole blood.

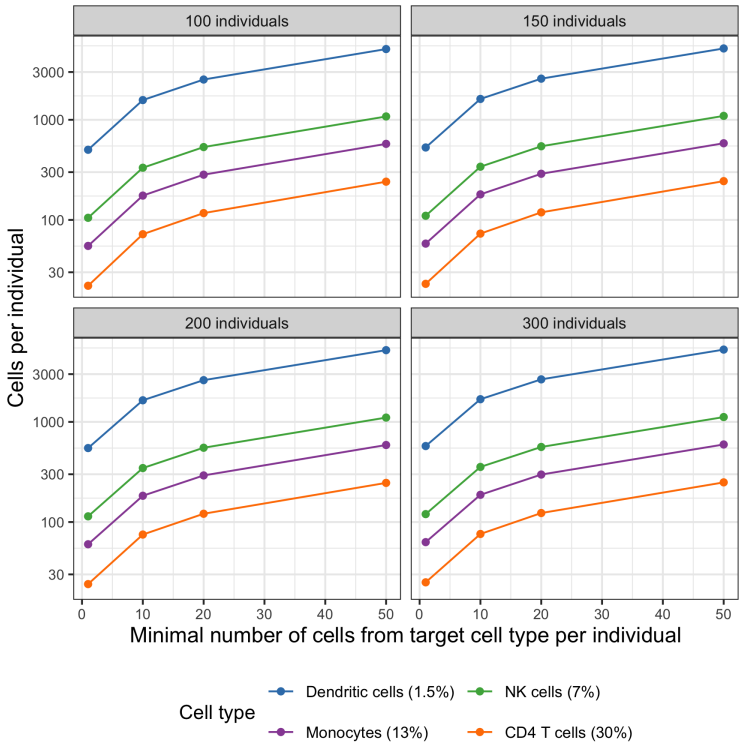


Figure 4.2.: **Cell type detection probability**
Required number of cells per individual (y-axis, log scale) to detect the minimal number of cells from a target cell type per individual (x-axis) with a probability of 95%. The probability depends additionally on the total number of individuals (subfigure header) and the frequency of the target cell type (line color). Figure and legend taken from [1].

The cell type frequency had a large influence on the required number of cells, high-

lighting the difficulty to capture enough cells from rare cell types. For example, 72 cells per individual are required to get at least 10 CD4+ T cells with a frequency of 30%, but 1,574 cells per individual are needed to get at least 10 Dendritic cells with a frequency of 1.5%. As expected, the more cells of the cell type one wants to capture, the more cells need to be measured in total (e.g. 72 cells to capture 10 CD4+ T cells, but 241 cells to capture 50 CD4+ T cells). In contrast, the cohort size had only a minor influence on the cell type detection probability.

4.2.3. Modelling the general detection power

Given that a certain number of cells per cell type and individual is detected, the overall detection power of the experiment can be estimated. It depends on the three main experimental parameters, the number of samples, number of cells per sample and the reads per cell, in combination with a set of priors (Figure 4.1). The overall detection power P_i of a gene i can be further deconstructed into two parts: first, a gene needs to be detected in the dataset, i.e. a certain number of counts for this gene needs to be measured in a certain fraction of individuals of the population. We call this part the expression probability $P(i \in E)$ in the following, so the probability that the gene i is in the set of expressed genes E. As single cell RNA-seq data is currently considerably sparser than bulk experiments, this provides a bottleneck for multi-sample analyses, because part of the lowly expressed genes will not be measured [156].

The second part of the overall detection power is the DE or eQTL power itself, so the power of the statistical test for gene i. In the following, we denote it as $P(i \in S)$, the probability that the gene i is in the set of significantly differentially expressed genes S.

Both the expression probability and the DE/eQTL power are affected by the expression distribution of the gene i, which is modelled as a negative binomial distribution with mean μ and dispersion ϕ (see Methods section 2.1.3). We will show in section 4.2.5, how the mean μ and dispersion ϕ of the distribution can be inferred for a specific experimental setting based on the number of cells n_c, the read depth r and an expression prior Θ_e that captures the cell type specificity. For the expression probability, additionally the sample size n_s plays a role, as the gene needs to be detected in a certain number of individuals. For the DE/eQTL power, additionally the expected effect size between the groups Θ_p, the chosen significance cutoff α (multiple-testing corrected) and again the sample size n_s are additional dependent factors.

We can decompose the overall detection power P_i into the product of the expression probability and the DE/eQTL power because P_i is conditioned on the aforementioned priors Θ_e and Θ_p as well as the three main experimental parameters (n_s, n_c, r):

$$\begin{aligned} P_i &= P(i \in E \wedge i \in S | n_s, n_c, r, \Theta_e, \Theta_p, \alpha) \\ &= P(i \in E | n_s, \mu(n_c, r, \Theta_e), \phi(n_c, r, \Theta_e)) \cdot \\ &\quad P(i \in S | n_s, \mu(n_c, r, \Theta_e), \phi(n_c, r, \Theta_e), \Theta_p, \alpha) \end{aligned} \tag{4.2}$$

The overall detection power P of the complete experiment is then defined as the average over all DE/eQTL genes in the set D:

$$P = \frac{1}{|D|} \sum_{i \in D} P_i \tag{4.3}$$

4.2.4. Estimating the expression prior

As explained in the last paragraph, the expression probability is an important component for the overall detection power. For this, we developed a new model that estimates the expression probability of a gene in a pseudobulk matrix (Figure 4.3 a). We define a gene as expressed if it has at least n counts in a fraction of k individuals. Alternatives to this expression definition are also covered by our model (discussed in section 4.2.6).

To estimate the expression probability for each gene in a planned experiment, the count distribution of the gene needs to be modelled. *scPower* does this based on the three main experimental parameters and a cell type specific expression prior that is derived from a matched pilot dataset with as similar conditions as possible, such as the same tissue and scRNA-seq technology. Furthermore, the matched pilot dataset should contain at best only samples from one group, usually the healthy control samples. Differences between the DE groups will then be modelled on top of the expression distribution in the second step, when the DE/eQTL power is estimated.

Figure 4.3 a shows how the expression prior can be inferred from a pilot dataset that consists of an annotated UMI count matrix with donor and cell type information. The goal of this process is to describe the general count distribution for each cell type with a small set of hyperparameters. The following steps describe this approach for UMI count matrices. For full-length read-based count matrices, a few small adaptations are necessary, discussed in section 4.5. Exemplarily, for visualization and evaluation purposes, expression priors were fitted for a pilot data set from the BECOME study [157], containing PBMC data of 14 healthy controls (Supplementary Figure A.9, preprocessing described in Method section 4.7.2).

First, the expression distribution for each gene i in each cell type c is estimated by fitting a negative binomial distribution over all corresponding single cell counts (see Methods section 2.1.3), leading to two parameters, mean $\mu_{i,c}$ and dispersion $\phi_{i,c}$, per gene and cell type. Modelling the distribution separately for each cell type ensures to capture the distribution differences between the cell types [158]. To further abstract the parameters across the genes in one cell type, the distribution of mean values $\mu_{i,c}$ is modelled as a mixture distribution based on a zero component $Z(x)$ and two left censored gamma distributions $\Gamma(x, r, s)$ parameterized by their mean and standard deviations (see Methods section 2.1.4):

$$f_{\mu_c} = p_1 Z(x) + p_2 \Gamma(x, \mu_{G1}, \sigma_{G1}) + p_3 \Gamma(x, \mu_{G2}, \sigma_{G2}) \tag{4.4}$$

This mixture distribution is an adaption of the distribution used in the single cell simulation tool *Splatter* [75], which applies a second gamma distribution to capture

highly expressed gene outliers. In contrast to *Splatter*, we model non-expressed or very lowly expressed genes more accurately by adding the zero component $Z(x)$ and the left censoring for both gamma distributions. The censoring point depends on the number of measured cells for the cell type n_c, as the smallest expression level that is measurable for the cell type is $\frac{1}{n_c}$. This means that the more cells from a cell type we capture, the more lowly expressed genes we can detect. This is also visible in the gamma fits of the PBMC pilot dataset (Supplementary Figure A.10).

The mixture distribution captures the gene expression means for a certain read depth. To apply the expression prior for different read depths, we subsampled the original count matrices and explored the effect on the mixture distribution. This showed clearly that the parameters of the mixture distribution - the mean and standard deviations of the two gamma distributions ($\mu_{G1}, \sigma_{G1}, \mu_{G2}, \sigma_{G2}$) and the probability of the zero component and first gamma distribution (p_1, p_2) - were linearly dependent on the mean UMI counts per cell (Supplementary Figure A.11), while the probability of the second gamma distribution (p_3) that captures the highly expressed outliers stays constant.

The relationship between the mean UMI counts per cell and the mean number of reads per cell showed a logarithmic fit when we subsampled different datasets (Supplementary Figure A.12). This represents the saturation curve of the single cell experiment: the more reads are sequenced, the fewer new UMIs are detected. The exact parameters of this logarithmic fit are however specific to the experimental setting and difficult to generalize. *scPower* provides saturation fits for different datasets, from which the users can choose.

In the last step of the expression prior generation, the dispersion parameters $\phi_{i,c}$ are also further abstracted, by modelling them dependent on the mean parameters $\mu_{i,c}$. For this, we applied the method from DESeq [104]. All these hyperparameters together (the mixture parameters modelled dependent on the UMI counts, the UMI-read depth fit and the dispersion-mean fit) comprise the expression prior.

4.2.5. Modelling the expression probability

Based on these approximated expression priors, the expression probabilities for all genes of a newly planned experiment can be estimated dependent on the chosen experimental parameters. As a short reminder, genes are defined as expressed if they have at least n counts in a fraction k individuals in the pseudobulk matrix.

In the pseudobulk approach, the single cell count matrix with counts $x_{i,j}$ for gene i and cell j is aggregated to a three-dimensional matrix of genes times individuals times cell types with the count $y_{i,s,c}$ for gene i, individual s and cell type c calculated as:

$$y_{i,s,c} = \sum_{j \in C \wedge j \in S} x_{i,j} \tag{4.5}$$

with C the set of all cells assigned to cell type c and S the set of all cells assigned to individual s.

As in the previous section 4.2.4, we assume for each cell type c a negative binomial distribution for all counts $x_{i,j}$. We can infer the corresponding parameters $\mu_{i,c}$ and $\phi_{i,c}$ from the expression prior of the cell type and the parameters of the new experiment. Based on this, the pseudobulk counts $y_{i,s,c}$ are also negative binomial distributed with shifted parameters dependent on the number of cells for this cell type and individual $n_{c,s} = |j \in C \wedge j \in S|$:

$$\mu'_{i,c,s} = n_{c,s} * \mu_{i,c} \tag{4.6}$$

$$\phi'_{i,c,s} = \frac{\phi_{i,c}}{n_{c,s}} \tag{4.7}$$

After estimating the parameters of the pseudobulk distribution for the cell type, the probability that the pseudobulk count $y_{i,c,s}$ for one individual s is greater than n can be inferred from the cumulative negative binomial distribution as:

$$p_{i,s} = P(y_{i,c,s} > n) = 1 - F_{NB}(n, \mu'_{i,c,s}, \phi'_{i,c,s}) \tag{4.8}$$

So, $p_{i,s}$ represents the expression probability of the gene i for one individual. The general expression probability for the whole cohort requires that this is true for at least a fraction of k individuals, so $k * n_s$ individuals in total. This follows a cumulative binomial distribution (Methods section 2.1.2):

$$P(i \in E) = 1 - F_{Bin}(k * n_s, n_s, p_{i,s}) \tag{4.9}$$

Afterwards, the expected number of expressed genes $\mathbb{E}(E)$ can be directly derived from the expression probability of all genes

$$\mathbb{E}(E) = \sum_{i \in G} P(i \in E) \tag{4.10}$$

with G the set of all genes.

We applied this formula to estimate the number of expressed genes in our pilot PBMC dataset for different read depths. The expected number of expressed genes matched well with the measured numbers across the different cell types and read depths. We evaluated both an expression cutoff of more than 0 counts in 50% of the individuals and more than 10 counts in 50% of the individuals and got correlation values of $r^2 = 0.994$ and $r^2 = 0.997$ for the depicted experimental batch in Figure 4.3 b. Also for the other batches, the concordance was very high (r^2 between 0.983 and 0.997).

Additionally, we predicted the number of expressed genes in an independent validation dataset, which had not been used for the expression prior generation [159]. Again, the number of expected and measured expressed genes matched quite well (for count > 0 $r^2 = 0.934$ and for count > 10 $r^2 = 0.971$) (Figure 4.3 c), despite being measured with a different read depth and for a different sample size as the pilot dataset for the prior. This supports the generalizability of our model across different datasets. In the *scPower*

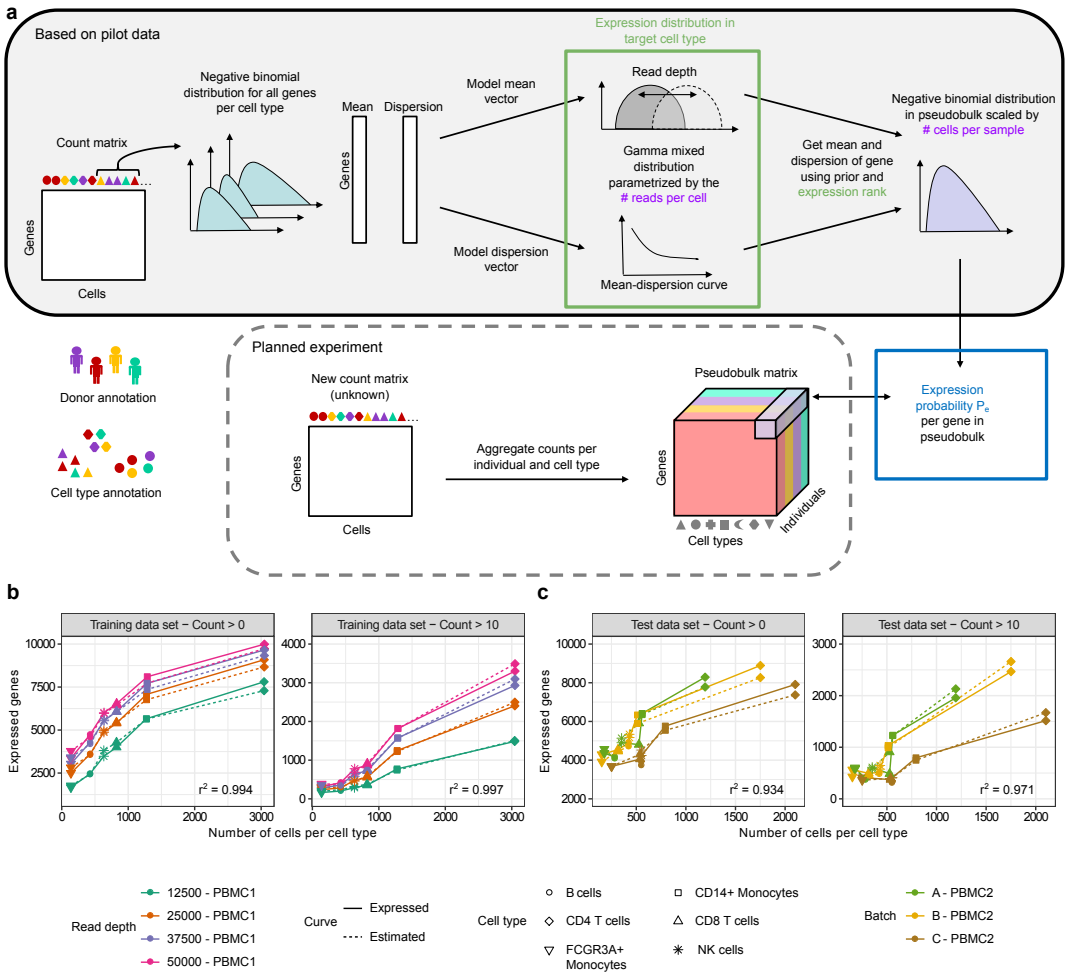


Figure 4.3.: **Expression probability model parameterized by UMI counts per cell**
a The expression probabilities for genes in pseudobulk of a newly planned experiment are estimated based on the expression prior and the planned experimental parameters. For this, the expression prior is derived from the mean and dispersion of gene-wise negative binomial distributions fitted from a matching pilot data set. **b** Using this approach, the number of expressed genes expected under our model (dashed line) closely matches the observed number of expressed genes (solid line) dependent on the number of cells (cell type indicated by point symbol) for one batch of the training PBMC data set. The data is subsampled to different read depths. The r^2 values between estimated and expressed genes were highly significant for both expression thresholds. **c** The model performed similarly well for the three batches of an independent validation PBMC data set. Used expression threshold: count > 10 (right panels of **b**, **c**) or count > 0 (left panels of **b**, **c**) in more than 50% of the individuals. Figure and legend taken from [1].

R package, we provided priors for 25 cell types in 3 different tissues and easy to use functions to extend the expression prior for all other cell types of interest.

4.2.6. Alternative parameterization of the expression probability

Of note, the exact thresholds to define a gene as expressed are debatable and users might prefer other definitions dependent on the use-case. Instead of an absolute threshold for the number of counts in the cell type, a percentage value could be used, for example defining genes as expressed if they are present in at least 50% of the cells with a value larger than 0. Vice versa, for the number of individuals that need to fulfill the count threshold, the percentage value could be replaced with an absolute value. Our tool covers additionally these cutoff definitions, which makes it overall very flexible.

For the first variation - updating the absolute expression cutoff with a relative one - the estimation needs to be done slightly differently. The probability that a gene is expressed with a count larger than 0 in a cell of a cell type c can be estimated based on the fitted negative binomial distribution $\mu_{i,c}$ and $\phi_{i,c}$ (the initial fit, not the transposed parameters from the pseudobulk):

$$p_{cell} = 1 - F_{NB}(0, \mu_{i,c}, \phi_{i,c}) \tag{4.11}$$

The probability that this happens in a certain fraction of cells q represents the expression probability for one individual $p_{i,s}$. It can be modelled using a cumulative binomial distribution dependent on the number of cells for this cell type and sample $n_{c,s}$:

$$p_{i,s} = 1 - F_{Bin}(q * n_{c,s}, p_{i,s}) \tag{4.12}$$

This probability $p_{i,s}$ can then be used to get the expression probability $P(i \in E)$ as defined before.

The second variation - updating the definition of how many individuals need to express the gene - is very straightforward. Looking at the final cumulative binomial distribution to calculate the expression probability, only the first parameter needs to be changed:

$$P(i \in E) = 1 - F_{Bin}(x, n_s, p_{i,s}) \tag{4.13}$$

with $x = k * n_s$ if a percentage cutoff k should be used or setting x directly to an absolute number if an absolute cutoff should be used.

For all downstream analyses, we will use the initially proposed model that the gene needs to be expressed in a certain fraction of samples with a certain absolute number of counts, as we recommend this parameterization in general. In contrast, the cutoff of "expressed in a certain percentage of cells" tends to define more expressed genes for less frequent cell types compared to more frequent ones, as fewer cells in total increase the probability that a gene is expressed in a certain fraction of these cells (Supplementary Figure A.13). This is conflicting with the empirical and statistical intuition that more measurement points (i.e. cells) will lead to more accurate estimations and so more genes

that can be looked at. Shallow sequencing of a large number of cells has been shown before to be beneficial for scRNA-seq experiments [153, 160]. Using a percentage cutoff would reverse this effect.

For the second part, defining a gene as expressed if it is expressed in a certain fraction of the population is a common strategy, used already in bulk. Genes that are expressed in only one individual or very few individuals have very low statistical power to be detected as DE or eQTL genes, but increase the multiple testing burden. For this reason, we propose to follow the suggestion as implemented edgeR [103]. They keep only genes that are expressed in at least one of the comparison groups of the DE study. This leads to the 50% cutoff for balanced groups with the same sample size, which we will use in the following.

As an equivalent for eQTL studies, we suggest that the gene should be expressed at least in the heterozygotes. Depending on the chosen minor allele frequency f_{ma} for the study, this leads to an expression cutoff of $2 * f_{ma} * (1 - f_{ma})$. We assume a MAF of 0.05, a typical value for an eQTL study, and set therefore in the following the percentage threshold for eQTL studies to 0.095.

4.2.7. Modelling the DE/eQTL power

After having established the model to calculate the expression probability, including different definitions of it, we can estimate the DE/eQTL power and combine both to get the overall detection power (see section 4.2.3). For the DE/eQTL power, the large advance of *scPower* is that it is an analytic power analysis method, as opposed to the simulation-based methods previously published [112, 113]. This makes our tool by orders of magnitude faster and memory efficient (more in Methods section 2.5).

To calculate the analytic power, the expected count distributions for a planned experiment need to be known, which depend on the chosen experimental parameters. We covered this issue already for our expression probability model, which provides us the cell type specific mean and dispersion parameters of the pseudobulk negative binomial distributions. Based on this information, analytic models developed for bulk can be applied also for pseudobulk models.

Other necessary prior information for the power estimation are the expected number of DE/eQTL genes combined with their expected effect sizes and the chosen significance threshold. While the general setup is the same for the DE and eQTL power, there are a few small differences: for the DE power, the underlying model is a negative binomial regression and the effect size is estimated as fold change between the groups (see Methods section 2.3.2). This regression type matches the count distribution for gene expression and is commonly used in established tools such as DESeq [104] and edgeR [103], which perform well in single cell DE benchmarking studies [113, 114]. We use the corresponding method to get the analytic power for negative binomial regressions developed by Zhu et al. [161].

For eQTL analyses, typically linear models are used on log-transformed count matrices, as these can be computed far more efficiently and facilitate testing a large search space

with many SNP-gene pairs (see Methods section 2.2 for details). For this reason, the power calculation needs to be changed to linear models [111]. Importantly, we observed that the analytic solution can be inaccurate for genes with small expression means (Supplementary Figure A.14), as the normalization via log-transformation does not work effectively here in all cases. Therefore, we simulate the power for genes with small mean values < 5 instead of using the analytic solution (details in section 4.7.8). As this affects only a small number of genes and the corresponding values can be precalculated, this is still far faster than pure simulation-based methods. Effect sizes are estimated based on R^2 values, which combine the allele frequency of the SNP and the beta value of the linear regression, both affecting the eQTL power.

DE and eQTL studies have in common that multiple testing correction of the significance threshold is required to control the number of false positives due to the larger number of tests (more in Methods section 2.4). *scPower* covers different approaches for this, either correcting the family-wise error rate (FWER) or the false discovery rate (FDR), which influence the significance threshold and so the resulting power.

For eQTL studies, FWER correction has been established. We follow the estimation of the GTEx consortium [32], that approximates 10 independent SNPs are tested per gene in a genome-wide cis eQTL analysis. This results in $10 * \mathbb{E}(E)$ tests with $\mathbb{E}(E)$ the expected number of expressed genes estimated from our expression probabilities. This leads to a Bonferroni corrected threshold of

$$\alpha = \frac{\alpha'}{10 * \mathbb{E}(E)} \tag{4.14}$$

for α' being the uncorrected significance threshold.

For DE studies, FDR correction is established. Here, the correction of the significance threshold is not as straightforward, because the p-value distribution of all tests is required, which is not obtained in the power analysis. Nevertheless, the correction is possible following the method of Jung [162]:

As discussed in Method section 2.4, the FDR is the fraction of false positives among all positive predicted tests. We assume that the p-values under the null hypotheses are uniformly distributed according to the probability integral transform. These p-values are derived from the tested non-DE/eQTL genes and their number is calculated based on the expected number of expressed genes and expressed DE/eQTL genes as $m_0 = \mathbb{E}(E) - \mathbb{E}(E_{DEG/eQTL})$. This leads to an expected number of $m_0 * \alpha'$ false positives for a corrected significance threshold of α'. The expected number of true positives can be taken from the power itself, abbreviated in the following $r_1(\alpha')$. Together, the true positives and false positives make up the positive predicted, leading to an FDR of:

$$FDR(\alpha') = \frac{m_0 * \alpha'}{m_0 * \alpha' + r_1(\alpha')} \tag{4.15}$$

The unknown raw p-value of α' which leads to an FDR value of $\alpha = FDR(\alpha')$ is inferred via numerical optimization in *scPower*.

In the following examples, we will use a significance threshold of 0.05 that we will correct for the DE cases using FDR correction and for the eQTL cases using FWER correction assuming 10 independent SNPs per gene.

4.2.8. Exploring the general detection for a few chosen examples

We tested the power analysis model, explained in the previous sections, in several example scenarios to explore the effect of different experimental parameter combinations on the power. In these use cases, we applied our expression priors from the PBMC training dataset and gathered realistic cell type specific effect sizes from FACS sorted bulk DE and eQTL studies.

For the DE cases, we used a study comparing chronic lymphocytic leukemia (CLL) subtypes [163] and a study exploring systemic sclerosis in macrophages [164]. *scPower* estimated an overall detection power of 74% for the CLL study (comparing iCLL vs mCLL), when measuring 3000 cells per cell type, a sample size of 20 (with balanced groups) and 20,000 reads per cell (Figure 4.4 a). The DE power in this scenario would be even higher with 98%, but the expression probability is only 74%. This means that only 74% of the DE genes are likely to be detected in the single cell dataset, reducing the overall detection power.

When exploring a range of parameter combinations, both an increase of the sample size and an increase of the number of cells led to a higher overall detection power, but increasing the number of cells had the far bigger effect (Figure 4.4 a). The weak effect of the sample size is probably caused by the small sample size of the reference study used for the prior ($N = 6$). A general limitation of using references studies to estimate the effect sizes is the power of the reference study. Potential reference studies with a higher sample size would have found additional DE genes, which in turn could reduce our power estimations for newly planned experiments.

We estimated with *scPower* similar power values for the comparison of the other CLL subtypes in the same study. Contrary, the power for the systemic sclerosis study was far lower with a power of maximum 30%, probably driven by the smaller absolute fold changes in this study (Supplementary Figure A.15).

For the eQTL effect sizes, we obtained values from a FACS sorted study of the BLUEPRINT consortium, which identified eQTLs in T cells and Monocytes [43]. We observed a maximal power of 64% for T cells (Figure 4.4 a) and 65% for Monocytes (Supplementary Figure A.15), when measuring 3,000 cells, 200 samples and 20,000 reads per cell. Here, the increase in power with more cells is again visible and additionally the effect of a larger sample on increasing the power is more pronounced.

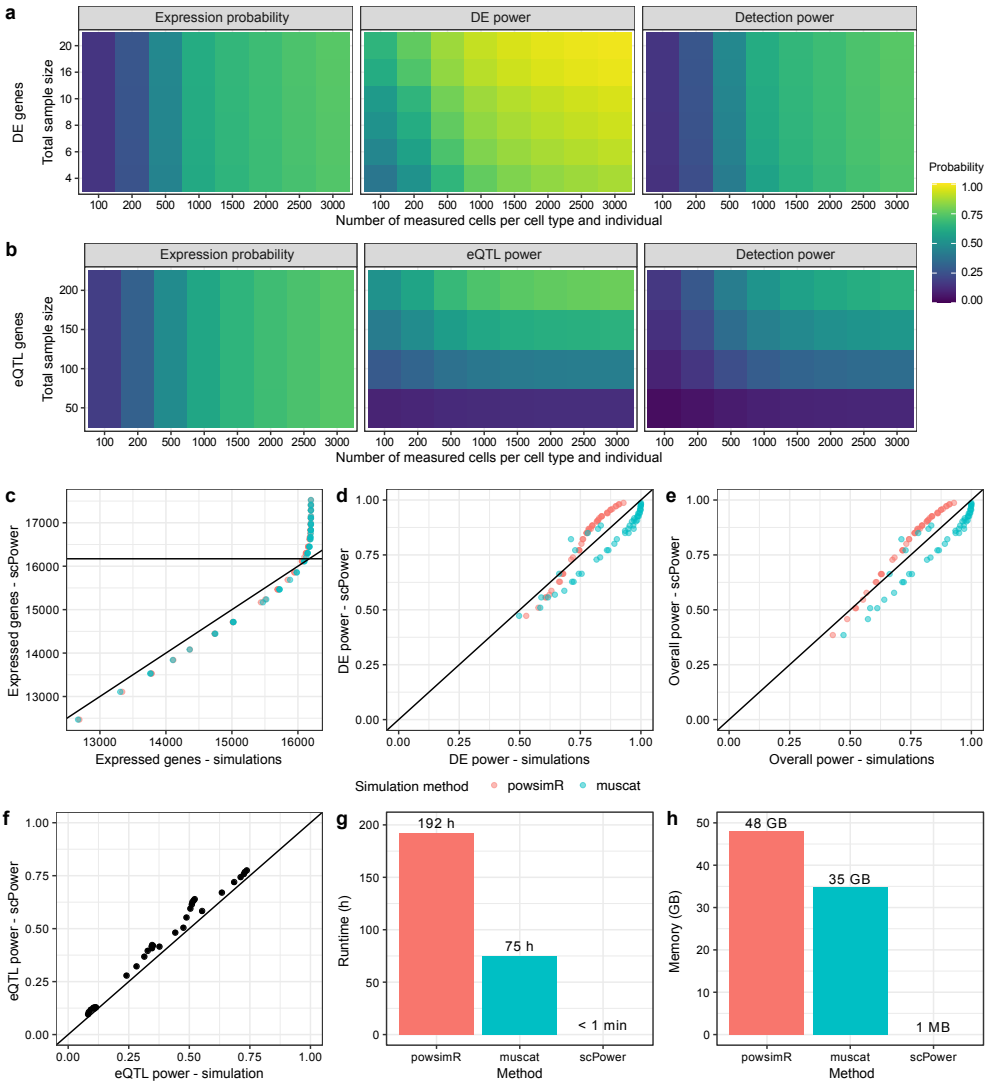


Figure 4.4.: **Estimating overall detection power and validation in simulation studies** **a,b** Power estimation using data driven priors for DE genes **(a)** and eQTL genes **(b)** dependent on the sample size and the number of cells. The effect sizes and expression ranks of the DE/eQTL genes were taken from published studies (**a**: the BLUEPRINT CLL study, **b**: BLUEPRINT T cell study). **c,d,e** The probabilities calculated in **a** were verified by the simulation-based methods powsimR and muscat with each point representing one parameter combination. **f** The eQTL power of **b** could be replicated with a self-implemented simulation. **g,h** Runtime **(g)** and memory requirements **(h)** were drastically higher in the simulations than for our tool *scPower* during the evaluations of **c–e**, showing the strength of our analytic model. Figure and legend taken from [1].

4.3. Comparison with simulation-based tools for validation and benchmarking

4.3.1. Validating the DE power calculation based on simulation

To validate our DE power analysis results, we compared the estimations of *scPower* with the published simulation-based tools *powsimR* [112] and *muscat* [113]. For both, small adaptions were necessary, for *powsimR* to incorporate the multi-sample setting and for *muscat* to input the same set of priors (see section 4.7.11). Following the optimal workflow identified in multiple benchmarking studies [113, 114], we processed the simulated pseudobulk matrices with established DE methods, namely *edgeR* [103], *DESeq2* [105] and *limma* [152].

The power estimations from our analytic solutions of *scPower* were very close to the simulated results of both *powsimR* and *muscat* with *edgeR* across different parameter combinations (Figure 4.4 c-e). We evaluated the different parts of our model separately and proved that all align well: the expression probability, here shown as number of expressed genes, the DE power and the overall detection power. Of note, the simulation-based methods can simulate at most the number of genes that were in their training dataset, while *scPower* can extrapolate how many additional genes could be identified with more cells/higher read depth. This causes the deviation in Figure 4.4 c, that *scPower* estimates in some cases more than 16,000 genes in this example.

Combining *powsimR* and *muscat* with different DE methods showed the influence of the analysis pipeline on the power (Supplementary Figure A.16). Despite these small differences in power depending on the DE method, all simulated power estimations were close to our estimations with *scPower* and the trends of how the power develops for different parameter combinations were accurately captured. These trends are essential to decide on the best parameter combination for an experiment. We included both parameter settings with FWER multiple testing correction and FDR correction in our evaluation, both worked well (Supplementary Figure A.16).

Notably, there were power differences between *powsimR* and *muscat* when using the same DE method, highlighting methodological differences between the different simulation-based methods. *scPower* lay in most cases between the estimations of both.

4.3.2. Validating the eQTL power calculation based on simulation

The first approach to simulate eQTL power, called *splatPop* [151], has been published only very recently, no method was available during the development of *scPower*. For this reason, we used our own simulation-based tool to validate the eQTL power (described in section 4.7.8). The expression probability model is the same as for the DE part, so this part was already validated before, and we focused on the eQTL power itself. When we compared the power of eQTL from *scPower* to our simulated values, we saw high concordance. Overall, the comparison with simulation-based power estimations supports our analytic model, both for DE and eQTL analyses.

4.3.3. Runtime and memory advantages of scPower

We quantified the computational advantage of our analytic model over the simulation-based approaches, by measuring the runtime and memory consumption for computing all parameter combinations shown in Figure 4.4 c-e. *scPower* was orders of magnitude faster (less than one minute compared to multiple days) (Figure 4.4 f) and had far lower memory requirements (Figure 4.4 e). This highlights *scPower* as a very user-friendly tool that can be quickly run on any personal computer. Furthermore, the fast calculation allows easy comparison of different experiment designs, which we will further explore in the next sections.

4.4. Optimization of experimental design with scPower

4.4.1. Extending the power framework to budget restricted optimization

In a realistic use case, the experimentalist is constrained by certain resources, often the budget. The question arising from this is how to choose the experimental parameters in a way to maximize the power while staying within the budget. The fast analytic calculations of *scPower* allow these optimizations of parameters under restricted budget.

For this, the costs need to be calculated dependent on the experimental parameters (sample size, number of cells per person and read depth). We estimated the total cost C_t of a 10X Genomics experiment by adding up the library preparation costs, which depend on the number of used 10X kits times the cost of a kit C_k, and the sequencing costs, which depend on the number of used flow cells times the cost of one flow cell C_f. A classical 10X kit has 6 lanes, so the number of kits is determined by the total sample size n_s divided by the number of samples loaded on one lane $n_{s,l}$, a parameter the user can choose (dependent on how much he wants to overload one lane). Higher overloading creates more doublets, but can still be more cost-efficient. We modelled the effect of doublets in *scPower*, as they reduce the number of usable cells and reads (section 4.4.2). For the sequencing costs, the number of reads per flow cell determine the number of used flow cells. Combing all parts together, this leads to a cost function parameterized by the experimental costs as:

$$C_t = \lceil \frac{n_s}{6 * n_{s,l}} \rceil * C_k + \lceil \frac{n_s * n_c * r}{r_f} \rceil * C_f \tag{4.16}$$

Realistic cost estimations for 10X Genomics experiments, which are used in the following, can be found in Table 4.1.

During the power optimization, two of the experimental parameters can be freely chosen, and the third one is uniquely defined given the other two parameters and the overall budget. The cost function is always solved with regard to one of the parameters, e.g. the sample size n_s as:

$$n_s = \lfloor C_t / (\frac{C_k}{6 * n_{s,l}} + \frac{n_c * r * C_f}{r_f}) \rfloor \tag{4.17}$$

This leads to a grid of possible parameter options given a certain budget, as visualized in Figure 4.5 a,b for a DE study and an eQTL study, respectively. Figure 4.5 a depicts an example evaluation for a DE study with an experimental budget of 10,000€, choosing the priors as in the previous evaluations (expression prior from T cells and effect sizes from the CLL study [163]). The highest power was reached with a high number of 12,000 cells per sample measured in 4 samples and a medium read depth of 30,000 reads per cell. Increasing the number of cells raised mostly the expression probability, while the DE power was actually decreasing slightly, probably due to the lower number of samples that can be measured within the same budget (4.5 c). As the expression probability grew much faster than the DE power declined, the overall detection power increased with higher number of cells. A similar trend, but with weaker effects, is visible for the read depth. More reads increased the expression probability, but lowered the DE power (4.5 d).

The same optimization was applied for an eQTL study (4.5 b) with a budget of 30,000€ in total (effect sizes from the BLUEPRINT T cells study [43] with corresponding expression priors). Here, the optimal parameter combination was a medium number of 1,500 cells and read depth of 10,000, but a high sample size of 242. In contrast to the example DE study, here the eQTL power dropped drastically with lower sample sizes (Figure 4.5 e,f), so a balance between the eQTL power and the expression probability was identified that led to the highest overall detection power. The users can easily run a customized version of this analysis with the corresponding plots over our web tool `http://scpower.helmholtz-muenchen.de`, where they can specify the desired priors and budget.

4.4.2. Overloading of cells per lane

As mentioned before, the number of cells that can be loaded on a 10X Genomics lane is not fixed, but instead a further parameter that can be chosen by the user. Loading more cells on one lane reduces the costs, but it leads to more doublets, so a larger fraction of cells will be lost for the analysis. We added this effect explicitly to our model, following the approach as implemented by Hafemeister et al. for the website "How Many Cells" [154]: the doublet rate d is estimated as a linear function of the number of cells per lane, which are the product of the number of samples per lane $n_{s,l}$ and the number of cells per sample n_c, and a scaling factor. We derived this scaling factor from the 10X Genomics User guide for the version 3 chemistry [165] as $7.67 * 10^{-6}$. Combining this leads to a doublet rate d of

$$d = 7.67 * 10^{-6} n_c n_{s,l} \tag{4.18}$$

This reduces the number of usable cells per individual to $n_u = (1 - d) * n_c$. Assuming a cell type frequency of f_c for your cell type of interest, this leaves $f_c * (1 - d) * n_c$ cells for the analysis itself and so the power estimation.

Of note, we realized that the doublet rates estimated by other studies [159, 55] and

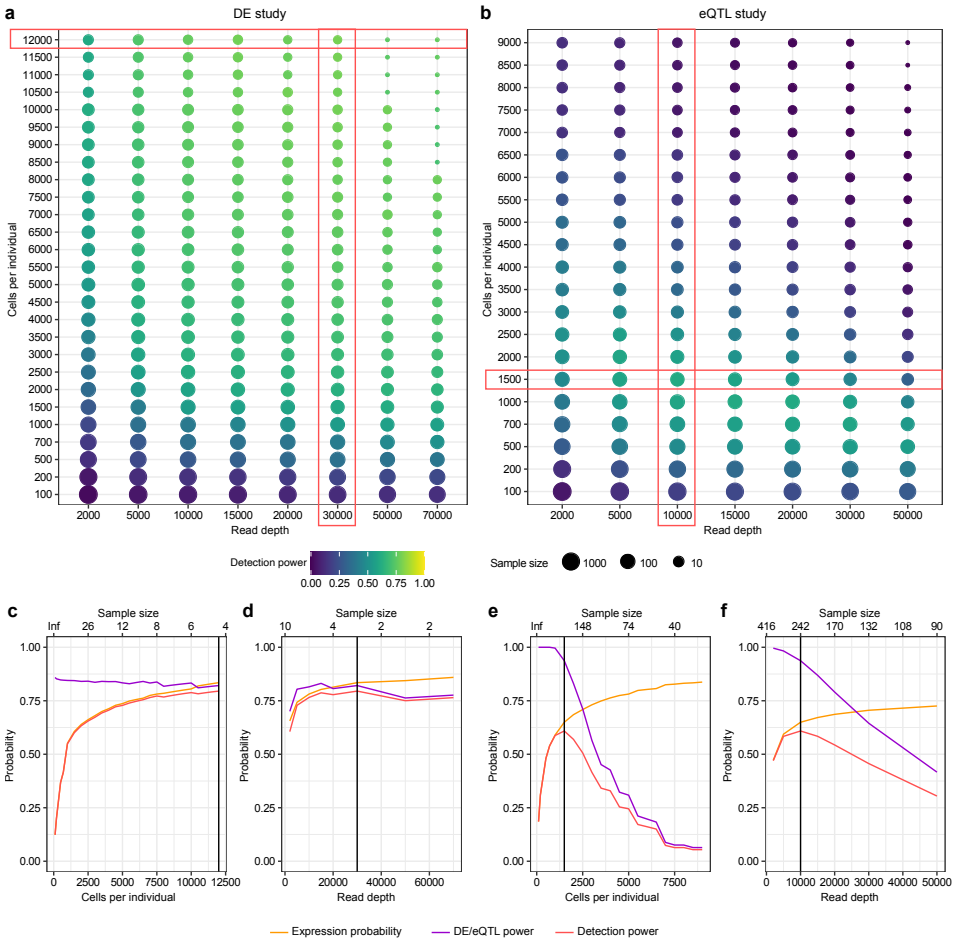


Figure 4.5.: **Parameter optimization for constant budget**
Maximizing detection power by selecting the best combination of cells per individual and read depth for a DE study with a budget of 10,000€ **(a)** and an eQTL study with a budget of 30,000€ **(b)**. Sample size is uniquely defined given the other two parameters due to the budget restriction and visualized using the point size. **c–f** Overall detection power dependent on cost determining factors. Influence of the cells given the optimized read depth **(c,e)** and of the read depth given the optimized number of cells **(d,f)**. **(c,d)** corresponds to the DE study in **(a)**, visualized in **(a)** by the red frames around the row with the optimal number of cells (for **(c)**) and around the column with the optimal read depth (for **(d)**). Same frames for **(e,f)** in the eQTL study **(f)**. The optimal sample size values are shown in the upper x axes for **(c-f)**. Vertical lines in the subplots mark the optimal parameter combination. Effect sizes were chosen as in Figure 4.4. Gene expression is defined as detected in >50% (DE analysis) or >9.5% (eQTL analysis) of individuals. Figure and legend taken from [1].

also by us, were above the doublets measured in the 10X Genomics user guide, so the exact doublets during overloading might be slightly higher than estimated in *scPower*, but we decided to use the official reference numbers in our tool.

Another point to take into consideration for the overloading is the effect on the read depth: during the sequencing, more reads will originate from doublets as from singlets, because doublets contain two cells. For this reason, the mean number of reads is diminished the more doublets are measured. We again followed the approach of Hafemeister et al. to model this [154]. They estimated a doublet factor f_d of 1.8 to describe the rate of reads from doublets and reads from singlets. The number of reads for singlets r_s will be reduced compared to the target read depth r as

$$r_s = \frac{rn_c}{n_u + f_d(n_c * n_u)} \tag{4.19}$$

The total number of usable reads per cell in the end depends on the mapping efficiency, which we set to 80%, leaving $0.8 * r_s$ reads in the power analysis.

For all budget analyses, we fixed the number of cells per lane $n_{c,l}$ instead of samples per lane $n_{d,l}$ to control the overloading rate directly. Then, the number of samples per lane is $n_{s,l} = \lfloor n_{c,l}/n_c \rfloor$, expecting that all cells from one sample are measured together on one lane and are not split into multiple lanes. If most doublets are successfully detected and the analysis is not falsified by them, overloading the cells is more cost-efficient, despite loosing some cells and reads. We validated the doublet detection in our pilot dataset using sex-specific genes and saw that their expression was very concordant with the sex of the donor after the removal of doublets (Supplementary Figure A.9 b). For this reason, we assume that a combination of different doublet detection tools performs well enough to identify most doublets. For the previous budget analysis in subsection 4.4.1 and all following budget analyses, we overloaded the lanes in our power analyses with 20,000 cells per lane, which leads to a doublet rate of 15.4%.

4.4.3. Exploring optimal parameters for different budgets

Building on this implementation for power optimization, we explored which experimental parameters were increasing if a higher budget was available (Figure 4.6). We included a large variety of different priors, exploring simulated priors (Figure 4.6 a,b) additionally to the observed priors used before (Figure 4.6 c,d). The simulated priors allowed us to explore the difference between high and low effect sizes as well as high and low expression ranks in specific prototypic scenarios. The same analysis was done for the DE cases (Figure 4.6 a,c) and the eQTL cases (Figure 4.6 b,d).

The detection power (first column of Figure 4.6) rises in all scenarios with higher budget, but the difference between the effect sizes is clearly observable. The detection power was higher for the same budget when the effect sizes are larger and/or the expression ranks are higher. The number of cells (second column) is either at the chosen maximum level already at the beginning or rises clearly when more budget is available, highlighting the value of a large number of cells for the detection power. For the eQTL

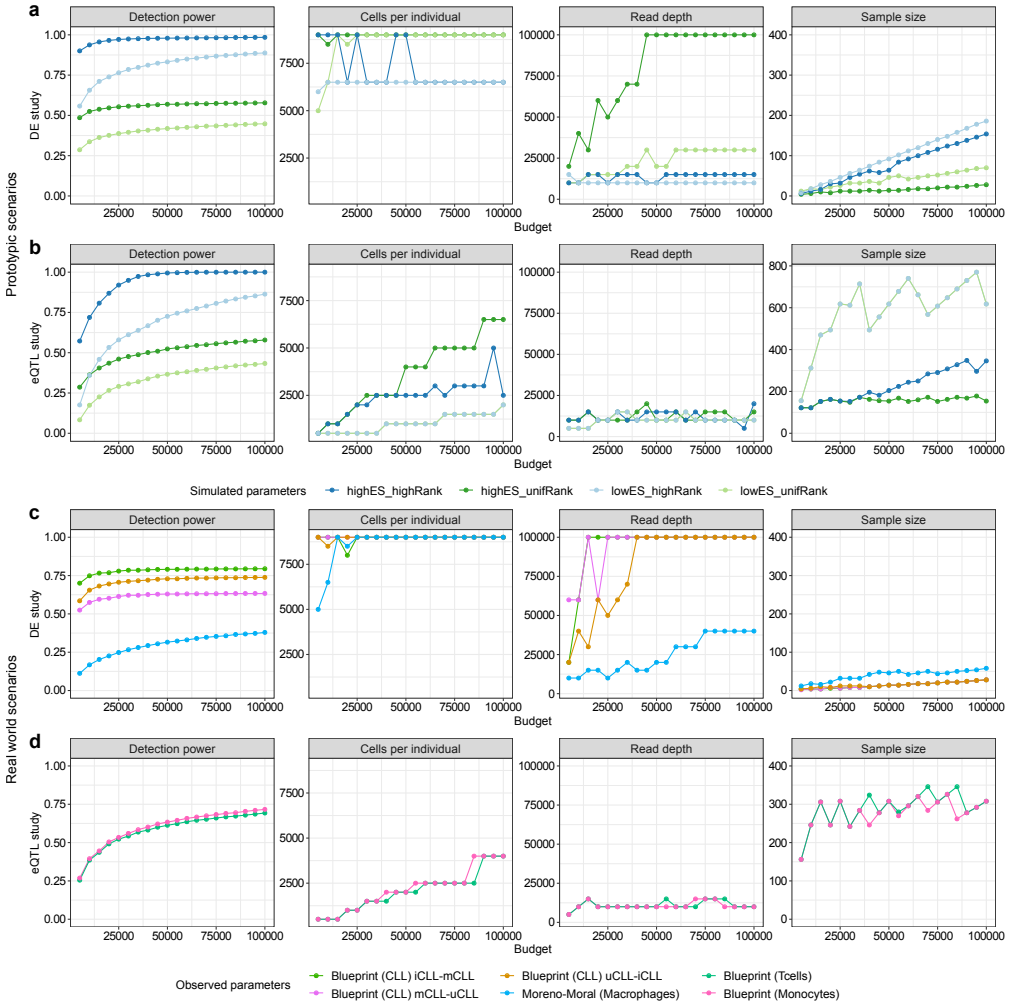


Figure 4.6.: **Optimal parameters for varying budgets and 10X Genomics data.** The maximal reachable detection power (column 1) and the corresponding optimal parameter combinations (columns 2–4) change depending on the given experimental budget (x-axis). The coloured lines indicate different effect sizes and gene expression rank distributions. Different simulated effect sizes and rank distributions for DEG studies **(a)** and eQTL studies **(b)** with models fitted on 10X PBMC data. highES = high effect sizes, lowES = low effect sizes, highRank = high expression ranks and unifRank = uniformly distributed expression ranks (always relative to effect sizes observed in published studies). Effect sizes and rank distributions observed in cell type sorted bulk RNA- seq DEG studies **(c)** and eQTL studies **(d)** with model fits analogously to **(a,b)**. Expression thresholds were chosen as for Figure 4.5. Figure and legend taken from [1].

cases, the number of cells is rising, but not as extreme as for the DE cases, probably because here a high sample size is essential for a high detection power, especially for low effect sizes (fourth column, Figure 4.6 b). The read depth is increased in the DE cases before the sample size, showing that for the chosen priors a relative small sample size is sufficient (third column, Figure 4.6). In contrast, shallow sequencing of a large number of cells and samples is most beneficial for the eQTL cases, as already suggested by Mandric et al. [160].

4.5. Generalization of scPower to other scRNA-seq technologies and tissues

While all previous analysis were based on 10X Genomics, one of the most frequently used single cell transcriptomics platforms, *scPower* can be likewise applied for power analysis and design optimization of experiments performed with other technologies. Other microfluidics-based methods can be used fully interchangeably, using expression priors fitted on a corresponding pilot dataset from the same technology. We showed this based on a lung cell dataset measured with Drop-Seq [72]. The only adaption we did here was setting the doublet rate to a constant factor, since we did not have the information on how overloading influences the doublet rate for Drop-seq.

Additionally, we chose a Smart-seq2 dataset from pancreas for testing *scPower* [166]. It is a plate-based method based on read counts instead of UMI counts, so a few more changes in our model were required. We adjusted for the gene-length bias by defining the expression threshold relative to one kilobase of the transcript, so taking a threshold of $\frac{n*1000}{l_i}$ for a gene of length l_i. These normalized counts are used to estimate the gamma mixed distributions, which are parameterized directly dependent on the reads per cell. The doublet rate was set to a constant factor, matching the technological differences to microfluidic-based systems with overloading.

For both datasets, the lung dataset measured with Drop-Seq and the pancreas dataset measured with Smart-seq, we fitted expression priors and validated the expression probabilities (Supplementary Figure A.17). It showed high concordance between the estimated and observed number of expressed genes ($r^2 = 0.995$ for the lung Drop-seq dataset and $r^2 = 0.991$ for the pancreas Smart-seq dataset). Furthermore, the power estimations for both technologies matched well the results of the simulation-based tools *powSimR* and *muscat* (Supplementary Figure A.18).

With the validated expression priors, we repeated the parameter optimization for different budgets, again with simulated and observed effect sizes (Figure 4.7; estimated costs in Table 4.1). For the simulated priors, we chose the same prototypic scenarios as in Figure 4.6. The observed effect sizes were taken from FACS sorted DE studies, covering asthma in lung cells [122] and aging in pancreas [167].

Across the different scenarios, the overall detection power for Smart-seq2 was not as high as it was for the same budget for Drop-seq (Figure 4.7). The Drop-seq curves showed in general very similar outcomes as for the 10X PBMC dataset, probably due

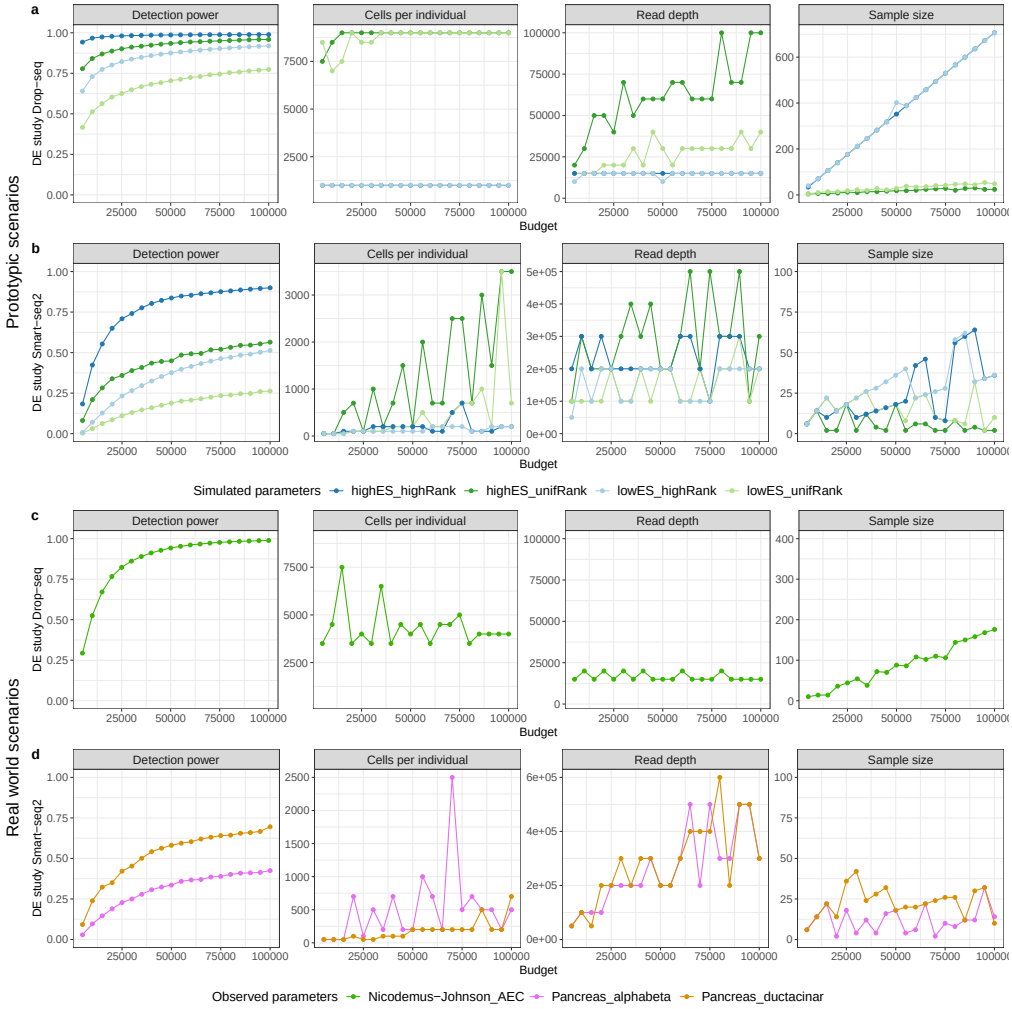


Figure 4.7.: **Optimal parameters for varying budgets and Drop-seq and Smart-seq2 data**
Maximal reachable detection power (y-axis, first column) for a given experimental budget (x-axis) and the corresponding optimal parameter combinations for that budget (y-axis, second till fourth column). The colored lines indicate different effect sizes and gene expression rank distributions. **a,b** Different simulated effect sizes and rank distributions for DE studies with models fitted on Drop-seq lung data **(a)** and Smart-seq2 pancreas data **(b)**. highES = high effect sizes, lowES = low effect sizes, highRank = high expression ranks and unifRank = uniformly distributed expression ranks (always relative to effect sizes observed in published studies). **c,d** Effect sizes and rank distributions observed in cell type sorted bulk RNA-seq DE studies with model fits analogously to **(a-b)**. Figure and legend taken from [1].

Technology	Library preparation costs per cell	Sequencing costs per 1 million reads
10X Genomics	0.05€ - 0.12€	3.42€
Drop-Seq	0.09€	3.42€
Smart-Seq2	13.00€	3.42€

Table 4.1.: **Experimental cost per technology.**
Library preparation cost estimation (per cell) and sequencing cost estimation (per 1 million reads) for three of the most common single cell RNA-seq technologies in Euro (€). For 10X Genomics, the cost depends on the number of cells per lane, an overloading of each lane with 20,000 cells generates costs of 0.05€ per cell, a loading with 8,000 cells per lane costs of 0.12€ per cell. Table and legend taken from [1].

to the large technological similarities. The number of cells per individual was the first parameter to be increased in most cases. In contrast, far fewer cells were determined as optimal for Smart-seq2 experiments and the sample size was far lower. Instead, the number of reads was increased.

We observed that the restriction of Smart-seq2 is not that the experiment is less powerful, but that it is less cost-efficient: measuring large number of cells and samples is far more expensive according to our estimated costs per cell (Table 4.1). In general, the evaluation showed that high-throughput technologies, such as 10X Genomics and Drop-seq, which can cheaply process large quantities of cells, are well suited for these large-scale multi-sample single cell transcriptomics experiments.

4.6. Project summary and outlook

In this chapter, we described *scPower*, a power analysis method for multi-sample single cell transcriptomics experiments that covers both interindividual DE analysis and eQTL analysis. The tool is specifically designed for single cell experiments because it takes the sparsity of the data into account by modelling the expression probability of a gene depended on the experimental parameters and the cell type prior. We validated the accuracy for different tissues and technologies, proving its generalizability to new settings.

scPower - as an analytic method - allows fast and memory-efficient evaluation of different scenarios in contrast to alternative simulation-based tools. This allowed us to incorporate budget optimization, which selects the optimal experimental design for a certain budget, without any large computational costs. In the end, we envision that *scPower* facilitates for all users the planning of powerful experiments, which are crucial for increasing the number of significant discoveries in future single cell DE and eQTL studies.

Run	Donors	Target number of cells	Target reads per cell	Number of cells	Mean reads per cell
Run 1	1 - 14	8,000	50,000	7,491	40,650
Run 2	1 - 7	8,000	50,000	5,989	127,685
Run 3	8 - 14	8,000	50,000	8,144	13,949
Run 4	1 - 14	8,000	50,000	7,429	35,417
Run 5	1 - 14	8,000	50,000	7,765	21,057
Run 6	1 - 14	25,000	50,000	20,126	51,792

Table 4.2.: **Experimental parameters of the 6 PBMC runs.**
In Run 1, 4, 5 and 6 all 14 donors were measured, in Run 2 only donor 1-7 and in Run 3 only donor 8-14. Run 6 was overloaded with 25,000 cells. The number of cells and mean reads per cell are taken from the cell ranger summary statistics. Table and legend taken from [1].

4.7. Materials and additional methods

4.7.1. Generation of the single cell RNA-seq PBMC dataset

The PBMC dataset that we used as a pilot dataset while building and evaluating *scPower* was generated by our collaboration partners at the Max Planck Institute of Psychiatry, Munich. It comprised 14 healthy individuals (7 male and 7 female) of the Biological Classification of Mental Disorders study (BeCOME; ClinicalTrials.gov TRN: NCT03984084) [157]. Six 10X Genomics runs (Chromium Single Cell 3 library and gel bead kit v2) were generated with different distributions of individuals, five aiming for 8,000 cells and one overloaded aiming for 25,000 cells (Table 4.2). The target read depth for all runs was 50,000 reads per cell, using HiSeq4000 from Illumina (150bp paired-end sequencing). Further information on the recruitment and dataset generation are found in the clinical trail protocol [157] and in the publication of *scPower* [1].

4.7.2. Processing the PBMC dataset

The sequenced reads from the 10X Genomics experiments were processed with *Cell-Ranger* (v2.0.0 and v2.1.1) [70], using the hg19 reference genome for mapping. We annotated cells to individuals using *Demuxlet* (v1.0) [159] after discarding doublets identified by either *Demuxlet*, which captures only doublets between individuals, and *Scrublet*, which captures all neotypic errors, i.e. doublets arising from cells with different expression profiles, e.g. originating from different cell types [168]. For *Demuxlet*, genes annotated as "doublet" or "ambivalent" were removed, for *Scrublet*, all cells above a doublet threshold of 0.28. Additionally, we discarded cells with less than 200 genes or more than 2,500 genes and with more than 10% counts from mitochondrial genes, following best practice suggestions for setting the thresholds [169]. We processed the single cell data for the filtering and the following cell type identification with *Scanpy*

Cell type	Markers
CD4+ T cells	IL7R, CD3D
CD14+ Monocytes	CD14, LYZ
B cells	MS4A1, CD79A
CD8+ T cells	CD8A, CD8B, CD3D
NK cells	GNLY, NKG7
FCGR3A+ Monocytes	FCGR3A, MS4A7
Dendritic cells	FCER1A, CST3
Megakaryocytes	PPBP
Plasma cells	CD79A

Table 4.3.: **Marker genes for cell type identifcation.**
Marker genes used to assign the Louvain clusters to the cell types. Annotations taken from van der Wijst et al., 2018 [55]. Table and legend taken from [1].

(v1.4) [170].

4.7.3. Cell type identification

After the initial filtering of cells, we annotated cells to different cell types. We started by discarding genes counted in less than 3 cells, normalized cells to 10,000 counts per cell and logarithmized them. We restricted the analysis to highly variables genes, followed by regression of total counts per cell and the mitochondrial percentage. We clustered cells by taking the first 40 principal components of the PCA on all cells, identifying the nearest neighbor graph with these, and performing Louvain clustering on the graph [171]. In the end, the Louvain clusters were assigned to nine main PBMC cell types based on marker gene expression (Table 4.3), as applied in van der Wijst et al., 2018 [55].

4.7.4. Subsampling counts for expression probability model

To parameterize our expression probability model on the read depth, we downsampled the fastq files of the 6 runs from the 10X Genomics PBMC pilot dataset to 75%, 50% and 25% of th reads using fastq-sample from *fastq-tools* (v0.8) [172]. Count matrices were again obtained by running *CellRanger* and all annotations - donors, doublets and cell types - were transferred from the results of the full count matrix.

4.7.5. Fitting expression prior from pilot data

The formulas to estimate the expression probability for a set of experimental parameters - sample size, number of cells and read depth - and the rationale behind are described in sections 4.2.4, 4.2.5 and 4.2.6.

We fitted an expression prior to validate our model on the processed PBMC pilot dataset. First, we estimated negative binomial distributions for each gene and cell type using *DESeq* [104]. Following this approach, the raw counts from the filtered UMI count matrix were normalized for library size using either the *DESeq* standard method or "poscounts" from *DESeq2* [105] that was specifically designed for sparser data. The normalization method was selected based on the quality of the fit. We chose the standard normalization for the 10X Genomics PBMC dataset, but "poscounts" for the Drop-seq lung dataset and the Smart-seq2 pancreas dataset. We focused on cell types with at least 50 cells for robust parameter estimation and fitted each batch separately to avoid overdispersion by batch effects.

Afterwards, we fitted a mixture distribution over all gene-wise mean values. The mixture distribution was made of a zero-component and two left-censored Gamma distributions, each censored at $1/n_c$ with n_c the number of cells in the cell type, as described in section 4.2.4. One mixture distribution is obtained for each cell type and subsampling step, represented in seven parameters: three parameters for the proportions between the distributions p_1, p_2 and p_3 and four parameters describing the two Gamma distributions (two mean and standard deviations).

The different mixture distributions for one cell type are combined when parameterizing the model via the mean UMI counts per cell: a linear regression is fitted for each mean and standard deviation of the two gamma distributions to explain it via the mean UMI count. The probability parameter p_3 that describes the frequency of the second Gamma distribution of highly expressed genes is kept constant, independent of the subsampling step. Here, the mean value over the different subsampling runs is chosen. The probability parameter p_1 is also modelled via a linear regression, with a minimal value of at least 0.01 to allow no negative values. The third mixture proportion can be inferred from the other two as $p_2 = 1 - p_1 - p_3$.

To incorporate the read depth in the end, we fitted a logarithmic curve to model the relationship between the mean UMI counts per cell n_{UMI} and the number of transcriptome mapped reads r as $n_{UMI} \sim log(r)$. This results in one curve per dataset. The UMI count - read depth function is difficult to transfer to other settings, as it is affected by many biological and technical factors. Nevertheless, the collection of functions from *scPower* help to assess how sequence saturation curves typically look like in different example studies.

The required mean-dispersion function is taken directly from *DESeq*. As the parameters are not affected by the mean UMI counts, i.e. are nearly constant over the subsampling steps, we average the parameters of the dispersion function across all runs and subsampled runs for one common mean-dispersion curve per cell type.

4.7.6. DE power

For the DE power, we used the analytic power analysis method for negative binomial models as implemented in the R package *MKmisc* [161]. The authors provided three different approaches, from which the third proved to be most accurate in their publication.

For this reason, we used method 3 for the power calculations. The mean and dispersion parameter for the power calculation are inferred based on our expression probability model (section 4.2.5), the dispersion parameter is thereby assumed to be the same for both groups. Other input parameters are taken from priors, in case of the fold change, or are chosen directly by the user, in case of the sample size, the significance threshold and the sample size ratio between both groups. In the following examples, a balanced design with a sample size ratio of 1 is chosen, but other ratios are possible without any adaptations.

4.7.7. eQTL power - Analytic power calculation

The eQTL power analysis is based on linear regression, the most frequently used approach here. In order to apply linear regression models, the residuals need to be normally distributed. This is ensured for the negative binomial counts through log-transformation which results in a constant variance independent of the mean [152]. However, we observed that the corresponding analytic power analysis method for linear regression models achieved only reliable results for genes with large mean values, which were obtained from simulations (both methods described in detail below; Supplementary Figure A.14). For lowly expressed genes, the analytic power was overestimated compared to simulation-based results, as the log-transformation did not work properly due to many zero values. For this reason, *scPower* applies the analytic power analysis only for genes with an expression mean > 5 and simulation-based power analysis for all other genes, which were too lowly expressed. The expression threshold of 5 was chosen based on the aforementioned comparison.

The analytic power analysis method is based on the F-test, as implemented in the function pwr.f2.test from the R package *pwr*. Effect sizes are quantified by the coefficient of determination R^2. It can be derived from a pilot study, using the regression parameter β, its standard error $se(\beta)$ and the study sample size N:

$$t = \frac{\beta}{se(\beta)} \tag{4.20}$$

$$R^2 = \frac{t^2}{N - 2 + t^2} \tag{4.21}$$

Other input parameters for the power are the sample size of the newly planned study n_s and the chosen significance threshold α.

4.7.8. eQTL power - Simulation for small means

The eQTL power simulation framework was implemented to benchmark the analytic power analysis solution (section 4.7.7) and is used in the final version of *scPower* for lowly expressed genes with means < 5, as discrepancy were visible here.

The same input parameters are required as for the analytic method, the effect size R^2, the planned sample size n_s and the significance threshold α. On top of that, the mean

count value μ_c of the gene is necessary, assuming that this is the mean of the genotype with the lower expression. The following simulation is performed 100 times to ensure a stable estimate:

First, the allele frequency f_a is sampled from a uniform distribution with the interval $[0.1, 0.9]$. The corresponding genotype vector g with $g_i \in \{0, 1, 2\}$ of length n_s is created, by following the Hardy-Weinberg equilibrium with f_a^2 times 0, $2 * f_a * (1 - f_a)$ times 1 and $(1 - f_a)^2$ times 2.

Next, the regression parameter β and its standard error $\hat{\sigma}$ can be inferred as:

$$\beta = \sqrt{\frac{R^2}{2f_a(1 - f_a)}} \tag{4.22}$$

$$\hat{\sigma} = \sqrt{1 - R^2} \tag{4.23}$$

The count vector x is sampled dependent on the genotype vector g. For each element x_i, the simulated count value is drawn from a negative binomial distribution with mean μ_i and dispersion ϕ_i. The mean is shifted dependent of the genotype:

$$\mu_i = e^{log(\mu_c) + \beta * g_i} \tag{4.24}$$

The dispersion parameter ϕ_i needs to be set in a way that variance of the log transformed counts is $\hat{\sigma}$. The Taylor approximation of the parameter did not work due to the small mean values [152], therefore the dispersion was estimated based on numerical optimization. To accelerate this part, *scPower* contains a table with precalculated values.

A linear regression model was calculated for the log-transformed simulated counts (plus one pseudocount) as $log(x_i + 1) \sim g_i$. The resulting p-value p_i for $H_0 : \beta = 0$ was used to estimate the simulation based power as

$$\sum_{i=1}^{B} (p_i < \alpha) \tag{4.25}$$

4.7.9. Overall detection power

As stated in section 4.2.3 in detail, the expression probability and the DE/eQTL power are combined to the overall detection power via multiplication, as both are conditionally independent given the expression mean and dispersion. For each DE/eQTL gene, the overall detection power is calculated as follows: from the pilot dataset, not only the effect size were taken as prior, but also the expression rank i. The mean expression of the gene μ_c was estimated as the quantile $i/|G|$ of the gamma mixture model parameterized by the mean UMI counts, with G the set of all genes. In our example cases, we focused on a total gene set of G of size 21,000. Based on the mean μ_c, the corresponding dispersion ϕ_c can be inferred from the expression prior and together the expression probability is calculated. For the DE/eQTL power, μ_c and ϕ_c are again required, together with the effect size of the gene and the significance threshold α. Importantly, α needs to be

corrected for multiple testing (more in section 4.2.7). We applied FWER adjustment for the eQTL scenarios and FDR adjustment for the DE scenarios, always using an α value of 0.05. Afterwards, both the expression probability and DE/eQTL power are multiplied to the overall detection power of the respective gene. The overall detection power of the complete experiment is then the average over the genewise overall detection power values.

4.7.10. Processing public pilot data sets for priors

Priors, which contain effect sizes and expression ranks of DE/eQTL genes, are required for the power calculations. We chose different public datasets to get realistic estimates, focusing on FACS-sorted studies for cell-type specific results [163, 164, 122, 167, 43]. In every study, expression ranks were obtained from FPKM normalized values and FDR<0.05 was chosen as the significance threshold for the DE/eQTL definition. When available, we took the published effect sizes, otherwise we reran the analysis with DEseq2 [105].

4.7.11. Adaptions for powsimR and muscat

The simulation-based power analysis methods *powsimR* and *muscat* were applied on the CD4+ T cells of the PBMC pilot dataset for different parameter combinations to validate the analytic power estimations from *scPower* [112, 113]. Both tools were run with 25 simulation rounds for accurate power estimations. The simulated count matrices were analyzed with the pseudobulk approach in combination with established DE methods. For powsimR, the median-ratio normalization of DESeq2 was used followed by DE identification with edgeR-LRT, DESeq2 and limma-voom. For muscat, DE identification was performed using edgeR, DESeq2, limma-voom and limma-trend. Both for *powsimR* and *muscat*, small adaptions were necessary to make their results comparable with *scPower*. Input parameters were changed so that the methods took a log-fold change vector together with a matching expression rank vector, instead of randomly selecting genes as DE genes. Furthermore, a pseudobulk version was added in case of *powsimR*, as it was not design for multi-sample power analysis.

To enable pseudobulk calculations with *powsimR*, we added a sample size parameter n_s for a balanced design ($n_s/2$ samples per group). The individual level effect sizes were set to the same values as the cell level effect sizes. After simulating the count matrix, the cells were distributed equally between the samples under consideration of the group structure. The counts per sample were summed up to the pseudobulk matrix, which was afterwards processed the same way as the single cell matrix for DE calling in *powsimR*.

muscat was run with specific settings to increase comparability with *scPower*: originally, one negative binomial distribution is fitted per sample in *muscat* to capture donor-specific differences. However, for our dataset, the reduced number of cells for the negative binomial fits decreased the number of expressed genes drastically and provided far

fewer genes for the simulation. For this reason, we decided to fit only one negative binomial distribution over all samples together, matching the approach of *powsimR* and *scPower*. Another point to consider was the DE simulation: *muscat* allows the simulation of more complex differential expression scenarios, for the comparison we used only the classical "DE" scenario with shifted mean expression between groups.

Accuracy of *scPower* was estimated using three different criteria: first, the expression probability of *scPower* was evaluated by comparsion of the expected number of expressed genes from *scPower* with the simulated number of expressed genes from *powsimR* and *muscat*. The expression probability threshold of *scPower* was set to more than zero minimal counts in order to match the expressed genes of *powsimR* and *muscat*.

Second, the DE power between the methods was compared. For *powsimR* and *muscat*, only the power of expressed genes can be estimated, because the simulated counts are required. For this reason, we restricted the DE power for *scPower* also to expressed genes by selecting only DE genes with an expression rank smaller than the expected number of expressed genes.

Third, the overall detection power is compared. This value is not automatically given by *powsimR*, as it focuses on the DE power for expressed genes and overestimates so the general experimental power. However, it can be derived easily, when by averaging the DE power over all simulated genes (including the not expressed ones, setting their DE power to 0). The same strategy was applied for *muscat*.

4.7.12. Evaluating doublet rate using sex-specific genes

In order to validate the donor assignment of *Demuxlet* and the doublet detection we performed using *Demuxlet* and *Scrublet*, we tested if sex-specific genes were only expressed in cells whose donor had the right sex. For male cells, we calculated the fraction of cells with an expression of Xist > 0 as the male-specific error. For female cells, we calculated the fraction of cells with more reads mapped to the Y chromosome than the q_f quantile of all cells as the female specific error. q_f is the fraction of cells assigned to a female donor among all cells. We normalized the reads mapped to chromosome Y in this step using transcripts per million (TPM), i.e. counting all reads mapped to chromosome Y except reads mapped to pseudoautosomal regions of the chromosome divided by the total read counts per cell times 1,000,0000. The female-specific error rate was defined a bit more lenient, as also in female cells, mis-mapping of single reads to the Y chromosome happens. We evaluated the male-specific and female-specific error once before the removal of doublets and once after, comparing if the values improved.

4.7.13. Generation of prototypic scenarios for budget evaluation

We extended our set of tested priors for the budget optimization to create additional prototypic scenarios. For this, we simulated effect sizes and expression ranks of different magnitudes, both for the DE and eQTL analysis. Specifically, we tested effect sizes that were higher or lower than our observed effect sizes from the prior studies and

expression ranks that were higher or lower. In each setting, we simulated priors for 250 DE genes or 2000 eQTL genes, respectively.

We sampled the DE log fold changes from a normal distribution with a standard deviation of 1, for the high effect sizes with a mean of 2 and for the low effect sizes with a mean of 0.5. R^2 values for eQTL scenarios were sampled from normally distributed Z-scores with a mean of 0.5 for high effect sizes and 0.2 for low effect sizes. The standard deviation was always 0.2. In each case, the normal distribution was truncated to contain only values above the mean, matching the observed distributions from the pilot datasets, and the values were afterwards transposed via inverse Fisher Z Transformation. The expression ranks sampled from an uniform distribution over the first 10,000 genes for the high ranks and over the first 20,000 genes otherwise. We combined effect sizes and rank distributions to four scenarios in total, called "highES_highRank", "lowES_highRank", "highES_unifRank" "lowES_unifRank". The abbreviations "highES" and "lowES" stand for high and low effect sizes, the abbreviations "highRank" and "unifRank" for a high rank distribution and a uniform rank distribution.

4.7.14. Applying scPower to Drop-seq and Smart-seq2 data

The general adaptions to run *scPower* with a Drop-seq and a Smart-seq2 dataset are already described in section 4.5. In the following, more details on how to deal with the gene length bias for Smart-seq2 datasets are given: negative binomial distributions were fitted on the raw counts without gene length normalization. Afterwards, the length normalized mean values were taken to fit the gamma mixed distribution. In order to derive the gene-length dependent mean and dispersion values for a newly planned experiment, the gene length was taken as an additional prior from the reference datasets together with the effect sizes and the expression ranks. Another change for the Smart-seq expression model is that the mean-dispersion function parameters displayed a linear relationship with the read depth for this dataset, in contrast to the droplet-based methods, which we additionally modelled for this reason.

Subsampling was applied again to incorporate the effect of the read depth. For the Drop-seq dataset, the same approach could be used as for the 10X Genomics dataset with fastq-tools [172] (see section 4.7.4). Following this, the pipeline described in [68] was used to create the count matrix. For the Smart-seq dataset, the step of modelling the relationship between the UMI counts and read depth could be skipped. Therefore, the read count matrix could be downsampled directly via the function downsampleMatrix from the R package *DropletUtils* [173].

A necessary assumption for our cohort level expression model is that the cell type frequencies of the different individuals are all in the same range. Further simulation studies for this and adjustment possibilities are shown in the publication [1]. Unfortunately, this assumption was not true for both the Drop-seq and the Smart-seq dataset and the cohort level expression probability could not be calculated accurately here. Instead, we used a simplified version of the expression probability: a gene is expressed if it has a certain number of counts per cell type over all individuals together.

5. Analyzing genetic influence on personalized networks with single cell transcriptomics

5.1. Advancing genetic variant interpretation through co-expression QTLs

The last two chapters 3 and 4 covered how population studies can be used to interpret the downstream effects of genetic variants, and how single cell data can improve these analyses, because they facilitate cell type specific analyses. On top of that, single cell data allows additional new approaches that were not possible with bulk (more in introduction chapter 1). One of these novel approaches is the identification of co-expression QTLs (co-eQTLs), genetic variants that influence the co-expression of two genes [55, 83]. Single cell data provides multiple measurement points per individual (and cell type), so the necessary individual specific co-expression values can be inferred from it. Previous analyses in bulk with only one measurement point per individual tried to overcome this with linear models including an interaction term of genotype times gene [46]. However, the bulk approach requires cohorts with extremely large sample sizes to gain enough power. Additionally, cell type specific analyses are generally difficult with bulk data, as discussed already before. For these reasons, performing co-eQTL studies in single cell data is a large improvement.

The identified co-eQTLs allow connecting the effects of genetic variants with the complex gene regulatory network. Standard eQTL analyses identify the downstream consequences a genetic variant has on the expression of a gene. However, the upstream regulatory processes that are disturbed by the variant can often not be pinpointed. In contrast, co-eQTLs identify which gene associations are changed dependent on the genetic variants and find so also the corresponding upstream regulators. Especially for disease-associated variants, it supports the identification of disease-associated biological processes and might in the future also help for drug development and personalized medicine. For example, if a co-eQTL affects an edge in the gene regulatory network that is crucial for the effectiveness of the drug, patients can be divided into two groups based on the co-eQTL: one group, where the drug will be effective, and one group, where it will not [174].

Recently, the first few small-scale and targeted co-eQTL analyses have been performed with single cell data [55, 83], but many open questions remain: first, the development of gene regulatory networks from scRNA-seq data is a very active research field with

many new algorithms in the last years [175, 176, 177]. However, a recent benchmarking study showed that the performance is very dataset specific [87]. In general, these evaluations have the issue that no ground-truth network exists for comparison, making all benchmarking studies difficult. A specific analysis that covers which association metric should be used for the co-eQTL identification is currently lacking and would be beneficial.

Second, the best workflow for the co-eQTL mapping itself has not been established. The search space of all possible triplets (SNP - gene - gene) is even larger than for a classical eQTL analysis and requires proper multiple testing correction. This limits however the detection power. To identify still many co-eQTLs, either large sample sizes are required, which are so far limited in the first single cell cohorts, or sophisticated filtering strategies to test only those triplets that are probably most relevant or interesting.

Third, the interpretation of co-eQTLs has never been done systematically before. Strategies how to best annotate the upstream regulatory processes and the regulating transcription factors need to be developed.

In this study, we addressed those questions and explored how to best identify and interpret co-eQTLs by conducting a cell type specific co-eQTL meta analysis with 173 scRNA-seq samples from PBMCs. For, this we combined three studies, naming each in the following after the first author of the study: the Oelen study with 104 healthy individuals [83], the van der Wijst study with 45 healthy individuals [55] and the van Blokland study with 38 individuals 6-8 weeks after hospital admission due to a heart attack [178] (Figure 5.1 a). All three studies measured PBMCs with 10X Genomics. Because the Oelen and the van Blokland study analyzed part of the samples with version 2 chemistry from 10X Genomics and part with version 3, we split both studies in two separate datasets each (called Oelen v2 and v3 dataset, and, respectively, van Blokland v2 and v3 dataset). The van der Wijst study was measured completely with version 2 chemistry.

Preprocessing of each dataset was done already in the corresponding studies, including cell type annotation. We performed all analyses in a cell type specific way, splitting the dataset into the six major cell types: CD4+ and CD8+ T cells, natural killer (NK) cells, monocytes, B cells and dendritic cells (DCs) (Figure 5.1 a).

We chose Spearman correlation as the association metric for the co-eQTLs as it showed robust results when comparing between different single cell and bulk datasets. Furthermore, highly correlated gene pairs matched well with associations identified from a CRISPR knock-out dataset (Figure 5.1 b). Comparison of Spearman correlation values between the cell types and between individuals within each cell type gave us insights into the general observed correlation structure identified with scRNA-seq.

For the co-eQTL mapping itself, we combined a stringent filtering of tested triplets with a permutation-based multiple testing correction, to increase the robustness of our results. For the filtering, first we identified cis eQTLs in the dataset and focused on these pairs of eQTL SNP and the associated gene, which is called eGene in the following. Then, we tested the SNP-eGene pair together with all other genes that are significantly

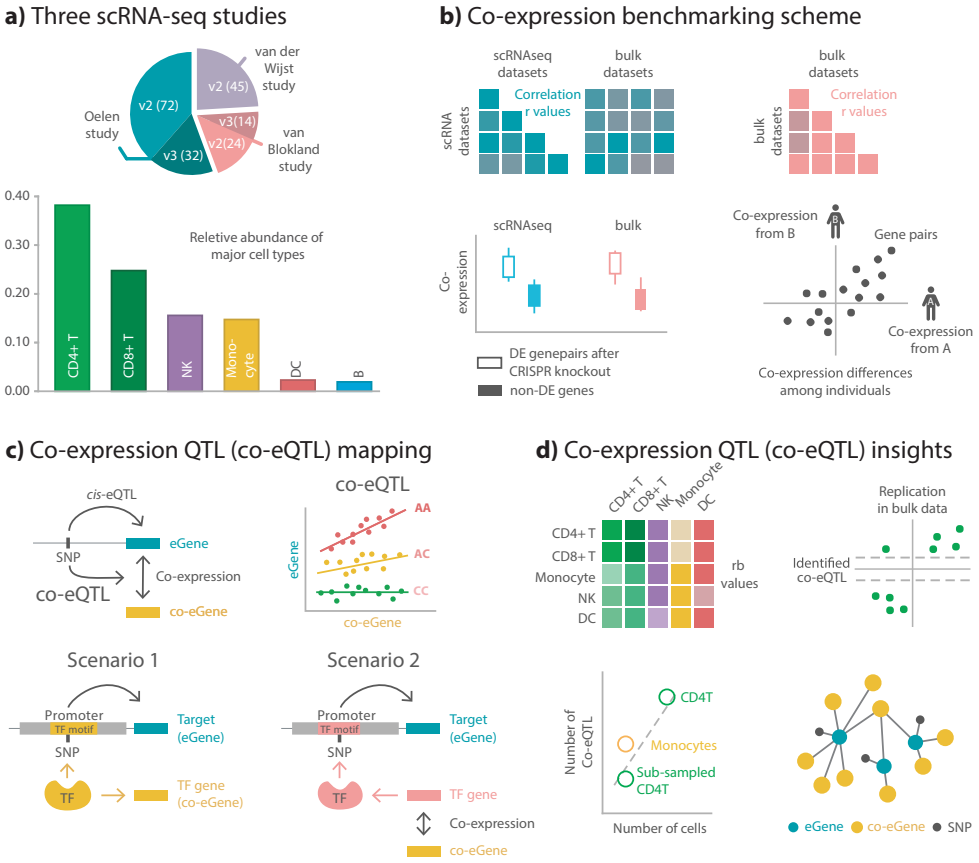


Figure 5.1.: **Study Overview**

a) Overview of the three PBMC scRNA-seq studies used in our analysis, including the chemistry version (v2/v3), number of individuals involved (number in the parenthesis), and relative composition of the major blood cell types used in this study. **b)** We first benchmarked co-expression patterns among the scRNA-seq studies and compared them to co-expression pattern in different bulk datasets and associations from a CRISPR knockout dataset. After benchmarking, we evaluated differences in co-expression patterns among cell types and among individuals within a cell type. **c)** Cell type specific co-eQTL mapping based on a novel strategy with strict filtering of tested SNP–eGene–co-eGene triplets: the SNP is required to be an eQTL for one of the genes and the genes show significant correlation in at least a certain number of individuals. **d)** Identified co-eQTLs were replicated in a bulk dataset. To evaluate technical influences, we assessed the impact of cell number, number of tests and the number of individuals on the number of significant co-eQTLs. Lastly, we interpreted the biological relevance of the co-eQTLs. Figure and legend taken from [3].

correlated with the eGene in at least 10% of the individuals. These selected genes are called co-eGenes in the following (Figure 5.1 c). This allows a large-scale co-eQTL analysis with sufficient power by decreasing the search space.

The identified co-eQTL associations can be caused by different biological scenarios that need to be distinguished for the interpretation (Figure 5.1 c). In Scenario 1 of Figure 5.1 c, it is depicted how co-eQTLs can represent upstream regulators of eQTLs: if the genetic variant is located in a TF binding site and changes its binding affinity, the co-expression between the TF and its target gene will become weaker for one genotype, resulting in a co-eQTL. Additionally, indirect interactions can be captured with genes in high correlation with the TF, for example genes that are also regulated by the TF or are at least part of the same pathway. These correlated genes will show the same correlation behavior as the TF and can therefore also be observed in co-eQTL triplets together with the eQTL (Scenario 2 in Figure 5.1 c).

Other scenarios are also possible that result in co-eQTLs. For example, if a genetic variant is associated with the sub cell type composition in a cell type, it can be associated with changes in correlation structure for sub cell type specific genes. This can be prevented by sub-cell type specific analysis, requiring however accurate annotations and a sufficient number of cells from each sub-cell type to gain enough power. For this reason, we explored sub-cell type effects only for Monocytes, and ran analyses otherwise on cell type level.

We explored the identified co-eQTLs with comparison between different cell types and replication in a large bulk study. Additionally, we analyzed technical factors influencing the detection power, such as the number of samples and cells (Figure 5.1 d). At last, we combined several enrichment analyses to identify the upstream regulatory pathways and the associated direct regulators from the set of co-eQTLs per eQTL, in order to differentiate better between the different scenarios described before (Figure 5.1 c).

The methodology, results and figures presented of this project have previously been published as a preprint in Li et al. [3] and are currently under review in Genome Biology (status November 2022). All code, including a description on how the figures were generated, is published in the GitHub repository associated with the publication `https://github.com/sc-eQTLgen-consortium/co-expressionQTLs`.

5.2. Benchmarking co-expression pattern obtained from single cell data

5.2.1. Exploring different association metrics for single cell data

The first step, before running the co-eQTL mapping itself, is choosing an association measure for the gene pairs. Previous analyses used Spearman correlation [55, 83], which is easy to apply and interpret. Different benchmarking studies identified however alternative measures that might work better specifically for single cell data [179, 87]. For this reason, we compared Spearman correlation with Rho proportionality [180] and

GRNboost2 [181], two of the suggested methods.

For Rho proportionality [180], we saw high correlation with Spearman values for genes expressed in at least 5% of the cells (r=0.68) (Supplementary Figure A.19 a). However, for some gene pairs, where both genes were very lowly expressed, very high Rho propensity values were calculated in contrast to Spearman correlation values around 0 (Supplementary Figure A.19 b). The authors of Rho propensity warned that the log-transformation during the estimation of Rho values can cause problems for sparse data with many zero values, potentially falsifying the results [180]. Additionally, the calculation of Rho values was far more computational demanding, so we decided that the Spearman correlation was the better choice.

Next, we evaluated GRNBoost2 [181] and Spearman correlation, by estimating how robust the gene pair associations were between similar datasets, which measured the same cell type (done later more extensively, see section 5.2.2). For this, we compared the Spearman correlation values identified in monocytes from the Oelen v2 dataset with the correlation values identified in a FACS-sorted bulk dataset of classical monocytes from the BLUEPRINT consortium [43]. Then we repeated the same for GRNBoost2 edge values. The correlation between single cell and bulk results was drastically lower for GRNBoost2 (r=0.17) (Supplementary Figure A.20) than for Spearman correlation (r=0.34) (Supplementary Figure A.21), suggesting that Spearman is more robust across datasets.

Many other single cell specific association measures could not be applied, as they required pseudotemporal ordering of cells [87], which was not possible to reliably infer for our adult PBMC datasets. We tested different methods for pseudotime ordering, RNA velocity [182] and SCORPIUS [183], on an extended version of the Oelen v3 dataset, which contained two additional measurement points, 3 hours and 24 hours after stimulation with pathogens. The inferred ordering from both algorithms did not agree well and separated only the time points clearly (Supplementary Figure A.22). This aligns with other studies that stated that the estimation of RNA velocity is difficult with a PBMC dataset [184]. Overall, as we could not validate any results and there are no large trajectories to expect in the untreated adult cells, we decided to not follow-up further on this.

Furthermore, we tested if a reduction of the single cell sparsity improves the Spearman correlation results. To achieve this, we grouped neighboring cells to meta-cells using both the original *MetaCell* algorithm [185] and our own implementation based on Leiden clustering [186] (see section 5.7.6). In both cases, we calculated the average expression for each gene per meta-cell and then performed Spearman correlation on the meta-cells. As planned, the sparsity was clearly reduced in the meta-cells. However, this led not to a better concordance with the correlation pattern estimated in BLUEPRINT [43], which we used again for validating the robustness (Supplementary Figure A.23). For this reason, also the meta-cell grouping was not used for the following co-eQTL analysis.

5.2.2. Evaluating robustness of Spearman correlation across datasets

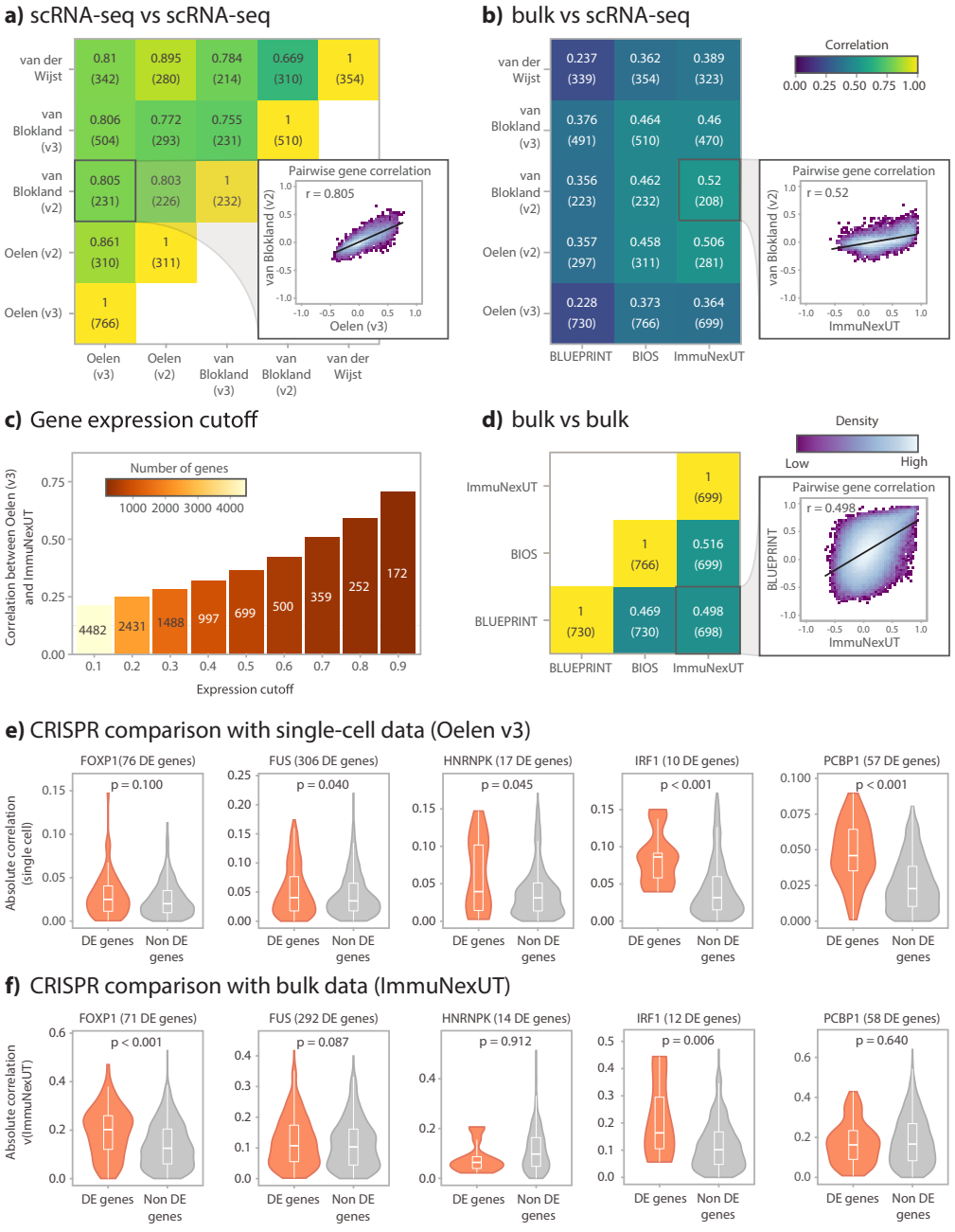
We extended the evaluation of the Spearman correlation values and explored how robust they were across different scRNA-seq and bulk RNA-seq datasets. In each comparison, we estimated the correlation values separately for each dataset and also separately per cell type, in case of the single cell datasets. Afterwards, we compared them across two datasets by calculating the Pearson correlation over the gene-pairwise correlation values, getting so one value for each comparison. A strict threshold was chosen thereby for the gene selection, with genes expressed in at least 50% of the cells in both datasets for each pairwise comparison. The influence of this expression cutoff is evaluated later in this section.

Across the five scRNA-seq datasets, we observed median correlation values between 0.86 for CD8+ T cells and 0.69 for monocytes (Figure 5.2 a, Supplementary Figure A.24). This high concordance was observable for all scRNA-seq datasets, for example for CD4+ T cells the correlations between datasets lay between 0.67 and 0.86.

For the comparison of single cell datasets with bulk datasets, we selected three large published bulk datasets, two with FACS sorted expression data from the BLUEPRINT consortium [43] and the ImmuNexUT consortium [44] and one whole blood dataset from the BIOS consortium [33]. The concordance across datasets was in general a bit lower than for the within single cell comparison. For the ImmuNexUT dataset, which covered all cell types annotated in our single cell datasets, the median correlation ranged between 0.570 for CD8+ T cells and 0.259 for DCs (Figure 5.2 b, Supplementary Figure A.21). Similar results were found for the comparison with the BLUEPRINT dataset, which contained only CD4+ T cells (median r= 0.356) and monocytes (median r=0.339). Likewise, the whole blood BIOS dataset got correlation estimates in a similar range (median r between 0.265 and 0.458 per cell types). Of note, the BIOS dataset did not contain cell type specific expression values, but had the largest sample size, which can improve the correlation estimates.

We identified different factors that influenced the concordance between bulk and single cell data: firstly, a stricter gene expression cutoff increased the concordance between Oelen v3 and ImmuNexUT for CD4+ T cells (r=0.71 for genes expressed in 90% of the cells) and vice versa, a less strict cutoff reduced it (r=0.21 for genes expressed in 10% of the cells) (Figure 5.2 c). As current single cell datasets tend to be sparser and noisier than bulk, a strict cutoff improves the correlation by selecting highly expressed genes that can be quantified more accurately. The drawback is, however, that only few genes could be analyzed here, at a cutoff of 90% only 172 genes remain. The cutoff of 50% that we applied in our previous analyses balances both extremes, so that part of the noisiest estimates was removed, but still several gene pairs could be explored.

Secondly, the cell type specific measurements did not match exactly between our single cell and bulk datasets, as the BIOS dataset is a whole blood dataset and BLUEPRINT and ImmuNexUT both captured only a subtype of CD4+ T cells and monocytes (naive CD4+ T cells and classical monocytes). Furthermore, due to the different processing steps, technical biases can lead to gene expression changes, for example during the


a) scRNA-seq vs scRNA-seq
van der Wijst
van Blokland (v3)
van Blokland (v2)
Oelen (v2)
Oelen (v3)
0.81 (342)
0.895 (280)
0.784 (214)
0.669 (310)
1 (354)
0.806 (504)
0.772 (293)
0.755 (231)
1 (510)
0.805 (231)
0.803 (226)
1 (232)
0.861 (310)
1 (311)
1 (766)
Pairwise gene correlation
r = 0.805
van Blokland (v2)
Oelen (v3)
b) bulk vs scRNA-seq
0.237 (339)
0.362 (354)
0.389 (323)
0.376 (491)
0.464 (510)
0.46 (470)
0.356 (223)
0.462 (232)
0.52 (208)
0.357 (297)
0.458 (311)
0.506 (281)
0.228 (730)
0.373 (766)
0.364 (699)
BLUEPRINT
BIOS
ImmuNexUT
Correlation
0.00 0.25 0.50 0.75 1.00
Pairwise gene correlation
r = 0.52
c) Gene expression cutoff
Number of genes
1000 2000 3000 4000
Correlation between Oelen (v3) and ImmuNexUT
4482
2431
1488
997
699
500
359
252
172
Expression cutoff
d) bulk vs bulk
Density
Low
High
1 (699)
1 (766)
0.516 (699)
1 (730)
0.469 (730)
0.498 (698)
Pairwise gene correlation
r = 0.498
e) CRISPR comparison with single-cell data (Oelen v3)
FOXP1(76 DE genes)
p = 0.100
FUS (306 DE genes)
p = 0.040
HNRNPK (17 DE genes)
p = 0.045
IRF1 (10 DE genes)
p < 0.001
PCBP1 (57 DE genes)
p < 0.001
Absolute correlation (single cell)
DE genes
Non DE genes
f) CRISPR comparison with bulk data (ImmuNexUT)
FOXP1 (71 DE genes)
p < 0.001
FUS (292 DE genes)
p = 0.087
HNRNPK (14 DE genes)
p = 0.912
IRF1 (12 DE genes)
p = 0.006
PCBP1 (58 DE genes)
p = 0.640
Absolute correlation v(ImmuNexUT)

FACS sorting.

Thirdly, as bulk expression counts are average values across many cells, the correlation values estimated from bulk can differ from correlation values obtained from the individual single cells. This divergence is called Simpson's paradox [187]. We aggregated our single cell values to pseudobulk to explore if this effect is visible in our data. Indeed, we identified several cases of highly expressed genes where the correlation was only visible in the pseudobulk data and not in the single cell data or the other way round (Supplementary Figure A.25). However, combing data points to pseudobulk can also reduce noise and lead to more accurate co-expression estimations, especially for lowly expressed genes. For this reason, identifying the Simpson's paradox for lowly expressed

Figure 5.2. *(preceding page)*: **Evaluation of correlation metrics**
a) Spearman correlation of different single cell datasets were compared with each other, always taking the CD4+ T cells and genes expressed in at least 50% of the cells in the corresponding datasets (number of tested genes in parentheses below Spearman correlation value). **b)** Comparison of the co-expression profiles between the single cell datasets and with the bulk RNA-seq datasets from BLUEPRINT, ImmuNexUT (both measuring FACS-sorted naive CD4+ T cells) and BIOS (whole blood). Again, we only assessed genes expressed in at least 50% of the cells for the single cell dataset (number of tested genes shown in parentheses below the Spearman correlation value). **c)** Relationship between the co-expression similarity between the ImmuNexUT naive CD4+ T cells and Oelen v3 dataset CD4+ T cells and increasing gene expression cutoffs (the ratio of cells with non-zero expression for a given gene). The number of genes tested are indicated by color scale and the numbers in the bar plot. **d)** Comparison of the co-expression profiles between the bulk RNA-seq datasets, taking the same gene subset as in **a,b** (number of tested genes in parentheses below the exact Spearman correlation value). **e)** Enrichment of correlated genes in scRNA-seq (Oelen v3 dataset) among associated genes identified by CRISPR knockout. For the enrichment, genes differentially expressed after knockout were identified and tested for enrichment. P-values in the plot show the significance level of the Wilcoxon rank-sum test. **f)** Enrichment of correlated genes in bulk RNA-seq (ImmuNexUT) among associated genes identified by CRISPR knockout. Figure and legend taken from [3].

genes is not always that clear, and we decided to not quantify the paradox overall.

As a last point, we observed that the gene-pair correlation estimates differed also between the bulk datasets (r=0.47 and 0.52 for CD4+ T cells and r=0.35 and 0.42 for monocytes) (Figure 5.2 d, Supplementary Figure A.26). This highlights that the bulk datasets are not the perfect ground truth to capture all gene relationships correctly. Putting it into perspective, the correlation between single cell and bulk datasets were in a similar range as the ones of bulk vs bulk datasets.

5.2.3. Evaluating Spearman correlation compared to associations from CRISPR knock-out data

Because the bulk datasets were not the perfect ground truth for the validation of the single cell correlation values, we additionally included associations from a published CRISPR-knockout scRNA-seq dataset in CD4+ T cells [188] in the comparison with the Spearman correlation values. In the CRISPR dataset, successfully perturbed cells were identified via single cell RNA barcodes using a tool called Mixscape [189] and afterwards differentially expressed genes between wild type cells and perturbed cells were analyzed. We restricted the analysis then to five knockout genes that had at least 10 DE genes.

In four of the five cases, we detected significantly higher correlation values between the DE genes and the knock-out gene compared to the non-DE genes and the knock-out gene using Wilcoxon rank sum test ($p<0.05$) and the CD4+ T cell correlation values from the Oelen v3 dataset (Figure 5.2 e). When comparing it with bulk correlation values from ImmuNexUT (naive CD4+ T cells) instead, only for two of the five knockout genes, the correlation distribution was significantly higher for the DE genes (Figure 5.2 f).

Additionally, we applied the same approach for comparing gene pairs listed in the STRING database [190] (i.e. who interact on the protein level) if their correlation values were higher than for non-interacting pairs. For both, single cell and bulk correlation values, the shift was significant based on Wilcoxon rank-sum test ($p < 0.05$) (Supplementary Figure A.27).

5.3. Exploration of cell type and donor specific co-expression

After we showed in the validation that the Spearman correlation values were robust across different datasets and matched associations identified in CRISPR data and in the STRING database, we analyzed the biological differences of the correlation values between cell types and donors within one cell type. We followed the same approach as for the comparison between the datasets: the gene-pair correlation values between two cell types were compared by calculating Pearson correlation across them. The identified pattern matched the biological relationships, with high correlation across the different lymphoid cell types (B, T and NK cells; $r > 0.73$), and lower correlations between the lymphoid and myeloid cell types (monocytes, DCs; $r < 0.45$) (Figure 5.3 a). The relatively

low correlation between monocytes and DCs might be caused by the low cell type frequency of the DCs, leading to less accurate co-expression estimates.

In general, the different scRNA-seq datasets showed more similar correlation pattern (median r=0.80, Figure 5.2 a, Supplementary Figure A.24) than the different cell types within the same dataset (e.g. for Oelen v3 dataset median r=0.64, Figure 5.3 a). This further assured us that the correlation values capture meaningful biological characteristics, such as cell type differences.

The overall correlation distribution differed slightly between cell types (Figure 5.3 b), but in general only few gene pairs had correlation values above 0.1 (median 12.4%). The DCs represented an outlier here with 32.3% gene pairs with r > 0.1.

As introduced in the beginning, multiple measurement points per donor in scRNA-seq data allow the calculation of donor-specific correlation values. Within one cell type, the concordance of correlations from different donors was quite high for the more frequent cell types (e.g. median r=0.56 for CD4+ T cells), but far lower for less frequent cell types (e.g. median r=0.06 for B cells) (Figure 5.3 c).

When we subsampled the cell types, we identified a clear association between the number of cells and the correlation between individuals (Figure 5.3 d). Hence, most of the differences between cell types were introduced by different number of cells, leading to more robust estimates for the more frequent cell types and so better concordance across donors. Some differences remained: when subsampling all cell types to the same level, the NK cells showed lower concordance than the CD4+ and CD8+ T cells. A logarithmic model that we trained for the four most frequent cell types (adjusted R^2 values between 0.86 and 0.98) predicted median correlation levels > 0.80 for 1,000 T cells or monocytes per donor and levels of 0.65 for 1,000 NK cells per donor (Supplementary Figure A.28). This highlights the importance of a sufficient number of cells from each donor to estimate cell type specific correlation values.

5.4. Approach for systematic identification of co-eQTLs

Following all these evaluations for the single cell Spearman correlation, we confidently applied the donor-specific correlation values in our co-eQTL analysis. To identify these genetic variants that change the correlation of a gene pair, we developed our own strategy that deals with the sparsity of the single cell data and the large multiple testing burden due to the large search space.

Previous co-eQTL analyses focused on a small set of preselected triplets (SNP-gene-gene) to avoid a large multiple testing burden [55, 83]. The goal of our analysis here was, however, a large-scale co-eQTL analysis to get a more comprehensive picture of the occurrence of co-eQTLs in different cell types. We estimated that our power would drop to 1.4% for a co-eQTL with a heritability of 10% if we test all genes expressed in monocytes and one SNP per pair (in total $1.98 * 10^8$ tests). For this reason, we pre-selected triplets for testing that we assumed are generally more likely to be co-eQTLs.

In the first step of our filtering strategy, we identified between 51 cis eQTLs for B

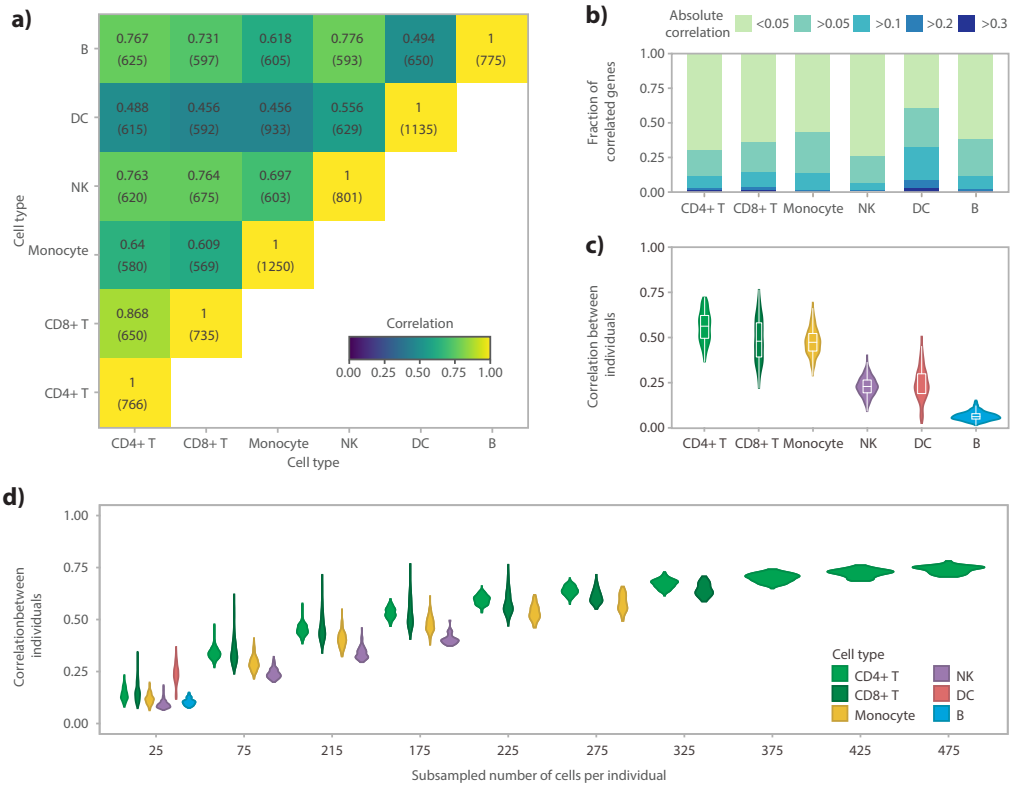


Figure 5.3.: **Comparison of correlation across cell types and donors**
Each analysis was performed in the Oelen v3 dataset for all genes expressed in at least 50% of the cells of the respective cell type. **a)** Comparing co-expression patterns across cell types within the Oelen v3 dataset for genes expressed in 50% of the cells for both cell types in each pair-wise comparison. The number of tested genes is shown in parenthesis below the Spearman correlation value. **b)** Correlation distribution within each cell type. **c)** Correlation between different individuals within each cell type showing the distribution of all pair-wise comparisons between individuals. **d)** Relationship between the number of cells per individual and cell type and correlation between individuals separately for each cell type. In each subsampling step, we assessed all individuals who have at least this number of cells and subsampled to exactly this number (this leads to removal of some individuals for higher number of cells and thus, a direct comparison with the correlation values in **c)** is not possible). Figure and legend taken from [3].

Cell type	Number of eQTLs
CD4+ T cells	917
CD8+ T cells	527
Monocytes	625
NK cells	376
DCs	145
B cells	51

Table 5.1.: **Significant eQTLs**
Number of significant eQTLs (FDR<0.05) identified via meta analysis in each cell type. Table adapted from [3].

cells and 917 cis eQTLs for CD4+ T cells (Table 5.1) via a cell type specific meta analysis (FDR<0.05; more in Methods section 2.2.3). We used here four of the five scRNA-seq datasets. The van Blokland v3 dataset was left out of this and all following analysis, as the small sample size led to very few variants that passed the minor allele frequency (MAF) cutoff of 10%. The motivation for this first filtering step was that variants that are co-eQTLs are expected to influence the expression of one of the genes also directly as an eQTL.

Selecting only variants that are eQTLs together with the associated gene as the eGene is not sufficient alone to reduce the testing burden enough. For example, it would still be over 12 million tests for CD4+ T cells. For this reason, we filtered additionally the second tested genes, the co-eGenes, and tested only genes significantly correlated (nominal $p < 0.05$) with the eGene in at least 10% of the individuals. The 10% cutoff was chosen based on the MAF filter, so that at least the individuals of one genotype group would have significant correlation values. The rationale behind was that we are interested in changes in correlation for co-eQTLs, but these changes are only meaningful if the correlation is significant in at least one group. In general, the filtering criteria were selected so that the approach balance both extremes, reducing the multiple testing burden but nevertheless still testing many potentially interesting gene pairs. Therefore, we chose a less strict filtering cutoff as for the first part, where we selected genes expressed in at least 50% of the cells (Figure 5.2, 5.3).

The co-eQTL mapping was performed again as a cell type specific meta analysis of our four selected single cell datasets in combination with a customized permutation approach per eQTL to deal with the correlation structure between the tests. We identified between 500 co-eQTLs for CD4+ T cells, containing 30 unique co-eQTL SNPs, and 35 co-eQTLs for B cells, all caused by the same co-eQTL SNP (Table 5.2).

Across all cell types, this led to a total of 72 unique co-eQTL SNPs associated with 946 unique gene pairs. The overlap between the cell types was rather small (Supplementary Figure A.29). However, reliable statements about the cell type specificity of co-eQTLs can not be inferred from this, as also the tested set of triplets differed between cell types due to our strict filtering strategy (Figure 5.4 a). The power issue for less frequent cell

Cell type	Number of tests	Number of co-eQTLs	Number of unique co-eQTL SNPs
CD4+ T cells	179,841	500	30
CD8+ T cells	73,017	420	22
Monocytes	304,707	281	24
NK cells	25,998	123	10
DCs	41,655	58	9
B cells	2,936	35	1

Table 5.2.: **Significant co-eQTLs**
Number of tested triplets (SNP-eGene-co-eGene) and number of significant co-eQTLs (FDR<0.05) identified via meta analysis in each cell type. Table adapted from [3].

types is again observable, with fewer eQTLs and co-eQTLs for the less frequent cell types.

When limiting the pairwise comparisons between cell types to triplets tested in both cell types, relatively high concordance of co-eQTL effect sizes is visible across cell types, with a median r_b value [191] of 0.85 (Figure 5.4 a, see section 5.7.16). The concordance was especially high between CD4+ and CD8+ T cells with values of 0.97 and 0.99.

The effect sizes across the different datasets of our meta analyses matched very well, supporting the robustness of our results (Supplementary Figure A.30). Moreover, the co-eQTLs replicated well in an independent dataset. We chose the large bulk dataset from the BIOS consortium [33] for the replication (n=2,491, removing samples also part of our single cell analysis), as the availability of other large-scale single cell datasets is still very limited. To get co-eQTLs from bulk, a linear regression model with an interaction term was applied on the BIOS dataset (more in Methods section 2.2.4). We found high concordance of effect sizes between the datasets with r_b values between 0.30 and 0.61 dependent on the cell type (Figure 5.4 b) The replication was especially high for the more frequent cell types, the highest for CD4+ T cells with an r_b value of 0.61. For the least frequent cell type, the B cells, the calculation of the r_b values was not possible, as only one co-eQTL SNP was identified here. This provided not enough independent measurement points for the calculation.

After we have proven that our chosen strategy with strict filtering provided us a set of robust co-eQTLs that replicated well in BIOS, we explored the effect of different technical factors. First, we skipped the last step of our filtering and tested each SNP-eGene pair against all other expressed genes. The higher number of tests led to a higher absolute number of co-eQTLs for the more frequent cell types, but to fewer co-eQTLs for B cells and DCs (Table 5.3). The replication rate in the BIOS dataset was drastically reduced in all cell types (Supplementary Figure A.31), highlighting that our original approach including the filtering step produced a more robust set of co-eQTLs.

We observed that the distribution of correlation values differed between co-eQTLs and

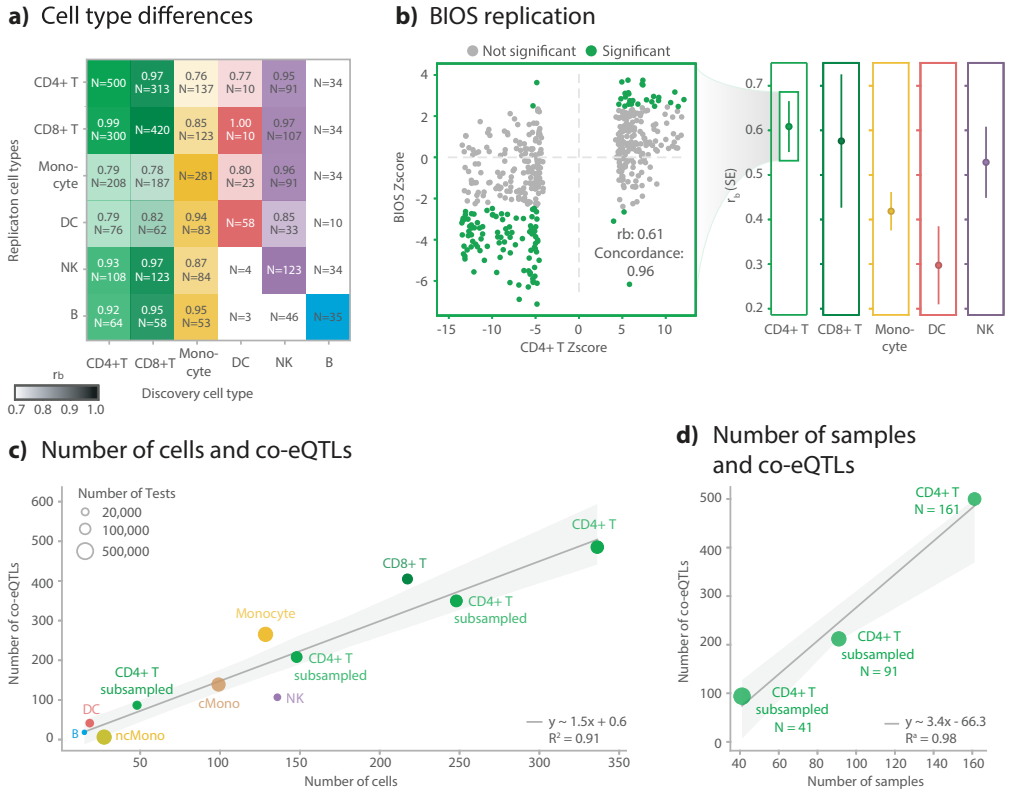


Figure 5.4.: **General characteristics of identified co-eQTLs a)** Replication of discovered co-eQTLs across the major cell types. Correlation of the effect sizes in replications among different cell types, measured by rb value. Text inside each block indicates the rb value and number of replicated co-eQTLs. Color intensity indicates rb value. **b)** Replication in BIOS dataset for different cell types, indicated by the rb values. Scatter plot shows the detailed Z-score comparison between the co-eQTL meta analysis and the Z-score from the BIOS replication for CD4+ T cells. **c)** Number of significant co-eQTLs for varying cell numbers. Dot color indicates the cell type, as indicated in the text next to each dot. "cMono" means classical monocytes. "ncMono" means non-classical monocytes. "CD4+ T Subsampled cells" means that this analysis was done for CD4+ T cells, but for every individual we randomly downsampled cells to the desired cell number as indicated in the x-axis. **d)** Number of significant co-eQTLs for varying sample numbers. "CD4+ T Subsampled Individuals" indicates that this analysis was done for CD4+ T cells, but we randomly subsampled for the individuals. Figure and legend taken from [3].

Cell type	Number of tests	Number of co-eQTLs	Number of unique co-eQTL SNPs
CD4+ T cells	12,137,281	15,433	249
CD8+ T cells	6,390,963	2,561	123
Monocytes	7,323,138	425	63
NK cells	4,261,756	960	79
DCs	1,272,690	8	1
B cells	390,150	20	1

Table 5.3.: **Significant co-eQTLs from the unfiltered approach**
Number of tested triplets (SNP-eGene-co-eGene) and number of significant co-eQTLs (FDR<0.05) identified via meta analysis in each cell type, when applying the unfiltered approach instead of the filtered. Table adapted from [3].

non-significant triplets, especially for the co-eQTLs identified with our original strategy including the co-eGene filtering step (Supplementary Figure A.32). The correlation mean and variance of these co-eQTLs were higher than for the non-significant triplets, as were the non-zero rates of the significant eGenes and co-eGenes. This further supports that we identified true biological associations here. The differences could be utilized in alternative filtering strategies to preselect triplets for co-eQTL mapping.

Furthermore, we evaluated potential confounding by sub-cell type composition in our co-eQTL set, exemplarily for monocyte co-eQTLs. The monocytes can be classified into two large subgroups, classical and non-classical monocytes. The assumption we tested was the following: genetic variants that change the distribution between classical and non-classical monocytes within an individual could potentially be identified as co-eQTL associated with sub-cell type specific genes, even without a direct relationship between the SNP and the genes. These sub-cell type specific co-eQTLs would only be detectable in the complete set of monocytes, not in the sub cell types separately. However, we found no general strong confounding of sub-cell type composition, as co-eQTL effect sizes were highly concordant between sub-cell types and all monocytes ($r_b \geq 0.9$) (Supplementary Figure A.33). Nevertheless, some individual co-eQTLs might still be explainable due to sub-cell type confounding.

At last, we explored the effects of different experimental parameters on the number of identified co-eQTLs by subsampling the number of cells and the number of samples (Figure 5.4 c,d) (methods in section 5.7.18). We observed that a higher number of cells led not only to more robust co-expression estimates, as shown before (section 5.3), but also to a higher number of co-eQTLs (Figure 5.4 c). The impact of increasing the number of samples was even stronger (Figure 5.4 d), highlighting the benefit of planned consortia with very large sample sizes such as sc-eQTLgen [88] to potentially identify far more co-eQTLs.

5.5. Interpretation of co-eQTLs

5.5.1. General results of enrichment analyses

The biological interpretation of the co-eQTLs can pinpoint upstream regulatory processes and direct regulators, which are disrupted by genetic variants. This gives very valuable insights into genetic regulation, especially also for disease variants. However, as discussed in the beginning (Figure 5.1), the interpretation is complicated by the mixture of direct regulators and indirectly associated co-eGenes in the co-eQTL set. As these indirectly associated co-eGenes are supposed to be strongly correlated with the unknown direct regulator, we made use of all co-eGenes and tried to identify the common upstream pathways via a combination of different enrichment analyses. We explored common gene ontology (GO) terms, TF binding sites and GWAS annotations separately for all co-eGenes associated with the same SNP-eGene pair (details in section 5.7.19).

We focused on the 25% of SNP-eGene pairs (in total 19 pairs) with at least five co-eGenes in at least one cell type. Most of them (18 of 19) were significantly enriched for at least one GO term, showing potential common biological functions or pathways for the co-eGenes. For seven SNP-eGene pairs, we additionally identified enrichment of TF binding sites in the promoter region of the co-eGenes using ChIP-seq annotations from the ReMap database [192]. These TFs represent likely common regulators of the shared processes. In four of the cases, part of the enriched TFs overlapped the co-eQTL SNP directly (or a SNP in high LD), strengthening the hypothesis that these TFs could be the direct regulators affected by the SNP.

Furthermore, we identified that many of the co-eQTL SNPS (41 out of 72) were associated with a GWAS loci (either directly or over a SNP in high LD). For two of these co-eQTL SNPs, the associated co-eGenes were enriched for the same GWAS traits, after annotating co-eGenes to GWAS traits via *MAGMA* [193]. These GWAS traits, mostly blood cell counts and immune-mediated diseases, support again a potential shared biological mechanism of the co-eGenes and a connection with the co-eQTL SNP itself.

In the following, we focused on different example SNP-eGene pairs where we found convincing evidence for a better interpretation of affected upstream processes of the SNP over GO and TF enrichment and downstream consequences based on GWAS annotations.

5.5.2. Interpretation of co-eQTLs associated with rs1131017–RPS26

The SNP-eGene pair rs1131017–*RPS26* was associated with the highest number of co-eGenes in total and was the only pair with significant associations in all six tested cell types. While co-eQTLs for rs1131017–*RPS26* were identified before in CD4+ T cells [55] and monocytes [83], our study, which combines a new approach with a large sample size, enabled a more in depth study of rs1131017–*RPS26* in all cell types. The rs1131017–*RPS26* eQTL is especially interesting, as rs1131017 was associated with several

autoimmune diseases before, including Type 1 diabetes [194]. If and how the eQTL gene *RPS26*, which is a ribosomal gene, is involved in this relationship remained unknown in previous studies [195].

Interestingly, the direction of effect differed between the cell types. Most co-eQTLs showed a positive direction of effect in monocytes, NK cells and B cells, i.e. the correlation increased for the less frequent genotype, concordant with the eQTL direction (Figure 5.5 a,b,c). In contrast, co-eQTLs with positive and negative direction of effects were nearly balanced in T cells (46% negative direction of effect in CD4+ T cells and 43% in CD8+ T cells) (Figure 5.5 a,c,d).

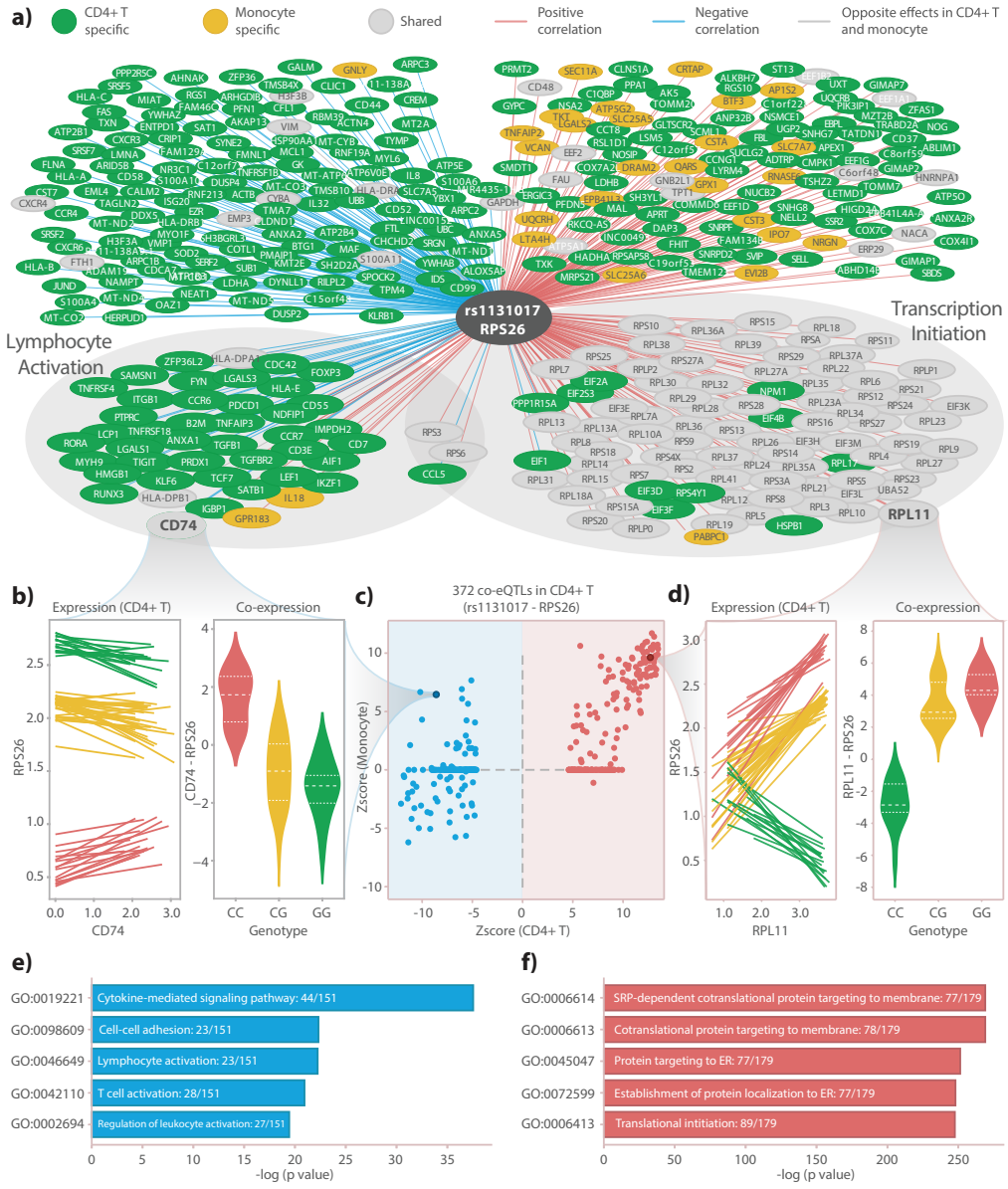
This was not only driven by different detection power across the cell types, as the replication of positively associated co-eQTLs was high across cell type, but not the negatively associated co-eQTLs (Figure 5.5 c, Supplementary Figure A.34). When running the GO enrichment analyses separately for positive and negative co-eQTLs, different functions turned up (Figure 5.5 e,f): positive co-eGenes were enriched for functions associated with translation (Figure 5.5 f), matching the fact that many positive co-eGenes were ribosomal genes as *RPS26* itself. Contrary, negative co-eGenes were enriched for immune response and T cell activation, further strengthening their specificity for T cells.

Six TFs were found in the TF enrichment analysis among co-eGenes that showed also binding directly at the genomic region of the co-eQTL SNP rs1131017 and that were co-eGenes themselves. From this set of potential direct regulators for the eQTL, five are involved in lymphocyte activity. Furthermore, *MAF* and *CD74*, two of them, were specifically enriched in the subset of negative co-eGenes. These two are especially interesting candidates for follow-up analyses.

Enrichment for different blood cell count traits were found in multiple cell types via our GWAS enrichment analysis. Additionally, only in T cells, there was significant enrichment of multiple immune-related diseases, among them rheumatoid arthritis (RA), Crohn's disease (CD), multiple sclerosis (MS) and hay fever. This matched the GWAS annotations of rs1131017. Taken together, the co-eQTLs and the different enrichment analyses indicate a role of *RPS26* in lymphocyte activation, which could explain how the downstream effects of rs1131017 are related to autoimmune diseases.

5.5.3. Other co-eQTL results

Also for other SNP-eGene pairs, the co-eQTLs provided additionally valuable insights into the genetic regulation. One of them is the SNP rs7806458 associated with the eGenes *TMEM176A* in monocytes (11 co-eGenes) and *TMEM176B* in monocytes (6 co-eGenes) and DCs (1 co-eGene). The SNP was associated with MS in previous studies [196]. Matching this, the GWAS enrichment of co-eGenes associated with rs7806458-*TMEM176A* showed an enrichment for MS. Additionally, the co-eGenes of rs7806458-*TMEM176B* in monocytes were enriched for the GO term complement component C3b binding. This is noteworthy, as MS was related to disturbances in blood coagulation before [197], with which complement component C3b binding is closely connected. Even though the TF enrichment analysis brought up no potential direct regulators for the


a)
CD4+ T specific
Monocyte specific
Shared
Positive correlation
Negative correlation
Opposite effects in CD4+ T and monocyte
rs1131017
RPS26
Lymphocyte Activation
Transcription Initiation
CD74
RPL11
b)
Expression (CD4+ T)
Co-expression
RPS26
CD74
CD74 - RPS26
CC
CG
GG
Genotype
c)
372 co-eQTLs in CD4+ T
(rs1131017 - RPS26)
Zscore (Monocyte)
Zscore (CD4+ T)
d)
Expression (CD4+ T)
Co-expression
RPS26
RPL11
RPL11 - RPS26
CC
CG
GG
Genotype
e)
GO:0019221 Cytokine-mediated signaling pathway: 44/151
GO:0098609 Cell-cell adhesion: 23/151
GO:0046649 Lymphocyte activation: 23/151
GO:0042110 T cell activation: 28/151
GO:0002694 Regulation of leukocyte activation: 27/151
-log (p value)
f)
GO:0006614 SRP-dependent cotranslational protein targeting to membrane: 77/179
GO:0006613 Cotranslational protein targeting to membrane: 78/179
GO:0045047 Protein targeting to ER: 77/179
GO:0072599 Establishment of protein localization to ER: 77/179
GO:0006413 Translational intitiation: 89/179
-log (p value)

process, these co-eGenes provided further information about the mechanistic connection between rs7806458 and MS.

Another interesting example is the SNP-eGene rs393727-*RNASET2*. We did not run an enrichment analysis for it, as it contained only four co-eGenes and we set a threshold of five for all enrichment analyses in general. Still, we found several connections with immune-mediated diseases for the co-eQTL set: based on our GWAS analysis, the SNP is connected with CD and inflammatory bowel diseases (IBD) over GWAS SNPs in high LD with it. The eGene is associated with IBD [198], one co-eGene *ITGB1* with CD [199] and another co-eGene *CRIP1* with gut immunity [200]. Especially interesting for this example is that all four co-eGenes are also negatively associated co-eGenes for rs11311017–*RPS26*. This might reveal common upstream regulatory pathways for both eQTLs and in general for the different immune diseases.

The interpretation of co-eQTLs is not always that straightforward, as shows the example of several co-eQTLs in the HLA locus. One of them is the SNP-eGene pair rs9271520–*HLA-DQA2*, which could be identified in T cells, monocytes and DCs (7-17 co-eGenes dependent on the cell type). The SNP was associated with several immune diseases that were also enriched among the co-eGenes. The exact mapping of causal relationships in this example is however challenging due to the LD structure in the HLA region. Proofing this, a removal of HLA genes from the GWAS enrichment analysis led to a loss of the signal.

Further SNP-eGene examples and their interpretation can be found in the publication [3] and its supplementary text.

Figure 5.5. *(preceding page)*: **Annotation of co-eQTLs**
a) Union network constructed with co-eQTLs found in CD4+ T cells or monocytes that are associated with the SNP–eGene rs1131017–*RPS26*. The two circled clusters contain co-eGenes that are in those corresponding GO terms. **b)** Example of one co-eQTL: rs1131017–*RPS26-CD74*. Left plot indicates the co-expression patterns from all individuals in the Oelen v3 dataset. Each regression line was fitted with expression data from one individual. Right plot indicates the co-expression values from the three genotype groups. **c)** Comparison between z-scores from monocytes and z-scores from CD4+ T cells, Red dots indicate positive co-eQTLs from CD4+ T cells and blue dots negative co-eQTLs. **d)** Example of one co-eQTL: rs1131017–*RPS26-RPL11* with the same layout as **(b)**. **e)** GO term enrichment results for the co-eGenes in negative co-eQTLs from CD4+ T cells. **f)** GO term enrichment results for the co-eGenes in positive co-eQTLs from CD4+ T cells. Figure and legend taken from [3].

5.6. Project summary and outlook

In this project, we explored the value of co-eQTL analysis, a new system genetics approach becoming now possible with growing scRNA-seq cohorts. Importantly, we evaluated the best way to map co-eQTLs. This included that we benchmarked the use of Spearman correlation as an association measure, we showed the effect of filtering tested SNP-gene-gene triplets to identify more robust co-eQTLs and highlighted the influence of number of cells and sample size on the detection power. With our selected strategy, we were able to identify 72 unique SNP co-eQTLs associated with 946 unique gene pairs in a cell type specific analysis for the six major cell types. At last, we showed strategies to interpret these co-eQTLs and identified with this for example a connection of the eQTL rs1131017-*RPS26* with T cell activation that can help to explain its association with immune-related diseases.

Rapid developments in the field of scRNA-seq are likely to improve co-eQTL results in the future. We hope that our experimental parameter and methods evaluation will support future studies, and we expect far more co-eQTLs from larger sample sizes. This will aid the identification of cell type specific co-eQTLs and the mechanistic interpretation of co-eQTL, which are both currently limited by the power of our study. Newer scRNA-seq technologies are likely to have a better capture efficiency and will extend the number of genes that can be considered, which is currently still restricted because of the sparsity of the data. Additionally, multi-omics technologies have the potential to advance association analysis of gene pairs and to allow the development of new methods on top of Spearman correlation. They enable an extended study of genetic regulation, which can not be completely captured at the transcript level.

5.7. Materials and additional methods

5.7.1. Single cell datasets

Three different studies were combined for the co-eQTL mapping and the preceding evaluation analyses to increase the total sample size and so the detection power. All three were published before in other studies (or are currently under review) and named after the first author: the Oelen dataset [83] with 104 healthy donors in total, the van der Wijst dataset [55] with 45 healthy donors and the van Blokland dataset [178] with 38 cardiac patients. Further information can be found in the respective publications, including technical specifications of the experiment and quality control. We used the already preprocessed datasets from each publication, but selected only one time point for the Oelen dataset (untreated) and the van Blokland dataset (6-8 weeks after hospital admission of the patients), choosing the time points that are expected to match the best.

For all downstream processes, count matrices normalized with *sctranform* are used [201]. Cell type annotations were obtained from the original publication in case of the Oelen dataset and inferred with *Azimuth* for the van Blokland and van der Wijst dataset [202]. Sub-cell type annotations were taken from *Azimuth* for all datasets.

5.7.2. Bulk datasets for evaluations

Three large published bulk dataset were taken for evaluation of the single cell association results, two of them containing cell type specific expression measurements from FACS sorting and one whole bulk measurements. The dataset from the BLUEPRINT consortium comprises FACS-sorted data from naive CD4+ T cells and classical monocytes from 197 individuals [43]. We regressed out the first principal component for the monocyte data to correct for unknown covariates. The second FACS-sorted dataset from the ImmuNexUT consortium is larger with 28 different cell types from 337 donors [44]. We followed the preprocessing as described in the publication. Briefly, we filtered lowly expressed genes, normalized using TMM with edgeR and corrected for batch effects via combat and removed outlier samples. The whole blood dataset from the BIOS consortium is the largest with 3,198 donors [33] and was corrected for 20 RNA alignment metrics before the analysis.

5.7.3. Rho proportionality

We estimated Rho proportionality values in Oelen v3 dataset monocytes for all genes expressed in at 5% of the cells (8,634 genes) using the *propr* R package [180] and compared the Rho values to Spearman correlation values by calculating one Pearson value across the gene pairwise association values. Additionally, we analyzed the concordance between both measures specifically for lowly expressed genes, which are especially critical for single cell data. To reduce the very high computational demand for Rho estimation here, we randomly selected 50 genes expressed in 0-5% of the cells and 50 genes expressed in at least 90% of the cells. We estimated Rho proportionality and Spearman correlation within the lowly and highly expressed genes and between each the groups.

5.7.4. GRNBoost2

GRNBoost2 [181] was applied on the Oelen v2 dataset monocytes for genes expressed in 50% of the cells and on the bulk dataset from BLUEPRINT classical monocytes. Edge weights between both were correlated using Pearson correlation. The result was compared to the Pearson correlation value when using Spearman correlation instead of the GRNBoost2 edge weights.

5.7.5. Temporal ordering of cells

Two different approaches for pseudotemporal ordering were applied on the subset of classical monocytes from the Oelen v3 dataset. The extended dataset with untreated and stimulated cells from the Candida stimulation was taken for this specific analysis to increase the differences between the cells in the dataset and obtain a ground truth of known time points.

The first method we tested is RNA velocity as implemented in *scVelo* [182] with dynamical mode and the 2,000 highest variable genes. The required input matrices with spliced and unspliced counts were produced with *velocyto* [203]. The second method was SCORPIUS, where we followed the default workflow as described in the manuscript [183]. Both temporal orderings were compared afterwards with each other.

5.7.6. Grouping cells to meta-cells

We tried to reduce the sparsity in our single cell data by grouping similar cells to metacells for the Oelen v3 dataset monocytes. First, we tested the original algorithm *MetaCell* [185], running it separately for each sample, but including also the stimulated conditions (in contrast to other analyses) to obtain larger heterogeneity for the clustering. As planned, the meta-cells divided the cells from different conditions nicely, and we could clearly assign the metacell condition dependent on the majority condition of the grouped cells. Average expression values across all cells in each metacells were taken as the expression values of the metacells.

We calculated the Spearman correlation across the untreated meta-cells, grouping the gene sets thereby in different expression bins (expressed in 20%-40% of the meta-cells, in 40%-60%, in 60%-80% and in 80%-100%). These values were compared with Spearman correlation values from the BLUEPRINT dataset classical monocytes. We repeated the same analysis using the single cell dataset directly instead of the meta-cells and explored which version had the higher concordance with BLUEPRINT.

As changing the granularity of the metacells, i.e. how many cells are grouped to one metacell, was limited in the original algorithm, we tested additionally our own implementation of a metacell algorithm using Louvain clustering of cells [186] with different resolution parameters.

5.7.7. Validation of Spearman correlation via comparison across datasets

Single cell correlation values were obtained separately for each of the five single cell dataset (split into study and 10X chemistry version) and per cell type, but combining all individuals. For the comparison across datasets, only gene expressed in at least 50% of the cells in both datasets were taken for each pairwise comparison. A summary value for each comparison was generated by calculating the Pearson correlation over all gene pairwise Spearman correlation values.

For the bulk correlation values, the same approach was chosen. The same set of genes was taken for a fair comparison: genes were required to be expressed both in the bulk dataset and in at least 50% of the cells in the compared single cell dataset.

Other thresholds from 10% to 90% were evaluated in the comparison of ImmuNexUT and Oelen v3 dataset and the comparison of BLUEPRINT and Oelen v3 dataset, both times for CD4+ T cells. The 50% threshold represented a good trade-off to capture a certain number of genes and still maintain some correlation between single cell and bulk results. For this reason, 50% was applied for all other evaluations.

5.7.8. Investigating the occurrence of the Simpson's paradox

We evaluated if we can find examples for Simpson's paradox with our single cell datasets. For this, we calculated the Spearman correlation once separately for each individual based on the single cells and once we first build a pseudobulk matrix, by averaging the counts per individual, and then calculated the Spearman correlation across donor. Then second approach mimics how a bulk version of the dataset would look like. We applied this strategy for genes expressed in at least 50% of the cells in monocytes from the Oelen v3 dataset. We explored the gene pairs with the largest difference between the pseudobulk correlation and the single cell correlation values.

5.7.9. Validation of Spearman correlation values using a CRISPR dataset

To define gene pair association from the CRISPR knockout dataset [188], successfully perturbed cells were identified using *Mixscape* [189]. We continued with five knockout genes that were expressed in at least 50% of the cells from the reference dataset (Oelen v3 dataset, CD4+ T cells) and had a sufficient number of perturbed cells. We defined genes as associated with the knockout gene if they were in the DE set of perturbed vs control cells (identified by *Mixscape*, FDR < 0.05), but not in the DE set of unperturbed vs control cells (again *Mixscape*, FDR < 0.05). We applied a one-sided Wilcoxon rank-sum test to identify if the correlation of these DE genes with the knock-out gene was higher than for other expressed genes. The same was repeated with naive CD4+ T cells from ImmuNexUT as a bulk reference set.

5.7.10. Validation of Spearman correlation values using the STRING database

Following the same logic as for the CRISPR dataset, we used a one-sided Wilcoxon rank-sum test to identify differences between Spearman correlation from gene pairs in the STRING database [190], i.e. pairs where the corresponding proteins interact, and other expressed gene pairs. We tested correlation estimates from single cell (Oelen v3 dataset, CD4+ T cells) and bulk (ImmuNexUT, naive CD4+ T cells). We chose the version 11 of the STRING database, downloading a preprocessed version of it curated by a benchmarking study [87].

5.7.11. Comparing Spearman correlation values across cell types

We followed the same strategy to compare the different cell types as we did to compare the different datasets for the same cell type. We calculated gene pairwise Spearman correlation values separately for each cell type, and then one Pearson correlation value across all gene pairwise correlations for the comparison in the end. We focused again on genes expressed in 50% of the cells for both cell types and chose the two larger single cell datasets for the evaluation, the Oelen v2 and v3 datasets. Furthermore, we compared the absolute distribution of correlation coefficients across cell types.

5.7.12. Comparing Spearman correlation values between donors

The approach to compared donors was the same as described in the last subsection 5.7.11 for the comparison of cell types, but splitting the dataset additionally per donor, i.e. calculating Spearman correlation values per donor and cell type. Every donor was then compared with every other donor within each cell type, which results in one correlation distribution over all donor comparisons per cell type.

We subsampled the number of cells per donor several times to evaluate what drives the differences between cell types. Donors that had less than the chosen number of cells were omitted from the comparison in each step. We started the subsampling at 25 cells per donor, increased the number in steps of 25 cells, and stopped when 75% of the donors had fewer cells than the subsampling number. The trend could be approximated by a logarithmic curve $mean(r_{individuals}) \sim log(n_{cells})$ for the four most frequent cell type (CD4+ and CD8+ T cells, monocytes and NK cells).

5.7.13. Power calculation

We estimated the power to detect a co-eQTL with a heritability of 10% (value selected based on previous study [83]) in our combined cohort with 173 samples based on the F-test, as implemented in *scPower* [1]. The significance threshold of 0.05 was corrected for multiple testing via the Bonferroni approach (see Methods chapter 2.4). The required number of tests was approximated assuming that we test one SNP per gene pair, but all combination of genes that are expressed in at least one cell in monocytes Oelen v3 dataset. Setting a higher expression threshold increases the detection power, in contrast, testing more SNPs per gene pair would decrease it.

5.7.14. eQTL mapping

For the cell type specific single cell eQTL mapping, we tested all SNP-gene pairs that were significant eQTLs in a very large bulk whole blood study [33], taking only the top SNP for each gene. This way, we limited the search space and increased the power. We performed the analysis separately for each cell type and for all genes expressed in this cell type. The eQTLPipeline v1.4.9 [204] was used, which combines results from the different cohorts using a fixed effect meta analysis (see Methods chapter 2.2.3), with a cis-window of 100 kb, 10 permutations for FDR calculation (strategy described in [33]) and a MAF of 0.1. Due to the small sample size, we omitted the van Blokland v3 dataset in this and all following analyses, as very few variants were above the MAF filter. This left us with four single cell datasets for the meta analysis (Oelen v2 and v3 dataset, van der Wijst dataset, van Blokland dataset).

5.7.15. Co-eQTL mapping

We explored two different strategies to select gene pairs to test in the co-eQTL mapping, which was always done separately for each cell type. In both cases, we restricted the

test set to SNP-eGenes that were significant in the corresponding eQTL analysis for the cell type. In the first approach, we tested these SNP-eGene pairs against all other genes that had significant Spearman correlation in at least 10% of the individuals (nominal p-value < 0.05). This additional filtering step was skipped for comparison later, taking instead all other expressed genes without any correlation requirements.

During the co-eQTL mapping, we decided to keep missing correlation values as missing instead of imputing them with 0. Our reasoning was that due to the sparsity of the single cell data, it is not clear if they represent really uncorrelated genes or are caused by lowly expressed genes not captured accurately.

We performed a meta analysis across the four single cell datasets (same strategy as for the eQTL meta analysis) followed by a sophisticated multiple testing strategy that considers the correlation structure among the tested gene pairs. For this, the permutation-based strategy from *fastQTL* [109] was adapted as follows: for each SNP-eGene pair, the SNP-eGene-co-eGene triplets were permuted between samples and the lowest p-value over all co-eGenes selected per SNP-eGene. This permutation was repeated 100 times, resulting in 100 p-values per SNP-eGene from which a beta distribution $Beta(n, k)$ is fitted. This is used as the extreme tail of the null distribution to get empirical p-values p_e for the nominal p-values p_n of the real (non-permutated) co-eQTL tests as $p_e = F_{Beta}(p_n)$ (see original algorithm [109] and Methods chapter 2.4 for more details).

The smallest p-value for each SNP-eGene pair over all co-eGenes was taken, transformed into an empirical p-value and Benjamini-Hochberg correction was performed over these p-values. To map this FDR correction back to all other p-values of the SNP-eGene pairs, which are not the smallest, the empirical p-value that is closest to FDR of 0.05 is selected. Over the respective inverse-beta distributions, this p-value is transferred back to a nominal p-value threshold, that will be different for each SNP-eGene pair, as each has another beta distribution. Afterwards, all co-eQTLs lower than this specific p-value threshold in the corresponding SNP-eGene are defined as significant.

5.7.16. Evaluation of concordance of effect sizes across cell type specific co-eQTL

The concordance of effect sizes between co-eQTLs of different cell types was quantified using the r_b measure [191]. We followed an approach to estimate the errors across gene pairs r_e, as suggested in the original manuscript [191] with

$$r_e = r_p * \frac{n_s}{\sqrt{n_i * n_j}} \tag{5.1}$$

Here, r_p represents the correlation of co-expression levels between two cell types in the overlapping samples, n_s the number of overlapping samples, n_i and n_j the number of samples in cell type i and j.

As we did not test all SNP-eGene-co-eGene triplets in all cell types due to our filtering strategy, we additionally reported the number of triplets that were tested in both cell types to put the r_b values into context. In cases, where less than 10 co-eQTLs were tested

in both cell types, all connected with the same SNP-eGene, no robust estimation of the r_b was possible, and instead the value was set to missing (NA).

5.7.17. Replication in BIOS

To replicate the single cell co-eQTLs in the whole blood bulk dataset from BIOS [33], we applied a linear regression approach with an interaction term as used previously [46] (see also Methods section 2.2.4).

$$eGene = \beta_0 + \beta_1 * SNP + \beta_2 * co\text{-}eGene + \beta_3 * SNP * co\text{-}eGene \quad (5.2)$$

The model was solved via the *statsmodel* Python package [205], using an ordinary least squares model. Effect sizes and Benjamini-Hochberg corrected p-values from the interaction term were taken for the co-eQTL evaluation. We again applied the r_b values as defined above (subsection 5.7.16) to quantify the replication with the single cell co-eQTLs. Donors that were part of the single cell datasets were removed from the BIOS dataset before the analysis.

5.7.18. Co-eQTL subsampling

We investigate the impact of the experimental parameters when we randomly subsampled the number of cells (to 50, 150 and 250 cells) and number of samples (to 50 and 100) for CD4+ T cells. For the number of cells, we kept all individuals with fewer cells than the subsampling cutoff and subsampled the other individuals to that level. Nine individuals with fewer than 10 CD4+ T cells were excluded in general. We then repeated the co-eQTL mapping on the subsampled datasets, including all filtering steps.

5.7.19. Enrichment analyses

Results from three different enrichment analyses were combined to interpret the identified co-eQTLs, GO enrichment, TF enrichment and GWAS enrichment. All analyses were conducted separately for each SNP-eGene pair over all co-eGenes and per cell type, including all pairs with at least five co-eGenes. Each time, multiple testing correction was done separately per SNP-eGene using FDR and a significance threshold of 0.05 was chosen.

GO enrichment was performed via the R package *clusterProfiler* (version 4.0.5) [53] using as background set all genes that were tested in the co-eQTL analysis of the corresponding cell type.

TF annotations for the enrichment were obtained from all blood-associated cell lines in the ReMap 2022 database [192], which contains processed ChIP-seq peaks. TFs were annotated to co-eGenes by checking for an overlap of the TF peaks with the promoter region of the co-eGenes (+/- 2 kB of the transcription start site). Fisher's exact test was used for enrichment with a background set containing all genes tested in the co-eQTL analysis of the corresponding cell type. We took as additional evidence for TFs to be

direct regulators if the enriched TF was a co-eQTL itself and if the co-eQTL SNP lies in a TF peak (or a SNP in high LD with $R^2 \geq 0.9$). We used SNiPA (1000 Genomes Project, Phase 3 v5, European population, GRCH37 and genome annotations from Ensembl 87) to extract the necessary LD information for this [206].

For the GWAS analysis, we first identified GWAS loci among the SNPs and SNPs in high LD ($R^2 \geq 0.8$) in the GWAS catalog (update from 3/1/2022) [6]. The LD information was obtained with *LDtrait* (1000 Genomes Project CEU, GRCH37) [207].

Afterwards, we analyzed if the co-eGenes were enriched for common GWAS traits using *MAGMA* [193]. The required GWAS annotations were taken from a GWAS analysis of the GTEx consortium, which processed summary statistics for 114 traits uniformly [208]. *MAGMA* analysis was run separately for each trait, conditioned on default gene-level covariates and using again the 1000 Genomes Project, European population, for LD information. The background set contained all genes tested as co-eQTLs.

Specifically for the co-eGenes of rs11311017–*RPS26* we tested the positively associated co-eGenes and negatively associated co-eGenes independently for enrichment in all three analyses, as we observed differences in direction between the cell types.

6. Discussion

Understanding molecular consequences of genetic variants and epigenetic factors and how they interact to form complex phenotypes remains a key question in systems genetics. In this thesis, we showed that both the analysis of DNA methylation in a large bulk cohort and the use of single cell technologies lead to the discovery of novel associations and allow the analysis of cell type specific effects. During the projects, we developed new strategies and computational methods, considering in particular the cell type specificity of associations:

In our DNA methylation study, we have done the first bulk deconvolution iQTL analysis for DNA methylation, which identified several cell type dependent meQTL effects, and we highlighted the large influence of cell type proportions on the identification of eQTMs. In the next project, we developed the first generally applicable and fast tool for cell type specific single cell eQTL and DE power analysis, called *scPower*, which we expect to be an important method for experimental design in the future. Finally, we benchmarked and improved the strategies to identify co-eQTLs with single cell data, a new analysis direction that identifies upstream regulatory processes affected by eQTLs and facilitates so the connection between eQTL effects and gene regulatory networks.

Additional to the rigid analysis and careful interpretation of the different biological datasets, the public accessibility of the developed computational methods was very important for us in all projects, visible for example in the R package for *scPower*.

In the following, we will discuss the challenges and opportunities of each project in more detail and give an outlook on future research directions based on them.

6.1. Uncovering cell type specificity of DNA methylation

In the first project, we explored the relationships between DNA methylation, genetic variants and gene expression. DNA methylation is an important epigenetic mark, associated with various disease [119, 120, 121, 122, 123] and affected by different environmental factors [124, 125], however its exact mechanistic role in gene regulation is not fully uncovered yet [130]. Associations between DNA methylation and genetic variants (meQTLs) as well as between DNA methylation and expression (eQTMs) have already been mapped in previous studies [209, 38, 191, 210], but we improved the identification and interpretation of these associations with our large multi-ethnic cohort and with our cell type specific analyses. Our strategies, the use of interaction models for iQTL detection and the adjustments of expression and methylation for cell type proportions, are important examples of how to overcome a large shortcoming of many

bulk datasets, the combined measurement of different cell types in a tissue. For eQTLs, it is generally established that these associations are often tissue and cell type specific [43, 32, 44] and therefore masked in classical bulk studies. In contrast, the cell type specificity of meQTLs is not well explored yet and requires further research. First comparative studies, which showed a high overlap across tissue, have limited explanatory power due to small sample sizes [191, 211, 212].

Overall, we identified 11,165,559 meQTLs using stringent statistical thresholds and replication testing across populations. A large fraction of the meQTLs seem to be generally found across tissues, evident in good replication rates also in non-blood tissues, in line with results from previous studies [191, 211, 212]. Nevertheless, we still identified several interaction meQTLs (iQTLs), where the effect sizes changed dependent on the cell type proportion of the individuals. Such interaction models were successfully applied for eQTL studies before [42, 46, 41], but we were the first to utilize this approach for interaction meQTLs. The iQTL of rs174548-cg21709803 is a good example of the additional value of iQTLs: the iQTL showed the T cells specificity of the meQTL rs174548-cg21709803 and provides a potential link of the meQTL with asthma, over which the fatty acid metabolism can be connected with the disease.

For the eQTMs, the cell type composition played an even larger role than for the meQTLs, likely because both DNA methylation and gene expression pattern are cell type specific. When we corrected for basic covariates and genetic background, we identified 90,666 cis eQTMs and 54,807,559 trans eQTMs, which we called genotype adjusted (GTA) eQTMs. However, when we corrected additionally for the cell type proportion, only 769 cis eQTMs and 97,281 trans eQTMs remained, which we called cell proportion adjusted (CPA) eQTMs. This shows that most of the eQTMs in the GTA set were driven by cell type composition. They capture not differences between individuals in the population, but instead differences between cell types visible due to different cell type compositions of the individuals. Previous eQTM studies typically focused only on CPA eQTMs, using our approach or likewise adjustment strategies [38, 136]. However, both sets of eQTMs give valuable insights into gene regulation, given that the context is considered in the interpretation: the GTA eQTM set shows how the interplay between DNA methylation and gene expression is important for the cell type identity, while the set of CPA eQTMs captures the actual variability in the cohort between donors.

In follow-up analyses of the GTA and CPA eQTMs, we analyzed the context-specificity of eQTMs in order to identify general rules which CpGs are associated with which genes, a currently open question [130]. We found strong evidence for a context-specificity of both GTA and CPA eQTMs, because our machine learning models that predict eQTMs based on different genomic features had high performance values both across cohorts and tissues. The idea of the eQTM prediction model is an adaption and extension of the prediction model from Bonder et al. [38], which we improved in several points. We evaluated not only the prediction of positively versus negatively associated eQTMs, but also of eQTMs vs non-eQTMs. Additionally, we extended the set of eQTMs by using a larger cis distance, compared to Bonder et al., so that for example more positive eQTMs

and more eQTMs with CpGs outside the promoter region were included. On top of that, we use a different feature set for the prediction. Bonder et al. used histone modifications as features, while we changed to ChromHMM states, which are derived from histone modifications and easier to interpret biologically [24]. We also incorporated additionally other types of annotations, for example the 3D chromatin structure and gene specific features, to get a more complete picture of potential influences. Finally, we extended the analysis also for cross-tissue prediction of muscle eQTMs based a model trained on PBMC eQTMs. This gave important insights in the generalizability of our model, suggesting that the general mechanisms of eQTMs are similar in all cell types, when taking cell type specific genomic annotations into accounts.

With our extended sets of eQTMs and features, we inferred several interesting biological relationships: for the CPA eQTM models, a strong connection of promoter methylation and gene expression was visible, matching previous studies [130]. For the GTA eQTMs, we identified a connection between enhancer methylation and gene expression. For both CPA and GTA eQTM sets, the models indicated an impact of the 3D chromatin structure on the eQTM probability of CpG-gene pairs, with pairs in the same TAD being more likely to be associated. Overall, the feature importance evaluation of our models highlighted large differences between the genomic context of GTA eQTMs and CPA eQTMs. The prediction performance was high for both types of eQTMs, showing that both types are connected with certain genomic contexts. Following our interpretation, that GTA eQTMs represent mostly differences between cell types, our analyses showed that cell type differences are associated with enhancer methylation. This finding is supported by previous studies that identified enhancers as crucial elements to drive cell type identity and showed how DNA methylation is defining the enhancer activation status [213].

Finally, our prediction models provide not only useful information about the genomic context of eQTMs, but can also be used to identify candidate genes connected with EWAS CpGs. These candidate genes can facilitate the interpretation of EWAS CpGs, as gene functions are much better annotated and understood than CpGs, allowing for example pathway enrichment analyses. Our model could successfully prioritize candidate genes that are likely associated with EWAS CpGs from a study about methylation changes in muscle after regular exercise training [137]. We showed the credibility of the identified candidate genes, as they were enriched among associated DE genes from the same study and for GO functions associated with muscle activity. Considering the fast-growing number of EWAS studies [29], this is an additional valuable use-case for our eQTM models.

The models that predicted the direction of effect of the eQTMs did not perform as well as the models that distinguished between eQTMs and non-significant CpG-gene pairs. We assume that this might be caused by the very general feature set that differentiates not enough between activating and repressing factors, for example for the transcription factors. Furthermore, the eQTM set contains not only directly interacting CpG-gene pairs, but also indirectly associated pairs. While the restriction on cis eQTMs probably

reduces the number of indirect interactions, these might nevertheless exist and make the prediction of the effect direction more difficult.

Another limitation of the current study is the detection power, despite the already large size of our cohort with in total nearly 7,000 samples. This is especially visible in our iQTL analysis, which is affected by a larger multiple testing burden of testing different interaction terms and many weak effect sizes of iQTLs. The global iQTL analysis with the larger multiple testing burden found fewer meQTLs than the targeted testing of cosmopolitan meQTLs. The large number of 64% independent global iQTLs suggests power limitations of our current dataset that hinder the identification of more iQTLs. This also needs to be considered when interpreting the small number of iQTLs associated with BMI and smoking, despite the known connection of DNA methylation and both environmental traits [124, 125]. Larger, more well-powered follow-up studies might identify additional associations and give a more complete answer, which fraction of meQTL associations is cell type specific and which is generally identifiable in all tissues.

For the future, the eQTM prediction models show great potential for the application in follow-up projects. While our models comprised already a large and diverse genomic feature set, further annotations could be included. For example, we focused on gene level quantification in our project. A similar model with transcript level quantification could investigate the potential role of DNA methylation in splicing and include splice sides and the intron-exon structure as features.

The rise of single cell data opens new possibilities for cell type specific analysis (more in the next sections), also in case of single cell DNA methylation [78]. However, large single cell cohorts are still scare and the coverage of single cell DNA methylation does currently not reach reliable quantification on CpG level. For these reasons, we expect that the cell type specific analysis approaches for bulk cohorts, introduced in this project, will remain very valuable in the next years to make use of the extensive resource of bulk datasets, with increasing sample sizes.

6.2. Development of a novel single cell power analysis method

The previous section about our DNA methylation study highlighted the importance of the cell type specific analyses when studying gene regulation, as was shown before already for eQTL analyses [41, 43, 44]. Compared to bulk studies, single cell transcriptomics facilitates cell type specific analysis clearly, as shown in several recent single cell eQTL analyses [57, 58, 83]. However, single cell analysis has currently still some drawbacks, in particular the higher sparsity of the datasets and higher costs during the cohort generation. Both issues can be compensated (to some degree) with proper experimental design, for which an appropriate power analysis method is required.

For this purpose, we developed *scPower*, the first analytic power analysis method for single cell multi-sample experiments, which enables fast and user-friendly calculations. Other single cell power analysis methods are based on simulations, for example *powsimR*

[112] and *muscat* [113] for single cell DE studies and *splatPop* [151] for single cell eQTL studies. The results of the different approaches match very well, as shown in our comparison of *scPower*, *powsimR* and *muscat*. However, the analytic approach of *scPower* makes it far more efficient than comparable simulation-based approaches. This allows also the estimation of power for large cohorts, where resource requirements are very extensive for simulation-based methods, as well as the easy and fast comparison of different parameter settings. Of note, the simulation-based methods have the advantage of greater flexibility, for example testing the effect of different normalization methods or additional batch effects on the data. For this reason, both power analysis strategies, analytic and simulation-based, have their application areas.

We designed *scPower* to be very flexible for a large variety of use cases. It covers both DE and eQTL analysis with data from different single cell technologies and can be easily customized with priors dependent on the analyzed cell type and expected effect sizes. We recommend the use of pilot datasets to estimate these priors, where possible. Alternatively, *scPower* provides the functions to simulate priors. In the end, the selection of appropriate priors by the user is crucial for good power estimations, a general requirement for all power analysis tools.

We leveraged *scPower's* efficiency by including a parameter optimization function. It finds the parameter combination for a certain budget that maximizes the detection power. While the exact optimal parameters depend on the experimental priors, the technology and the costs, we saw in our evaluations a general trend that measuring a large number of cells at low read depth is in many cases the best setting. This is in line with the findings of Mandric et al. [160] who did a first analytic investigation of effective sample sizes for single cell eQTL studies. In contrast to Mandric et al., *scPower* estimates the power directly instead of the effective sample size and provides a generalizable method for a wide range of adjustable settings. Additionally, we saw in our evaluations that microfluidics-based single cell technologies, such as 10X Genomics and Drop-Seq, reached higher power for the same budget compared to plate-based technologies, such as Smart-seq, because a high number of cells can be measured more cost-efficiently. In the end, the optimal experimental design depends on the specific experimental question and given resources of the user, which all can be easily adjusted in *scPower*.

scPower was designed specifically for multi-sample single cell transcriptomics experiments based on the pseudobulk approach. This restriction to pseudobulk analyses is necessary, as each power analysis method is tightly coupled with a certain statistical testing procedure. Other analyses, for example the co-expression QTLs (chapter 5), and also other omics layers, for example single cell ATAC-seq data, are currently not covered by *scPower*. The same is true if a continuous cell type definition should be used, as pseudobulk analysis requires discrete cell types. Here, linear mixed models, such as CellRegMap [85], provide an alternative.

To further increase the usability of our tool, we plan to extend the default set of cell type specific expression priors of our R package by integrating the large collection of tissues from the Human Cell Atlas [71]. Other possible extensions for our tool in the

future would be the inclusion of power calculations for allele specific expression analysis and perturbation experiments, both of course again specifically for single cell data.

Currently, many new single cell cohorts for different contexts and with larger sample sizes are generated, for example within the sc-eQTLgen consortium [88]. We expect *scPower* to be here an important support for experiment design of these studies. The value of our tool is also visible in several review articles about best practices in single cell studies that recommend *scPower* for study design [214, 215]. All together, we envision that *scPower* will further promote the design of more well-powered single cell cohort to facilitate cell type specific analysis of gene regulation.

6.3. Detection and interpretation of single cell co-eQTLs

Simplified cell type specific eQTL analysis based on the pseudobulk approach is an important new direction that becomes possible with single cell technologies. On top of that, single cell data allows completely new types of analysis, such as the construction of individual specific gene regulatory networks using the multiple measurement points per individual [86]. This can improve the interpretation of genetic variants, because regulatory processes influenced by the variants can be better pinpointed. Exploring these new possibilities, we mapped of co-expression QTLs (co-eQTLs), i.e. genetic variants that change the correlation between two genes, in a large single cell cohort, in order to identify upstream regulators affected by eQTLs.

First targeted co-eQTL analyses that focused on a small number of selected loci have been done before [55, 83], but a more broad systematic identification of co-eQTLs has not been performed so far. Several open methodological questions hindered a large-scale co-eQTL analysis: although benchmarking studies evaluated single cell association metrics and network construction algorithms, no clear best performing method was identified there [179, 87]. In particular, nobody explored the methods for the specific use-case of co-eQTL identification yet. Additionally, sophisticated pre-selection strategies, which choose SNP-gene-gene triplets for co-eQTL testing, are necessary to reduce the huge search space and so the very large multiple testing burden. Furthermore, proper approaches for the downstream interpretation of the identified co-eQTLs have not been determined yet.

For this reason, we developed a novel strategy to identify and interpret co-eQTLs large-scale, addressing the aforementioned questions. We tested it on a single cell cohort of 173 human PBMC samples, where we identified 72 independent co-eQTL SNPs associated with 946 gene pairs in the six major cell types.

We chose Spearman correlation as the association metric to quantify the gene co-expression, as we got very robust results across different datasets, including single cell RNA-seq, bulk RNA-seq and CRISPR knockout associations. Additionally, it was very efficient to calculate for many gene combinations and easy to interpret. Other association methods could not improve our results, when we tested Rho proportionality [180] and GRNboost2 [181], which were both among the top performing methods in

different benchmarking studies [179, 87]. The same is true for alternative strategies, in particular merging cells to meta-cells [185] and ordering cells by pseudotime [183, 182]. However, a full benchmarking of all existing association metrics was beyond the scope of this project and all validations were specifically tested for our dataset.

An interesting result during the association benchmarking was that the single cell datasets correlated better with each other than the bulk datasets. We identified different aspects influencing these observations: first, single cell data needs to be strictly filtered for these results, as robust correlation estimates depend crucially on the removal of lowly expressed genes. Second, we showed how the Simpson's paradox could potentially falsify part of the bulk correlation results. Third, also the cell type composition is expected to influence the results, but the associations from the FACS-sorted, i.e. cell type specific, bulk datasets were not closer to the single cell associations than the associations from the whole-blood dataset. This could be caused by technical effects during FACS sorting, which can distort the expression quantification, or by mismatches in the sub cell types between the FACS sorted datasets and the single cell datasets, e.g. Naive CD4+ T cells and CD4+ T cells. In general, it needs to be kept in mind that scRNA-seq and bulk RNA-seq capture different types of variability, between-cell and between-person variability, respectively. A previous study showed that this can lead to the identification of different gene modules [216]. A full overlap of found associations between bulk and single cell is therefore not expected.

After choosing Spearman correlation, we added to our co-eQTL detection approach a strict filtering of SNP-gene-gene triplets, based on eQTL associations and the correlation strengths of gene pairs. The value of this filtering was proven by higher replication rates of the filtered co-eQTL set in a second independent cohort. Dependent on the co-eQTL analysis goal, other filtering approaches could be applied, e.g. using prior knowledge on gene-TF pairs, or targeting a certain subset of SNP-gene-gene triplets associated to a specific biological question. However, our goal was a broad analysis of co-eQTLs without requiring additional prior information.

The interpretation of the co-eQTLs is complicated by the fact that Spearman correlation captures both direct and indirect interactions. However, we utilized these associations and identified common pathways and TF binding sites using enrichment analyses. In the case of rs1131017-*RPS26*, we found so a potential explanation for the relationship between the SNP rs1131017, which is associated with several autoimmune diseases, and the ribosomal gene *RPS26*. Co-eGenes with positive effect sizes were associated with translation and well replicable across cell type. In contrast, co-eGenes with negative effect sizes were associated with lymphocyte activity and only found in T cells, several of them also directly linked to autoimmune diseases. Furthermore, we identified five co-eQTLs that could potentially be direct regulators of the process, as their binding sites were enriched among all co-eQTLs of the T cells. Taken together, this indicates that *RPS26* plays not only a role in translation, but is also involved in T cell regulation, as supported by another current study [217]. This additional function of *RPS26* provides a logical connection of the eQTL with autoimmune disease, which was debated before

[195]. All this information can only be gained by the additional knowledge about the co-eQTLs, not by the eQTL alone, proving the value of co-eQTL analysis to analyze genetic variants.

In many of the other cases, the interpretation of co-eQTLs was less clear. Overall, we see that our study is still underpowered, as we found a relatively small number of co-eQTLs compared to the expectations based on the targeted co-eQTL analyses [55, 83] and large eQTL analyses [33]. This complicates the enrichment analyses, which were in many cases not possible at all. Furthermore, the sparse scRNA-seq data restricted the set of tested genes. Specifically, TFs are often lowly expressed and might not be included in the co-eQTL analysis. Because of this, the TFs that are the direct regulators of the eQTL might have been missed. Nevertheless, we can still infer information about it and the underlying biological pathways over our enrichment analyses.

A disadvantage of our strict filtering approach is that a different set of triplets is tested in each cell type, which makes the comparison between cell types quite difficult. However, we saw that already the different cell type frequencies have a huge impact on the number of significant co-eQTLs, i.e. far fewer co-eQTLs are detected for less frequent cell types. Hence, the small overlap of co-eQTLs between cell types can not be interpreted reliable and larger cohorts with higher power are needed.

We plan to apply our improved co-eQTL strategy in future single cell studies. To generally aid the experimental design of future single cell co-eQTL cohorts, we explored the effects of different experimental parameters on the number of co-eQTLs using a subsampling approach. Both a higher number of cells and a higher sample size led to a clear increase of the number of identified co-eQTLs. This can also explain the low number of co-eQTLs for less frequent cell types as a power limitation of our current analysis. Of note, the subsampling showed also that the number of cells was important to get more robust correlation estimates for each sample, as a higher cells numbers resulted in higher concordance between individuals. The experimental design considerations for co-eQTLs matched with our observations for eQTL power analysis from *scPower*, which showed the large importance for both sample size and number of cells to increase the number of identified eQTLs.

For this reason, we expect that a far larger number of co-eQTLs is detectable with larger cohorts in the future, for example with data from the OneK1K cohort [57] and the sc-eQTLgen consortium [88]. We envision that our first evaluations in this project will provide the groundwork to properly design new cohorts, identify more robust co-eQTLs and interpret the found associations.

6.4. Conclusion and outlook

When comparing the different projects, both analysis directions - the integration of multiple omics layers and the advance from bulk to single cell technologies - have shown their value to increase our understanding of genetic and epigenetic gene regulation. A combination of both, single cell multiomics analysis, is expected to bring everything

together and enable far better insights into the molecular set-up of each cell.

Single cell transcriptomics alone is not sufficient to capture the full picture of all regulatory processes within each cell. As shown in our co-eQTL project, association measures such as Spearman correlation can be used to identify connections between genes, but the identification of direct regulating transcription factors remains difficult. It was recently shown that the combination of single cell RNA-seq and single cell ATAC-seq data can improve the construction of regulatory networks, as it includes information about open chromatin and so about likely active TF binding sites for each gene [218, 219]. Similar strategies could be used to improve our co-eQTL pipeline, for example to preselect gene-TF pairs for testing or to validate identified associations. While currently mostly single cell transcriptomics datasets are generated on population level, promising developments have been made for single cell ATAC-seq to measure open chromatin for large number of individuals [76, 77] and even to measure both scATAC-seq and scRNA-seq simultaneously in the same cell [82].

Furthermore, the combination with single cell methylation data, for which no cohort level datasets exist yet, will improve our understanding of genetic regulation. DNA methylation is an important epigenetic regulator for gene expression and very cell type specific, as shown in our DNA methylation project. With single cell data, this can be taken better into account and the set of cell type specific meQTLs and eQTMs can be extended.

For each new analysis, proper power analysis methods, such as *scPower*, are crucial for the success. The current algorithm behind *scPower* can build the basis for scATAC-seq power analysis and others.

The different projects in the thesis highlight many promising novel developments: the technical improvements of measuring omics layers on single cell level, followed by new computational methods to make use of these large and complex data sets. Together, they have the potential to increase our understanding about genetic and epigenetic regulation in healthy individuals and disease patients drastically in the coming years.

A. Supplement

A.1. Supplementary Figures and Tables for chapter 3

Annotation	Resource	Described component	Cell type specificity
ChromHMM states (15 state model)	Roadmap Epigenomics Project [24]	CpG	yes
DHS sites	ENCODE [35]	CpG	yes
TFBS	ENCODE [35]	CpG	yes
TFBS	ReMap [220]	CpG	yes
Chromatin interaction points	4DGenome data base [221]	CpG	yes
Super-enhancers	Hnisz et al. [222]	CpG	yes
CpG islands and CpG island shores	UCSC genome browser[223]	CpG	no
TATAbox and CpG island promoter	FANTOM5 project [224]	gene	no
House-keeping genes	Eisenberg et al. [225]	gene	no
TSS ChromHMM state	Roadmap Epigenomics Project [24]	gene	yes
Distance between CpG and TSS	Biomart [226]	pair	no
CpG within the gene body	Biomart [226]	pair	no
Pair within the same TAD or connected via HiC contact	Javierre et al. [227]	pair	yes
CTCF binding site between the CpG and the gene	ENCODE [35]	pair	yes

Table A.1.: **Overview over genomic annotations for ML models**
Genomic annotations used for the machine learning models with their source, the component of the eQTM they describe (the CpG, the gene or the CpG-gene pair together) and if it is a cell-type specific annotation.

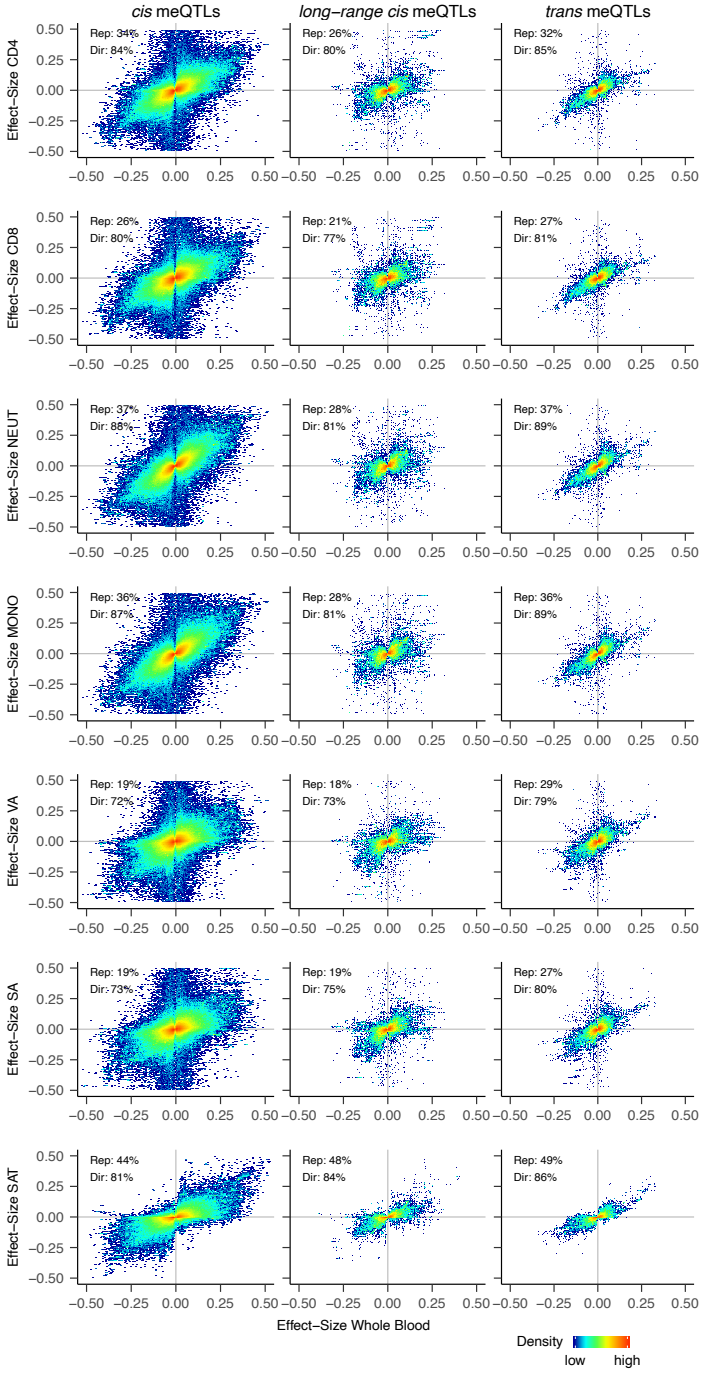


Figure A.1.: **Replication of meQTLs in different cell types and tissues.** Replication of meQTLs in **a-d** isolated white cell subsets (CD4+ and CD8+ T cells, neutrophils and monocytes), **e,f** isolated visceral and subcutaneous adipocytes (VA and SA) and **g** whole subcutaneous adipose tissue (SAT). Figure taken from [2].

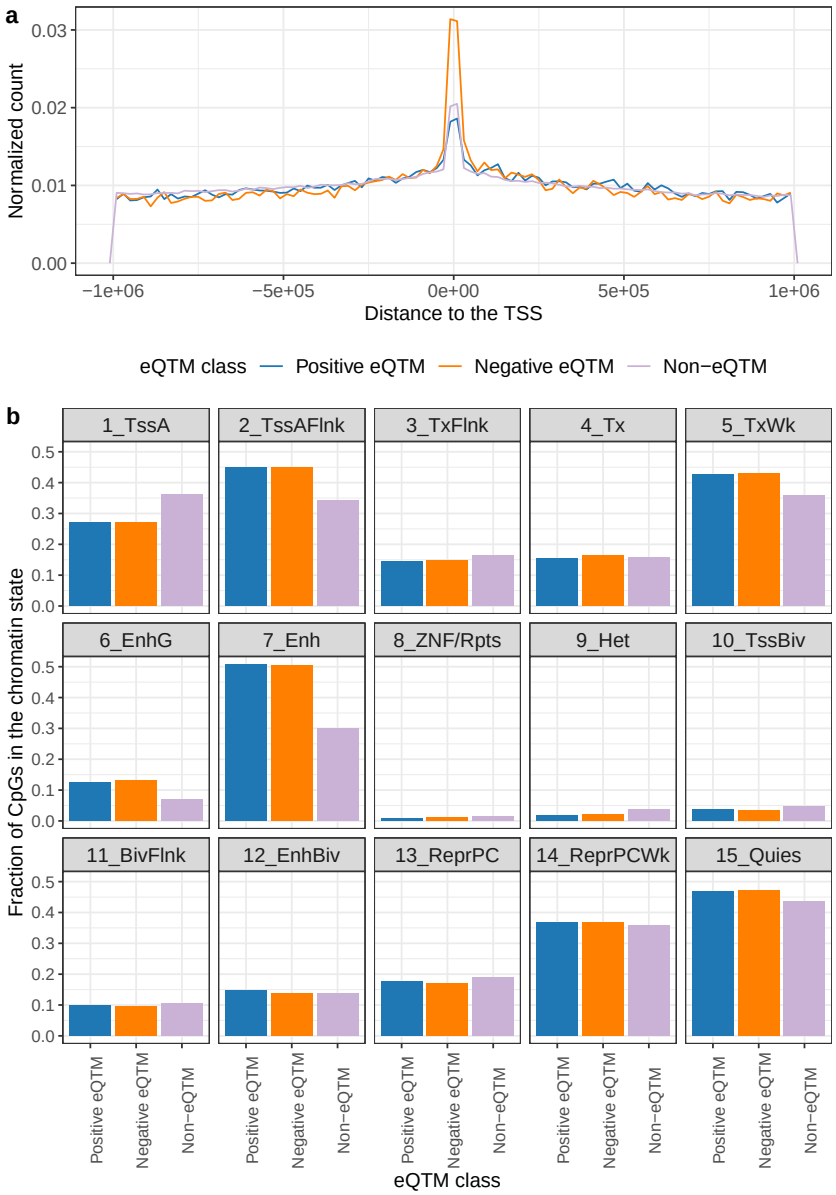


Figure A.2.: **Distribution of machine learning features among the eQTM classes.** **a** Distance distribution (CpG to TSS of the gene) for positive eQTMs, negative eQTMs and Non-eQTMs (non-significant eQTMs) based on the GTregressed eQTMs of the KORA cohort and an FDR threshold < 0.05. **b** Distribution of CpGs over the 15 different ChromHMM states for the same eQTM set. A CpG is defined to be in a state if it is in the state in at least one of the investigated cell lines.

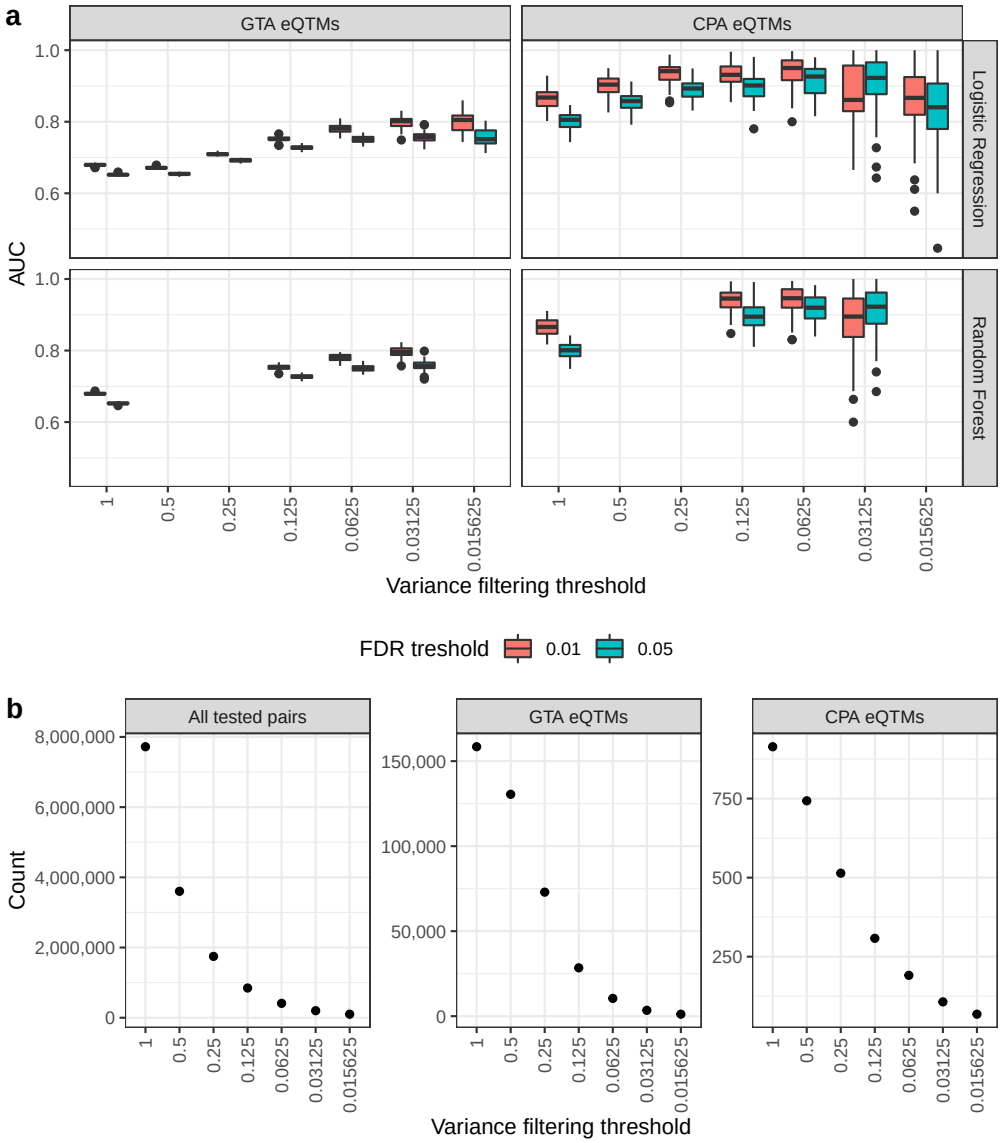


Figure A.3.: **Effect of FDR and variance filtering on the model performance.** **a** AUC performance for the eQTM prediction (eQTM vs non-significant pairs) in a 10-fold cross validation on the KORA dataset for different FDR thresholds (colors) and variance filtering thresholds (x axis). The variance filtering threshold represents the fraction of highly variable genes, which is kept after the filtering. The box plots show the AUC values across the different folds of the cross-validation. Models were trained for both the Logistic Regression and Random Forest, as well as for the GTA eQTMs and the CPA eQTMs **b** The total number of CpG-gene pairs and the number of significant eQTMs that remain after the variance filtering.

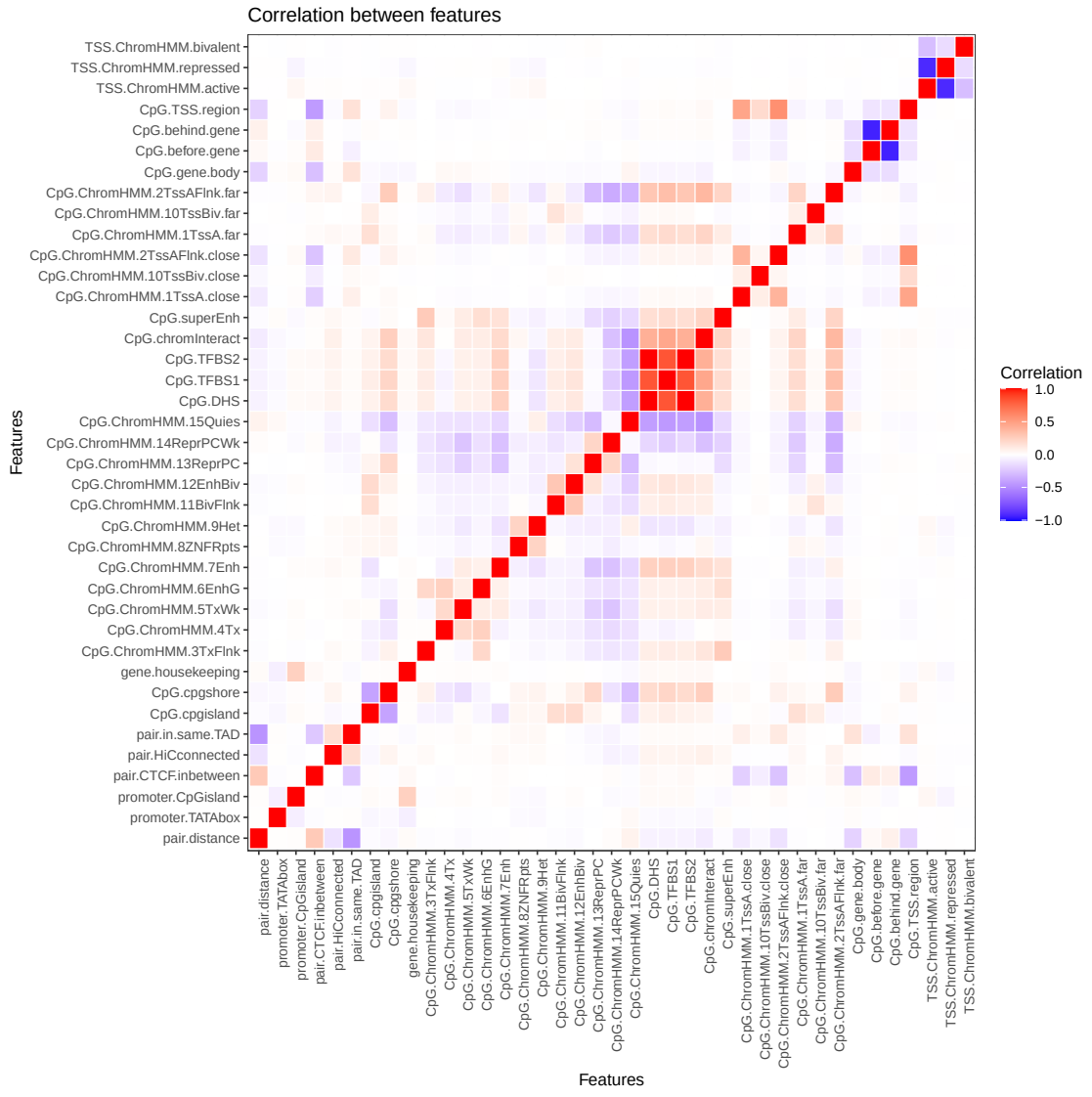


Figure A.4.: **Correlation matrix between all features**
The correlation is calculated on the full set of CpG-gene pairs (not filtering on significant eQTMs), but considering only pairs where the CpG was selected after the variance filtering (for the KORA cohort, GTA model).

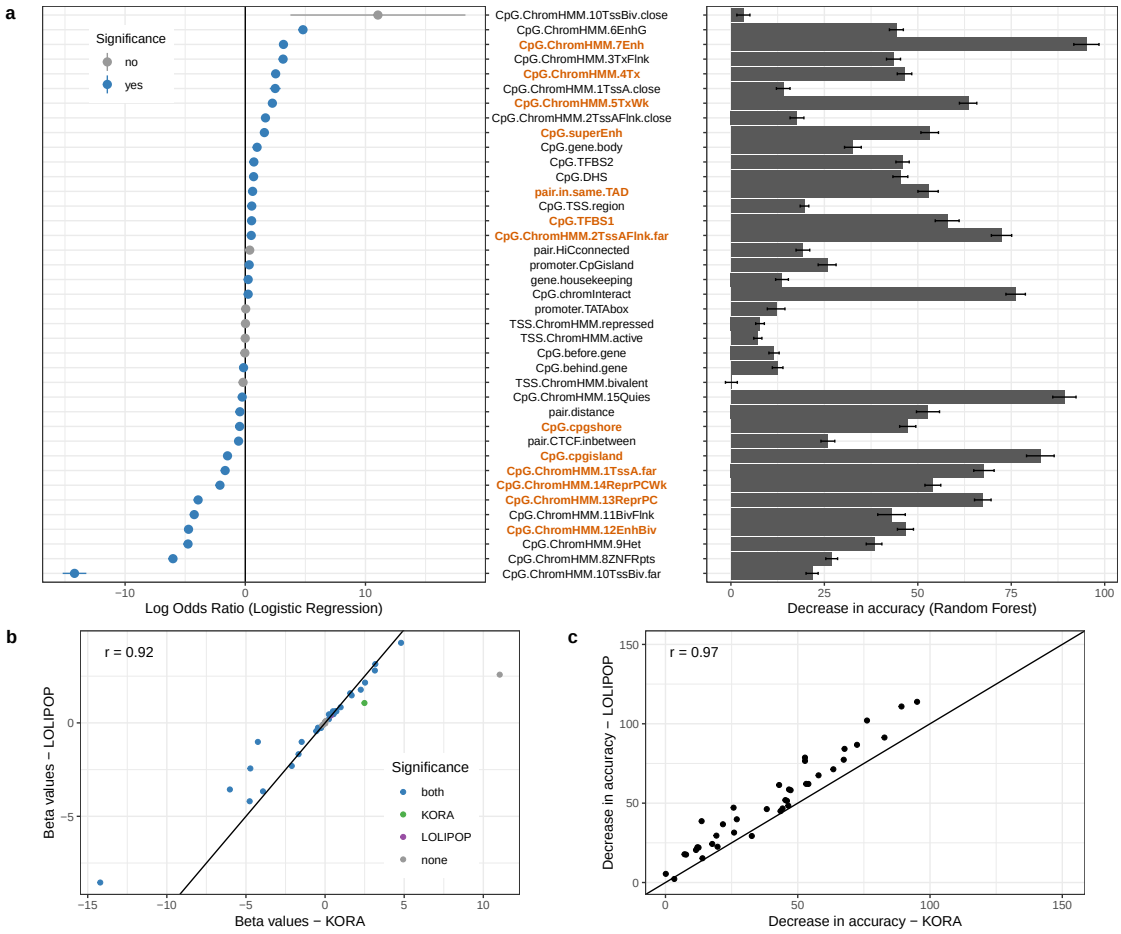


Figure A.5.: **Feature importance of the model eQTMs vs non-significant pairs for the GTA eQTMs**
Extended version of Figure 3.4, showing all features (without filtering the rare features).

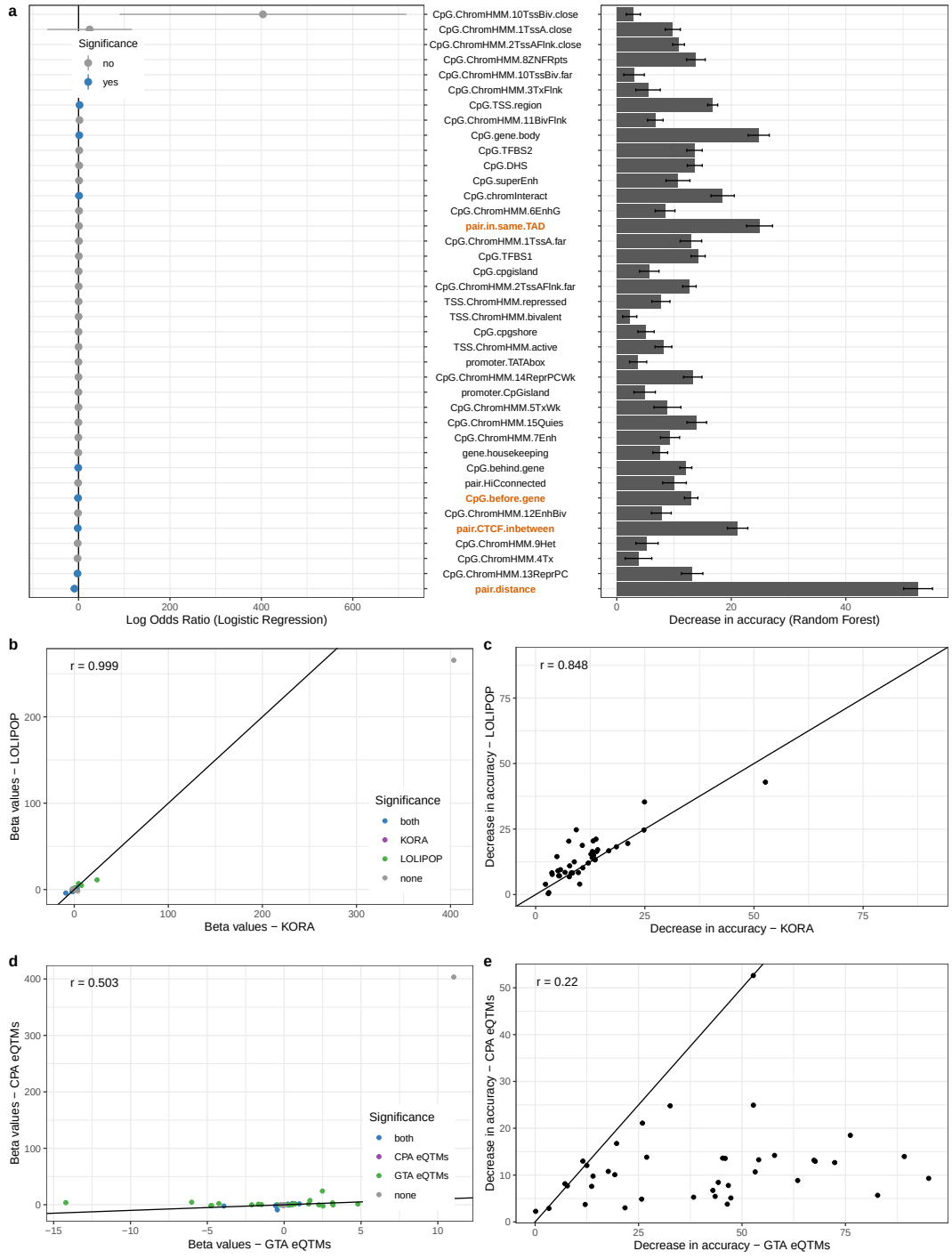


Figure A.6.: **Feature importance of the model eQTMs vs non-significant pairs for the CPA eQTMs**
Extended version of Figure 3.5, showing all features (without filtering the rare features).

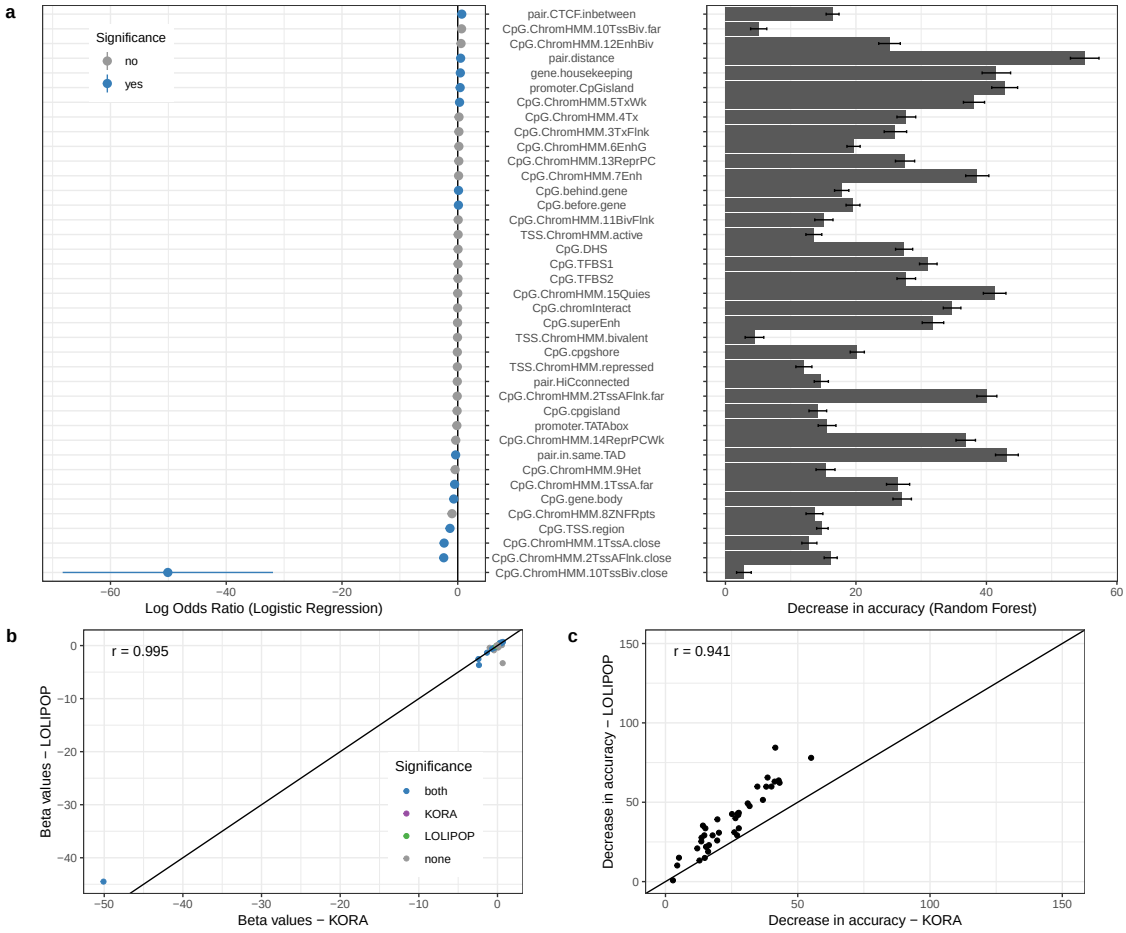


Figure A.7.: **Feature importance of the model positive eQTMs vs negative eQTMs for the GTA eQTMs**

a The feature importance of the KORA models with GTA eQTMs. For the Logistic Regression, bivariate models were trained for each feature plus the distance, for Random Forest, multivariate models. **b,c** Replication of the feature importance in the LOLIPOP model for Logistic Regression (**b**) and Random Forest (**c**). r values (top left of each plot) are the Pearson correlations between the scores from LOLIPOP and KORA.

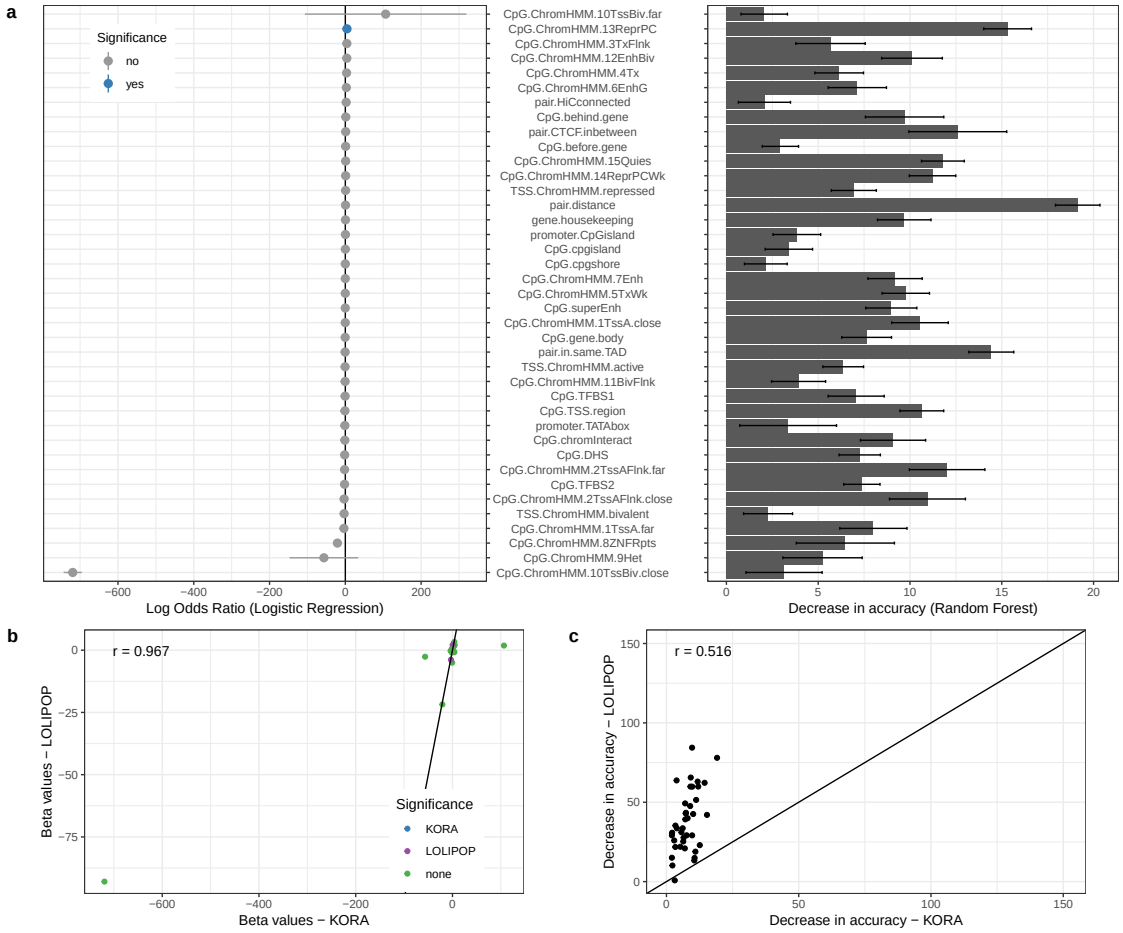


Figure A.8.: **Feature importance of the model positive eQTMs vs negative eQTMs for the CPA eQTMs**

a The feature importance of the KORA models with CPA eQTMs. For the Logistic Regression, bivariate models were trained for each feature plus the distance, for Random Forest, multivariate models. **b,c** Replication of the feature importance in the LOLIPOP model for Logistic Regression (**b**) and Random Forest (**c**). r values (top left of each plot) are the Pearson correlations between the scores from LOLIPOP and KORA.

A.2. Supplementary Figures for chapter 4

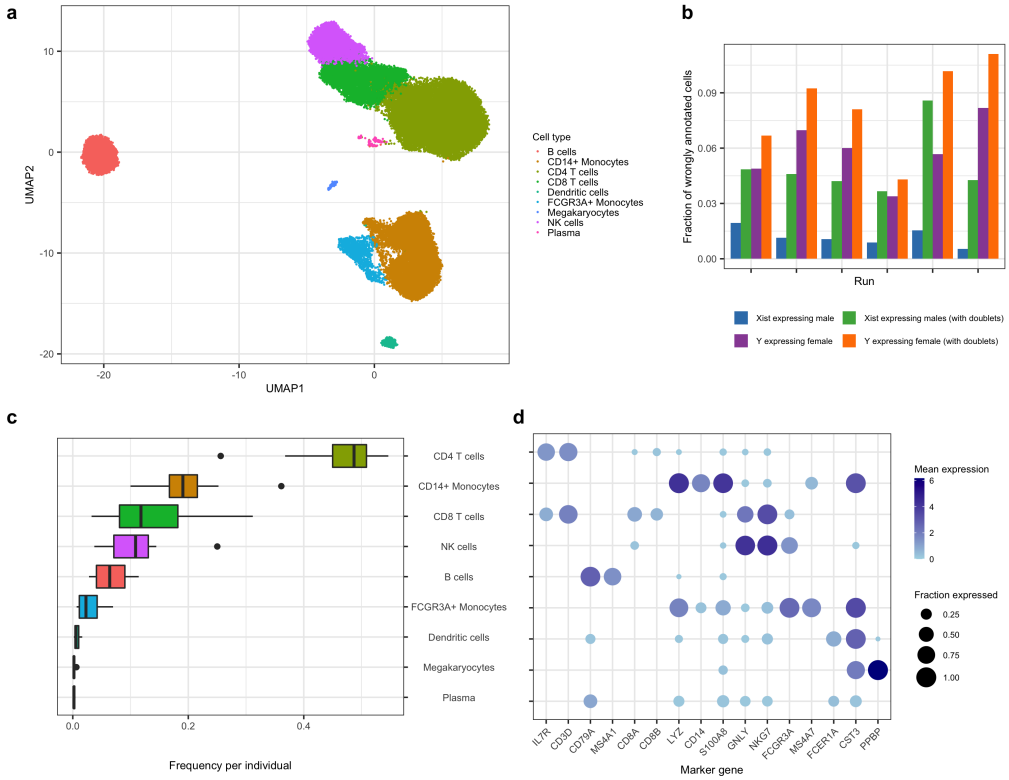


Figure A.9.: **Pilot PBMC data set to show expression prior estimation**
a UMAP visualization of all 6 runs, clustered using Louvain and annotated to cell types using marker genes. **b** Evaluation of Demuxlet assignment to the different individuals by testing the expression of sex specific genes per run. The error rate "Xist expressing male" shows which fraction of cells is assigned to a male donor from all cells expressing Xist. The "Y expressing female" shows which fraction of cells is assigned to a female donor from all cells having more reads mapped to chromosome Y than the median value. Both error rates decrease when Demuxlet and Scrublet doublets are removed. **c** Cell type frequencies for each individual (n=14 biologically independent samples). Box plots show medians (center lines), first and third quartiles (lower and upper box limits, respectively), 1.5-fold interquartile ranges (whisker extents) and outliers (black circles). **d** Marker gene distribution over the Louvain clusters. The color of the point visualizes its mean expression in the cluster, the size of the dot in how many cells of the cluster it is expressed (expression level larger than 0). Figure and legend taken from [1].

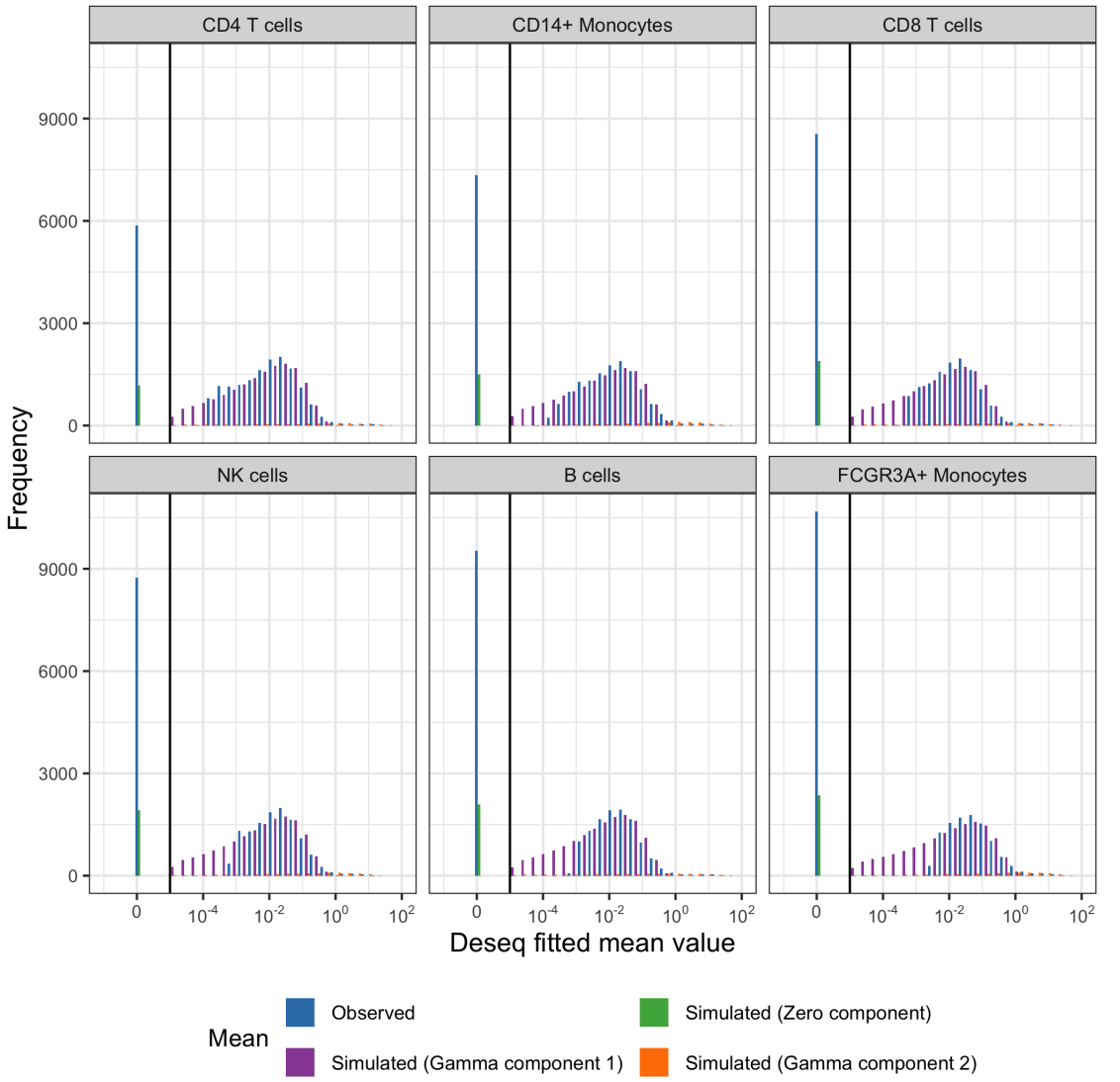


Figure A.10.: **Evaluation of gamma mixture fits for the expression means**
Gamma mixture fit (components in violett, green, orange) over all gene expression means for one batch of the PBMC pilot dataset compared to the observed distribution (blue). Fitted separately for each cell type (see panel titles), showing the 21,000 highest expressed genes for each cell type. Figure and legend taken from [1].

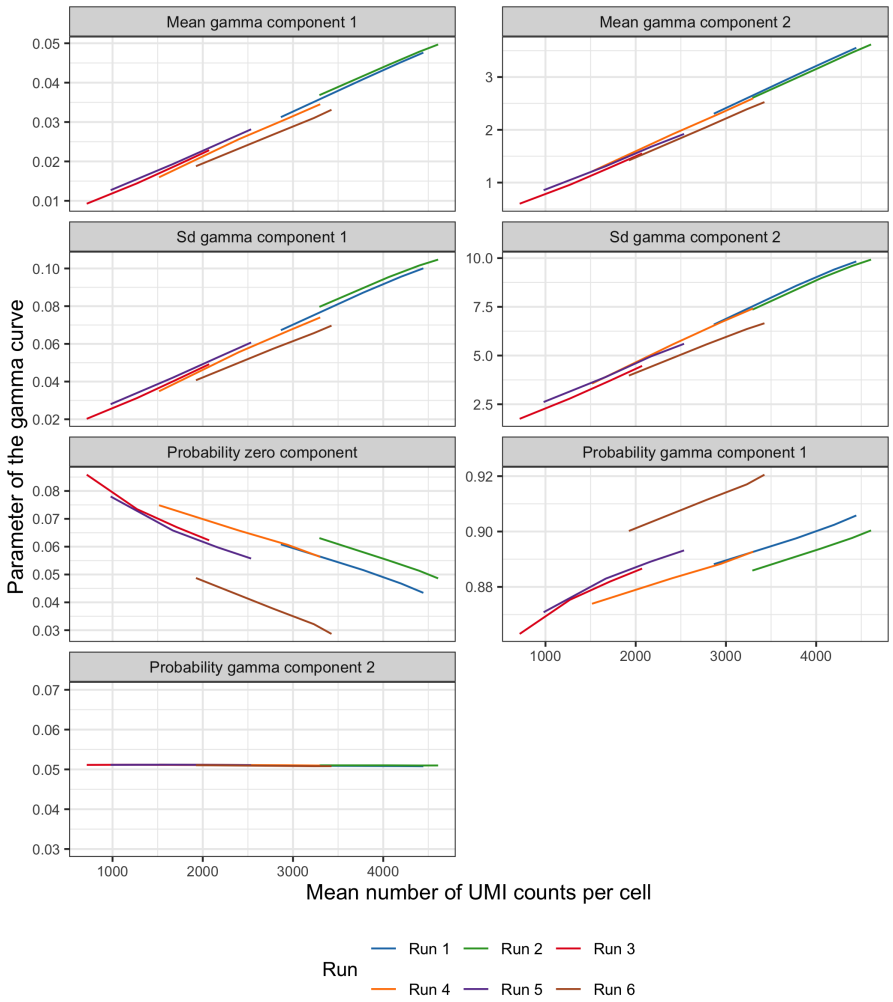


Figure A.11.: **Relationship between the parameters of the mixture distribution and the mean number of UMI counts per cell**
The two left censored gamma components of the mixture distribution are parametrized over their means and standard deviations (row 1 and 2), additionally there are three probability parameters (row 3 and 4) showing the proportion of each of the three components, the zero component and the two gamma components. The fits were performed for each cell type separately, here shown for the CD4 T cells. There is a linear relationship between the mean and standard deviation parameters of the gamma components and the mean UMI counts (row 1 and 2). The probabilities of the zero component and the first gamma component show a linear relationship to the mean UMI counts (row 3). The probability parameter of the second gamma component stays constant (row 4). The other cell types show the same pattern. Figure and legend taken from [1].

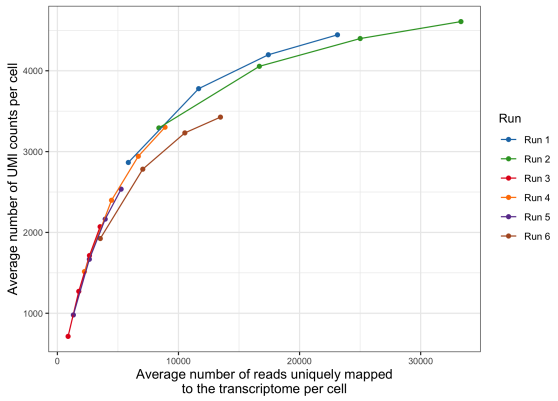


Figure A.12.: **Relationship between UMI counts per cell and average number of reads that were uniquely mapped to the transcriptome per cell**
Visualizing the relationship between the mean UMI counts per cell and number of reads for the different runs from the PBMC pilot dataset. The relationship can be described with a logarithmic fit. Figure and legend taken from [1].

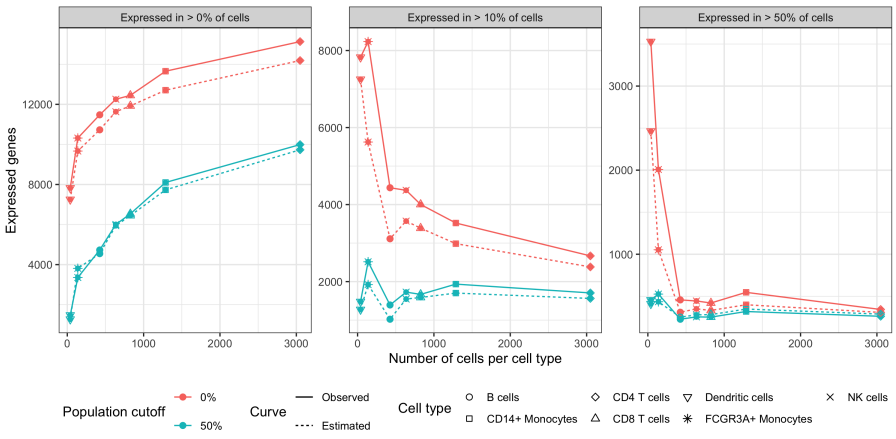


Figure A.13.: **Expression model with percentage cutoff**
Expression model when applying a percentage cutoff (expressed in x% of the cells) instead of an absolute cutoff (> x counts in pseudobulk) for the individual level threshold. Calculated for the same data set as in Figure 4.3 B-C. The observed number of expressed genes (solid lines) closely match the ones estimated with scPower (dashed lines). Curves were calculated once with a population cutoff of expressed in > 50% of the individuals (blue lines) or once without any population level filtering (red lines). Figure and legend taken Reviewer's Response of [1].

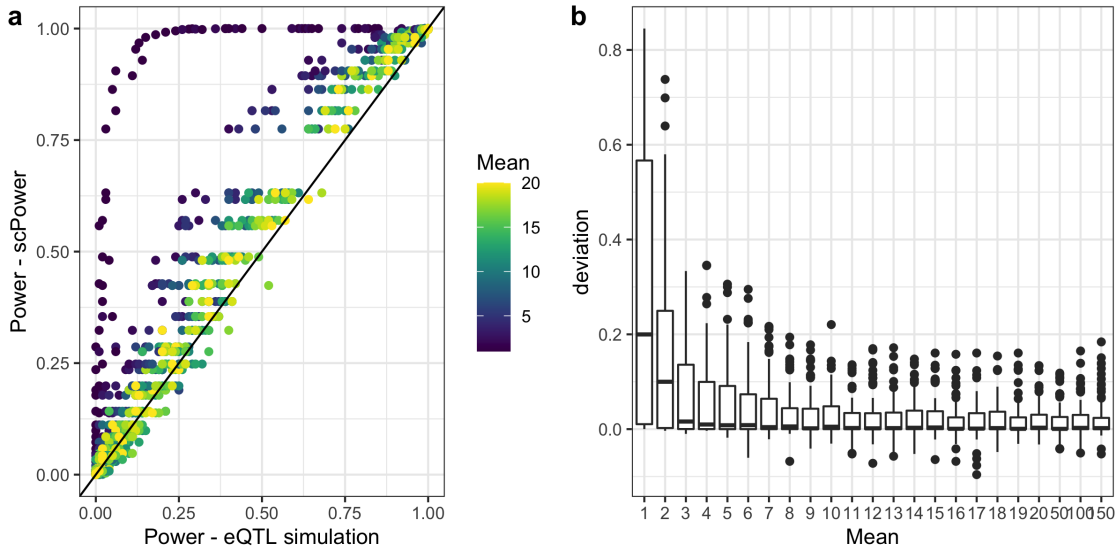


Figure A.14.: **Relation between eQTL power and expression mean in a simulation study**
a The simulated eQTL power is compared to the analytic power calculated with scPower for a range of effect sizes (R^2 between 0.1 and 0.6), sample sizes (between 20 and 150) and FWER-corrected p-values (between 0.05/(10*1,000) and 0.05/(10*10,000)). The color coding shows the mean count value. **b** The deviation between the analytic power and the simulated power is stratified by the expression mean used in the simulation. Box plots show medians (center lines), first and third quartiles (lower and upper box limits, respectively), 1.5-fold interquartile ranges (whisker extents) and outliers (black circles). Figure and legend taken from [1].

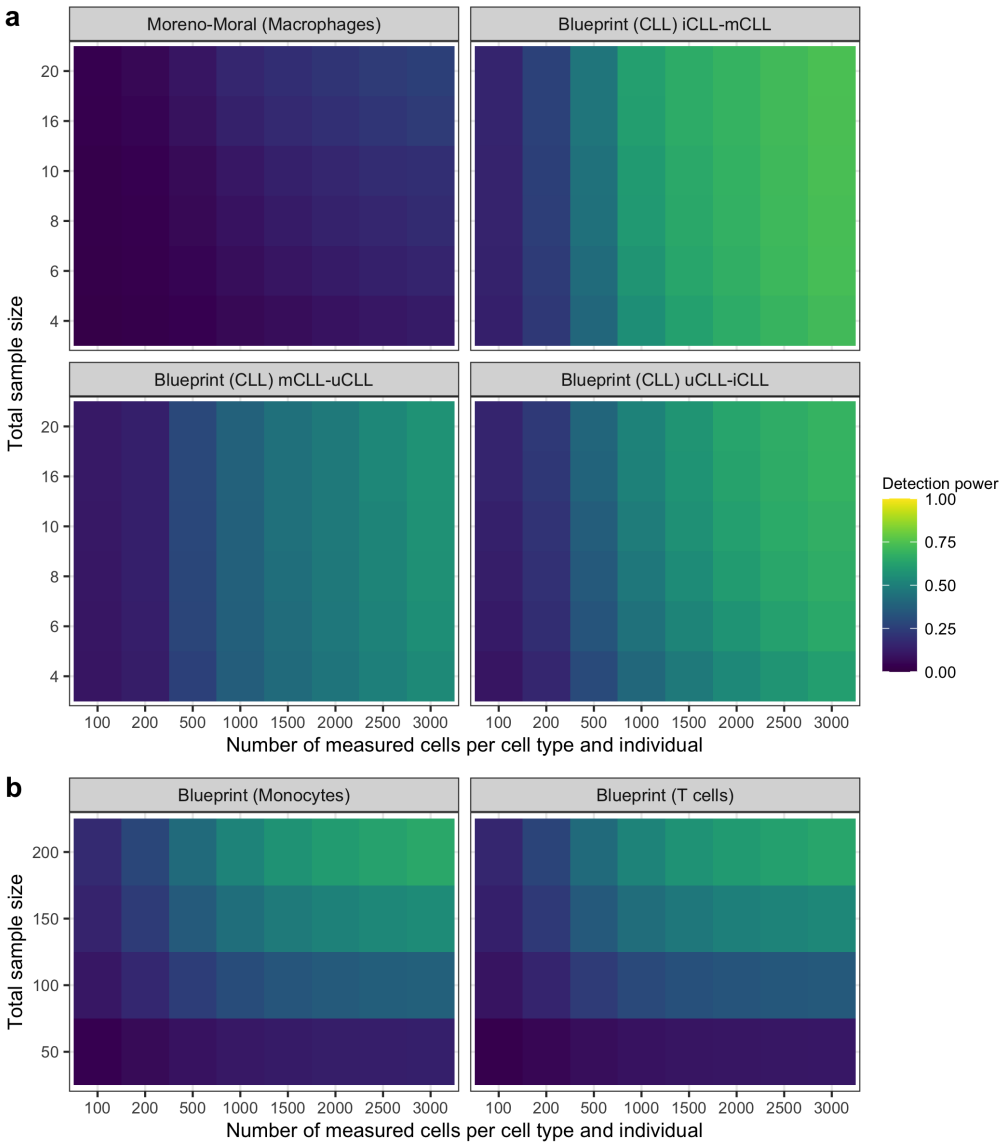


Figure A.15.: **Detection power using observed priors from reference studies**
Detection power for DE genes **(a)** and eQTL genes **(b)** dependent on the study, total sample size and the number of measured cells per cell type for a transcriptome mapped read depth per cell of 20,000. The fold change for DE genes and the R^2 for eQTL genes is taken from published studies, together with the expression rank of the genes (studies shown in panel titles). The expression profile and expression probabilities in a single cell experiment with a specific number of samples and measured cells was estimated using our expression prior, setting the definition for expressed to > 10 counts in more than 50% of the individuals. Figure and legend taken from [1].

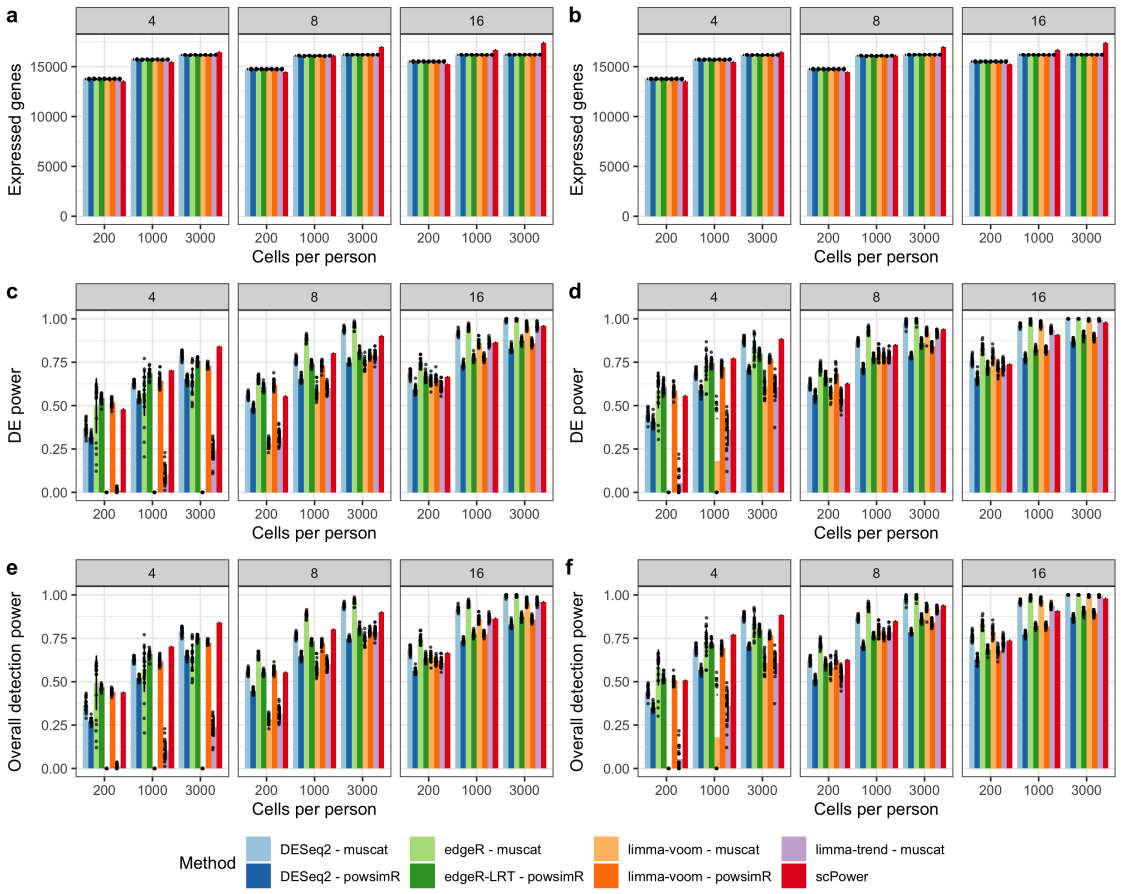


Figure A.16.: **Comparison of *scPower* with the simulation-based methods *powsimR* and *muscat* in combination with different DE methods**
Difference in number of expressed genes **(a,b)**, DE power **(c,d)** and overall detection power **(e,f)** between *scPower* and simulations. The adapted version of *powsimR* was run with *DESeq2*, *edgeR-LRT* and *limma-voom*, using the mean-ratio method (MR) for normalization. The adapted version of *muscat* was run with *DESeq2*, *edgeR*, *limma-trend* and *limma-voom*. The power was evaluated for sample sizes of 4, 8 and 16 (see panel titles) and for 200, 1000 and 3000 cells per person (x axis). Both FWER adjusted power **(a,c,e)** and FDR adjusted power **(b,d,f)** were evaluated. The bar plots represent the mean power over n=25 simulation runs of *powsimR* and *muscat*, the error bar shows the standard deviation and the points represent each individual simulation run. *scPower* as an analytic solution always provides the same result (so n=1 here). Figure and legend taken from [1].

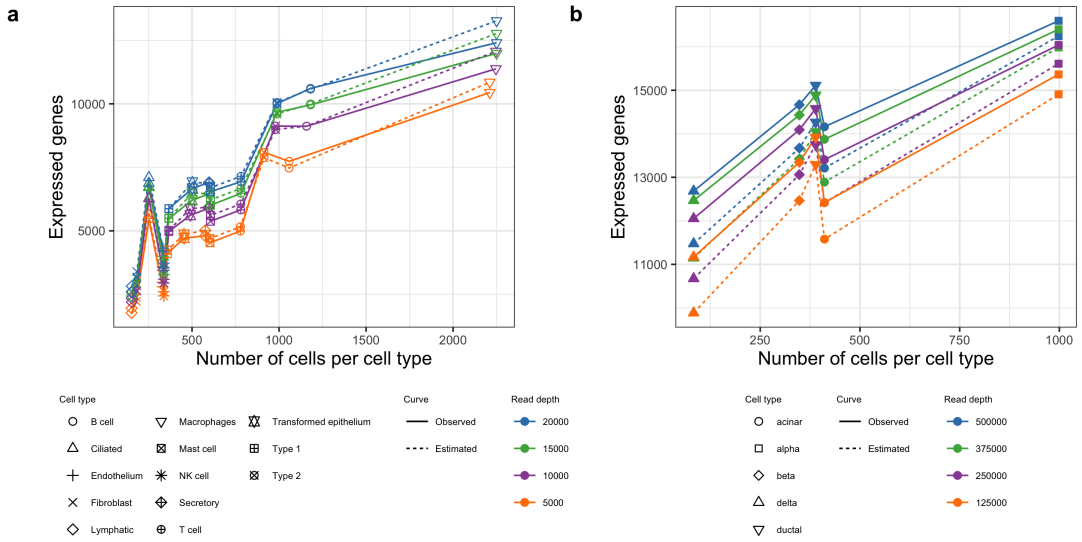


Figure A.17.: **Gene curve fits for different single cell technologies**
Evaluation of the expression probability from *scPower* for a lung data set measured with Drop-seq **(a)** and for a pancreas data set measured with Smart-seq2 **(b)**, both subsampled to different read depths (represented by line color). The solid lines represent the observed gene curves, the dashed lines the fitted curves. The point symbol visualizes the cell type. Gene expression criteria are chosen as UMI counts > 10 in all cells for Drop-seq **(a)** and read counts > 10 per kilobase transcript in all cells for Smart-Seq2 **(b)**. Figure and legend taken from [1].

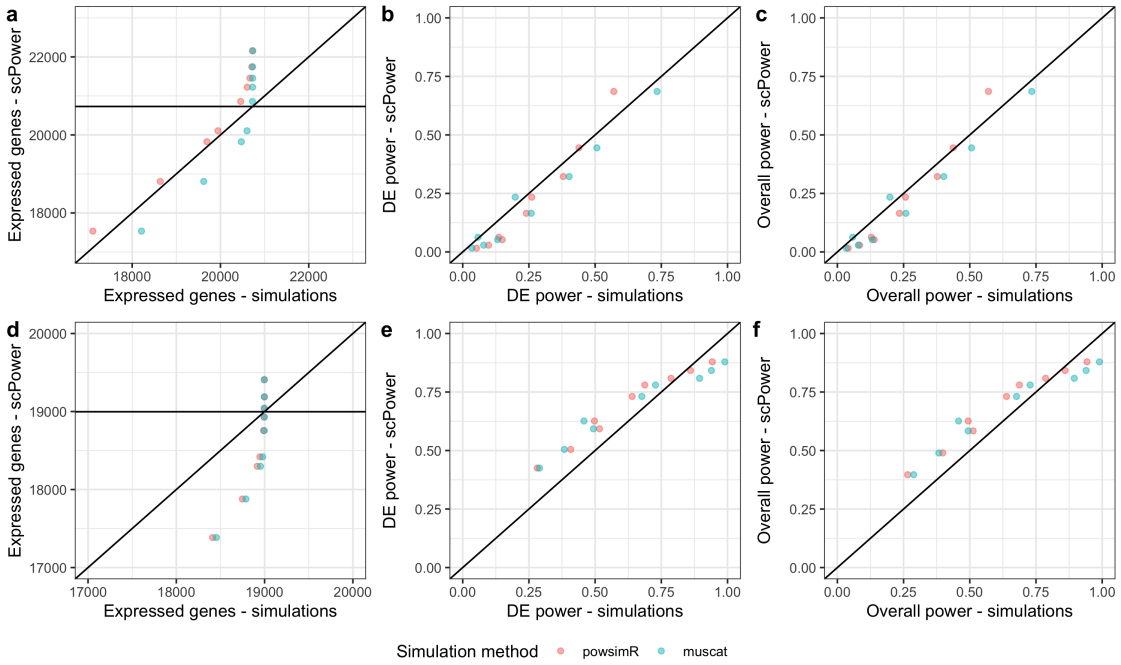


Figure A.18.: **Comparison of *scPower* with the simulation-based methods *powsimR* and *muscat* for other single cell technologies**
Evaluating expression priors of a lung dataset measured with Drop-seq **(a-c)** and a pancreas dataset measured with Smart-seq2 **(d-f)** by comparing difference in number of expressed genes **(a,d)**, DE power **(b,e)** and overall detection power **(c,f)** between *scPower* (y axis) and simulations of *powsimR* and *muscat* (x axis). The adapted version of *powsimR* was run with *edgeR-LRT* using the mean-ratio method (MR) for normalization and the adapted version of *muscat* with *edgeR*. The power was evaluated for sample sizes of 4, 8 and 16 and for 200, 1000 and 3000 cells per person, always using FDR adjusted p-values. Figure and legend taken from [1].

A.3. Supplementary Figures for chapter 5

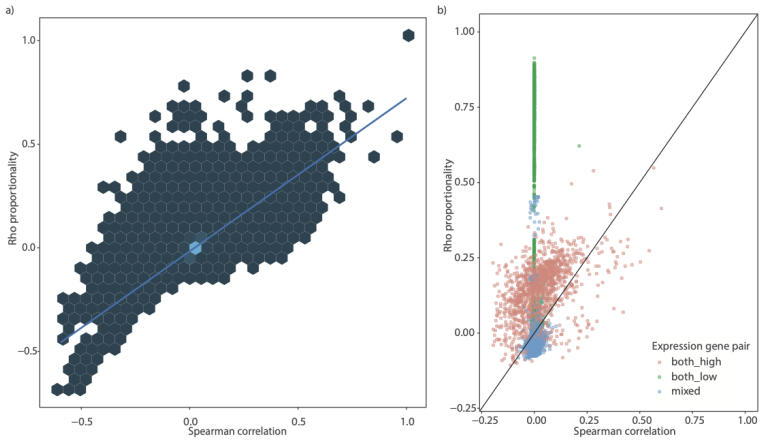


Figure A.19.: **Comparison between Rho proportionality and Spearman correlation** **a)** Comparison for genes expressed in at least 5% of the monocytes in Oelen v3 dataset. Colors indicate the density (light color for higher density). **b)** Comparison for very highly expressed genes (expressed in at least 90% of the cells) and very lowly expressed genes (expressed in 0-5%), both times sampling 50 examples to increase visibility of the scatter plot. Colors represent type of gene pair (either both highly expressed, both lowly expressed or one highly and one lowly expressed). Figure and legend taken from [3].

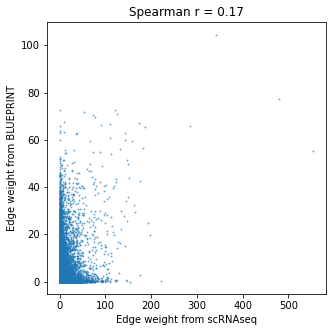


Figure A.20.: **Comparison of GRNBoost2 correlation between single cell and bulk** Comparison of the co-expression profiles (edge weights) gained from GRNBoost2 between the single-cell monocytes from Oelen v2 dataset with the bulk RNA-seq dataset from BLUEPRINT (classical monocytes). Only genes were taken that were expressed in at least 50% of the cells for the single-cell dataset. Figure and legend taken from [3].

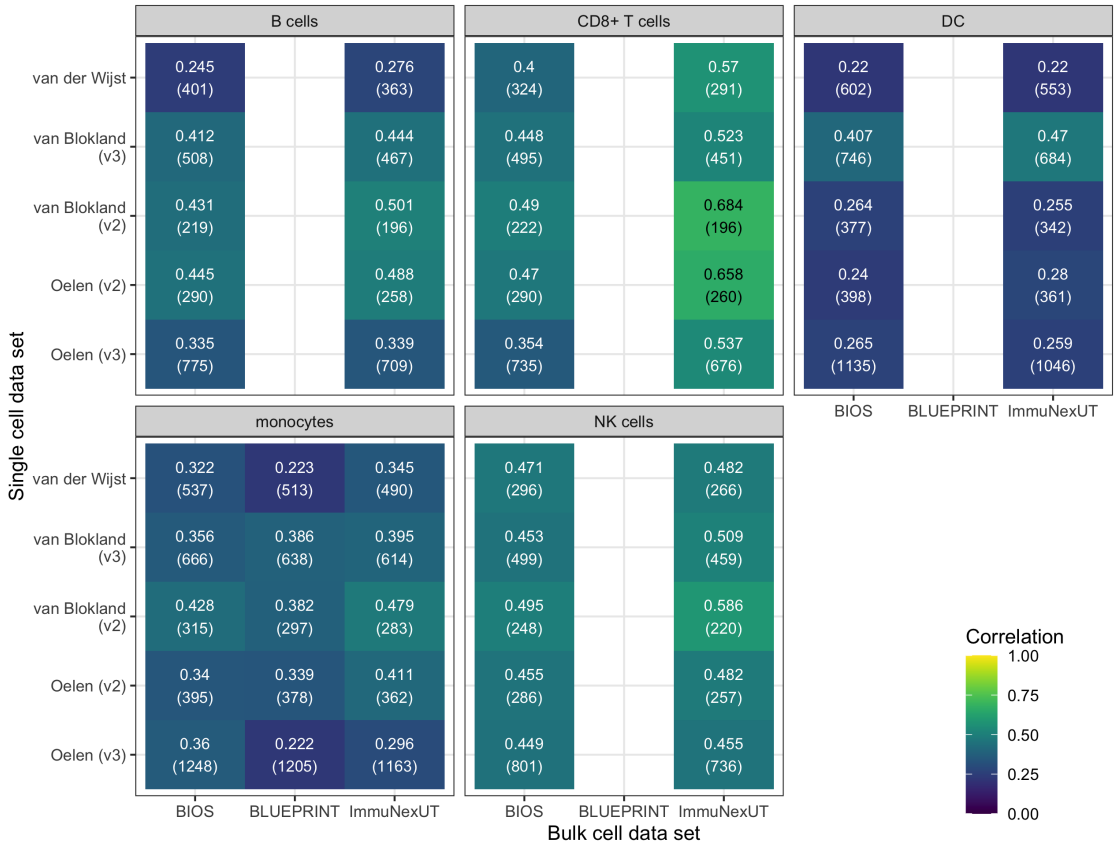


Figure A.21.: **Comparison of Spearman correlation between single cell and bulk datasets for different cell types**
Extension of Main Figure 5.2 b for other cell types: Comparison of the co-expression profiles between the single-cell datasets with the bulk RNA-seq datasets from BLUEPRINT, ImmuNexUT (both measuring FACS sorted cell types) and BIOS (whole blood). For BLUEPRINT, classical monocytes were measured, for ImmuNexUT, the compared cell types were naive B cells, naive CD8+ T cells, myeloid DCs, classical monocytes and NK cells. Again, only genes were taken that were expressed in at least 50% of the cells for the single-cell dataset. The number of tested genes is shown in brackets in each square below the exact Spearman correlation value. Figure and legend taken from [3].

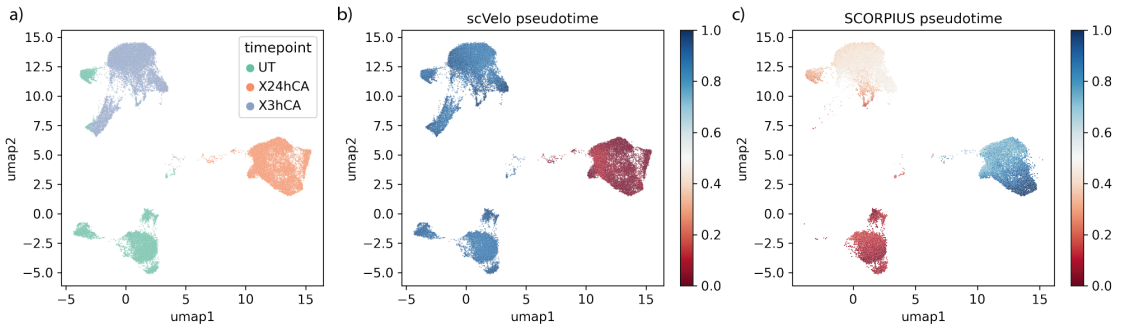


Figure A.22.: **Exploring different pseudotime methods**
Comparison of sampled time points during the stimulation experiment (a), temporal ordering from RNA velocity with scVelo (b) and ordering from SCORPIUS (c), plotting cells each time in the same UMAP. Estimated for Oelen v3 dataset, classical monocytes. Abbreviations for the stimulation experiment: UT: untreated, 3hCA: 3 hours after CA stimulation, 24hCA: 24 hours after CA stimulation. Figure and legend taken from [3].

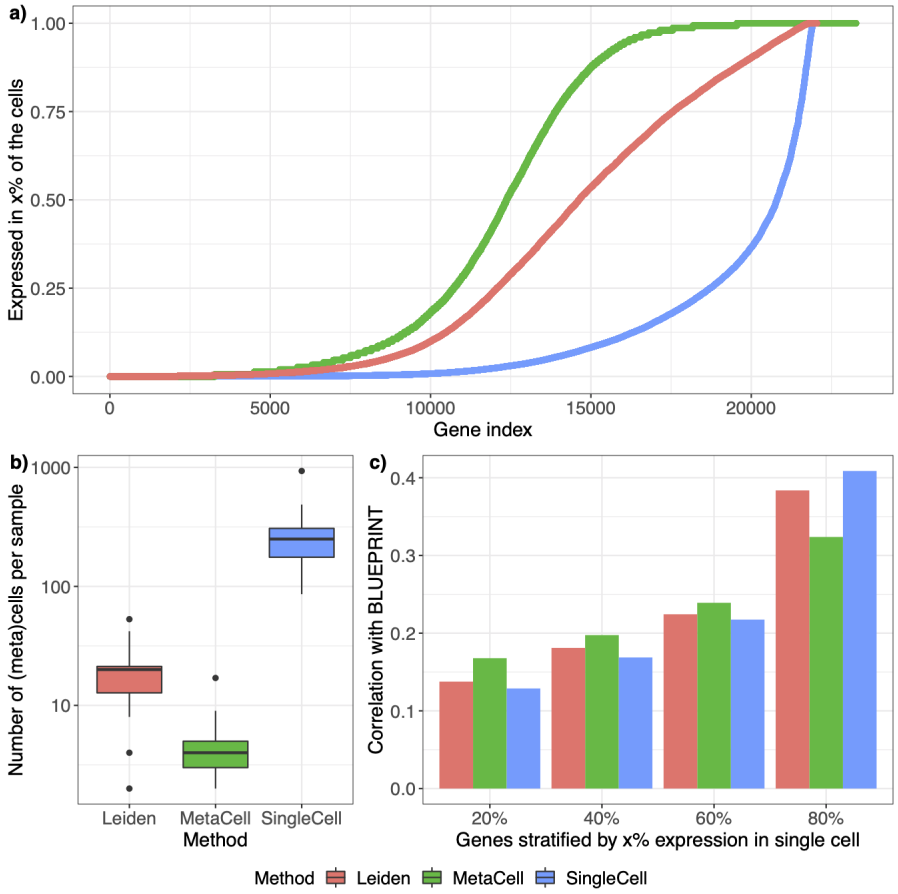


Figure A.23.: **Grouping cells to meta-cells**
Similar cells were grouped to meta-cells, using either the original MetaCell algorithm (parameters shown in the plot: K=20, minCells=10) or our own implementation based on the Leiden algorithm (parameters shown in the plot: resolution=20) (see Methods for detail). All methods were applied to the Oelen v3 dataset, Monocytes. **a)** Both meta-cells generated from Leiden clustering and from the MetaCell algorithm lead to more genes expressed in at least x% of the cells compared to the original single cell data (visualized here via a cumulative density function). **b)** In contrast, the number of (meta)cells per sample is reduced with both algorithms drastically, this way reducing the number of measurement points to infer the correlation per sample. **c)** To benchmark the performance, the correlation with the BLUEPRINT bulk data set was calculated. To evaluate how lower expressed genes are affected by the meta-cell grouping, the correlation is calculated separately for gene pairs were the non-zero expression level of both genes is between 20%-40%, 40%-60%, 60%-80% and 80%-100% of the cells (showing the first number on the x- axis). Figure and legend taken from [3].

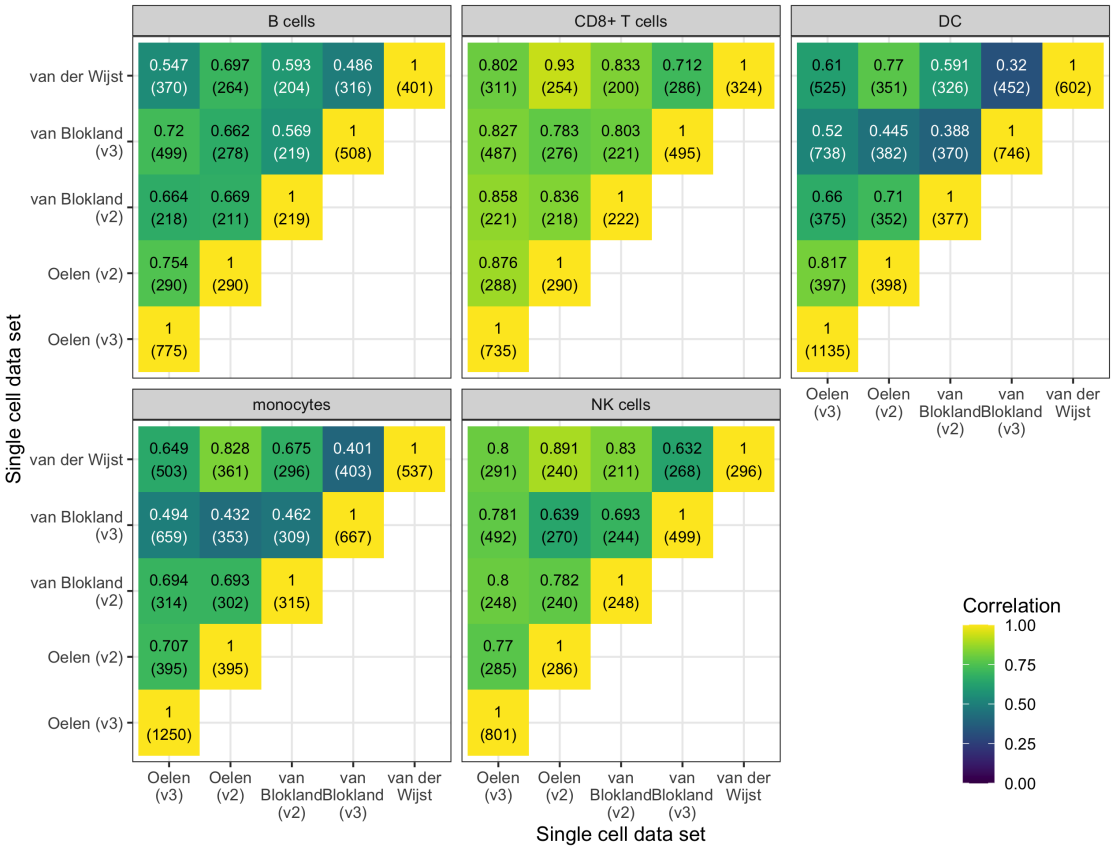


Figure A.24.: **Comparison of Spearman correlation across single cell datasets for different cell types**
Extension of Main Figure 5.2 a for other cell types, comparing different single-cell datasets. Spearman correlation of the Oelen v3 and v2 datasets, the van Blokland v2 and v3 datasets and the van der Wijst dataset were compared with each other, taking genes expressed in at least 50% of the cells in the corresponding datasets. Figure and legend taken from [3].

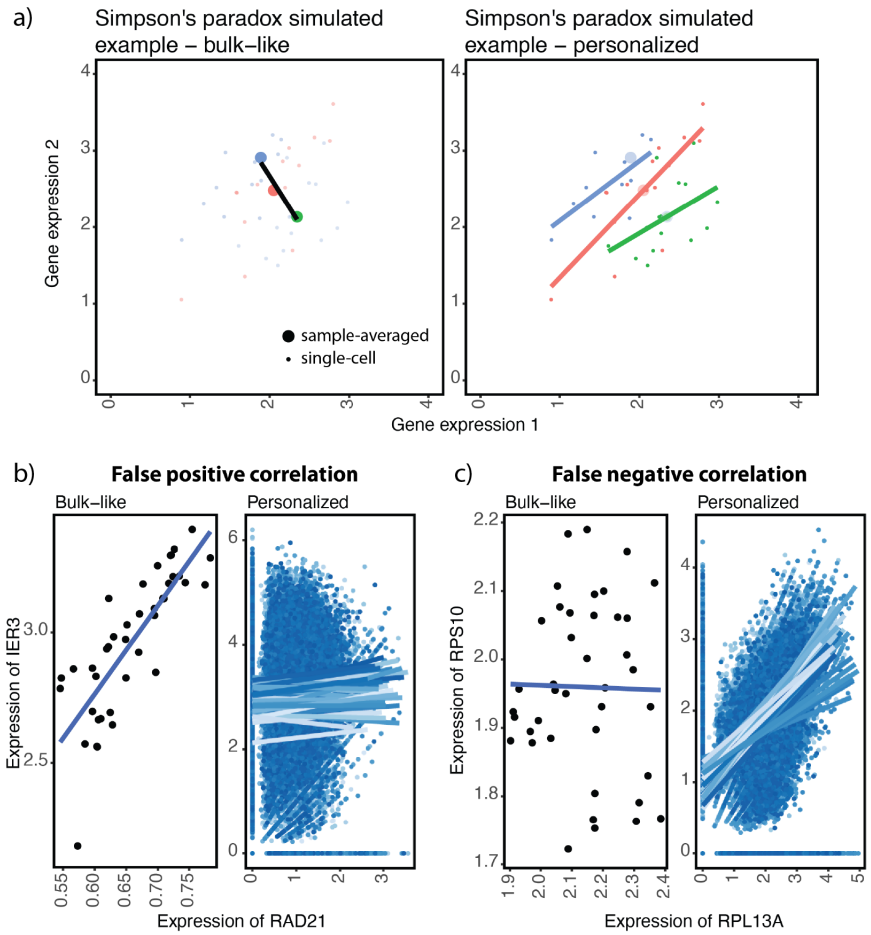


Figure A.25.: **Simpson's paradox in expression data**
a) A simulated example showing how Simpson's paradox can appear in expression data. The colors of the dots represent the sample. Each small dot represents a cell, and the large dots represent the average expression across cells for that individual. The line in the left figure is the regression line for the sample-averaged dots, and the lines on the right are the regression lines for all cells for each of the three samples. **b)** An example showing a false positive correlation identified by the pseudobulk expression data but not identified in the personalized manner **c)** An example showing a false negative correlation not identified if aggregating the scRNA-seq data with the pseudobulk approach, but a true correlation identified by the personalized expression data. Figure and legend taken from [3].

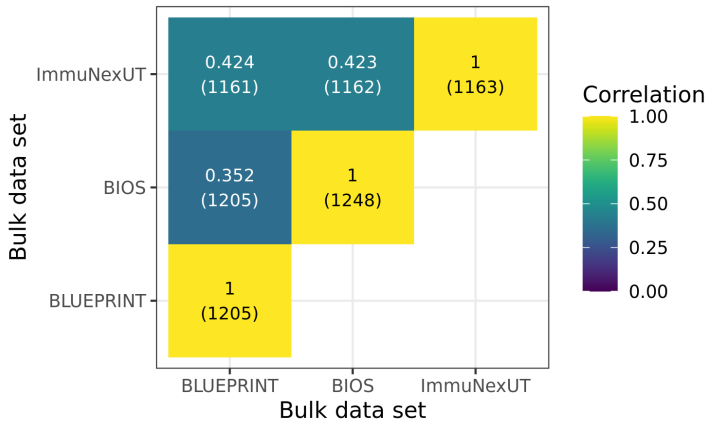


Figure A.26.: **Comparison of Spearman correlation across bulk datasets for monocytes** Extension of Main Figure 5.2 d for monocytes: Comparison of the co-expression profiles between the bulk RNA-seq datasets from BLUEPRINT, ImmuNexUT (both measuring FACS sorted classical monocytes) and BIOS (whole blood). In all datasets, only genes expressed in 50% of the cells from the Oelen v3 dataset were selected, to make it comparable with Supplementary Figures A.24 and A.21. The number of tested genes is shown in brackets in each square below the exact Spearman correlation value. Figure and legend taken from [3].

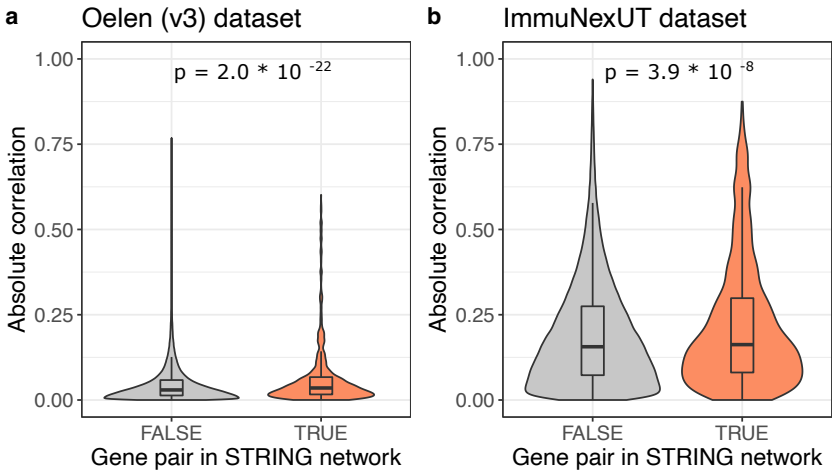


Figure A.27.: **Validation of correlation pattern with the STRING database** Enrichment of correlated genes among gene pairs whose proteins are interacting according to the STRING database, taking correlation values from Oelen v3 single-cell dataset in **a)** and ImmuNexUT bulk dataset in **b)**. P-values in the plot show the significance level of the Wilcoxon rank sum test. Figure and legend taken from [3].

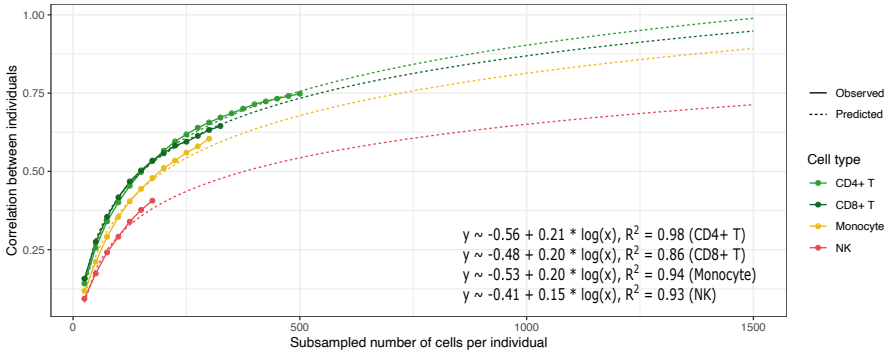


Figure A.28.: **Relationship between number of cells and concordance between donors** Fitting a logarithmic curve (based on the natural logarithm) for the four most frequent cell types (CD4+ T cells, CD8+ T cells, monocytes, NK cells) to explain the correlation value between individuals by the number of cells per individuals (estimated curves and adjusted R2 values for each cell type in the text). The dotted line shows the extrapolation of this fit to predict correlation when increasing the number of cells up to 1,500 cells per individual and cell type. Figure and legend taken from [3].

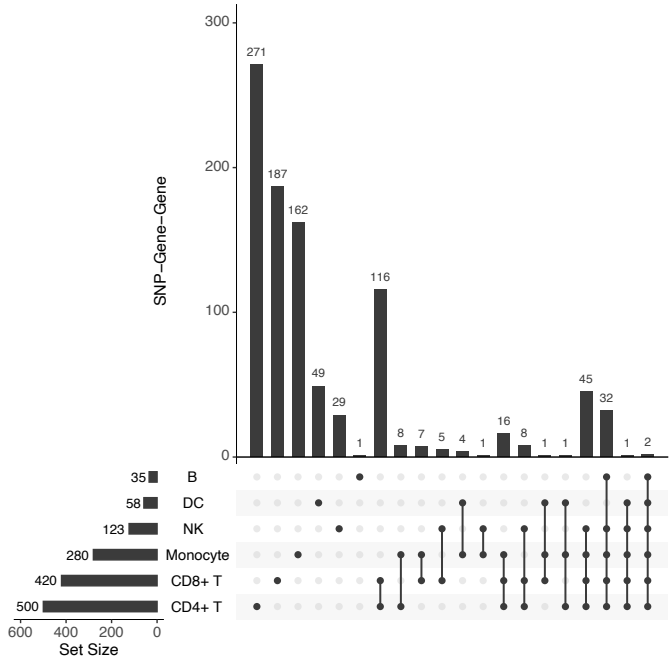


Figure A.29.: **Overlap of significant co-eQTLs between cell types** Number of significant co-eQTLs after filtering and meta-analysis and overlap between all cell types, showing that only a small fraction of co-eQTLs is identified in more than one cell type. Figure and legend taken from [3].

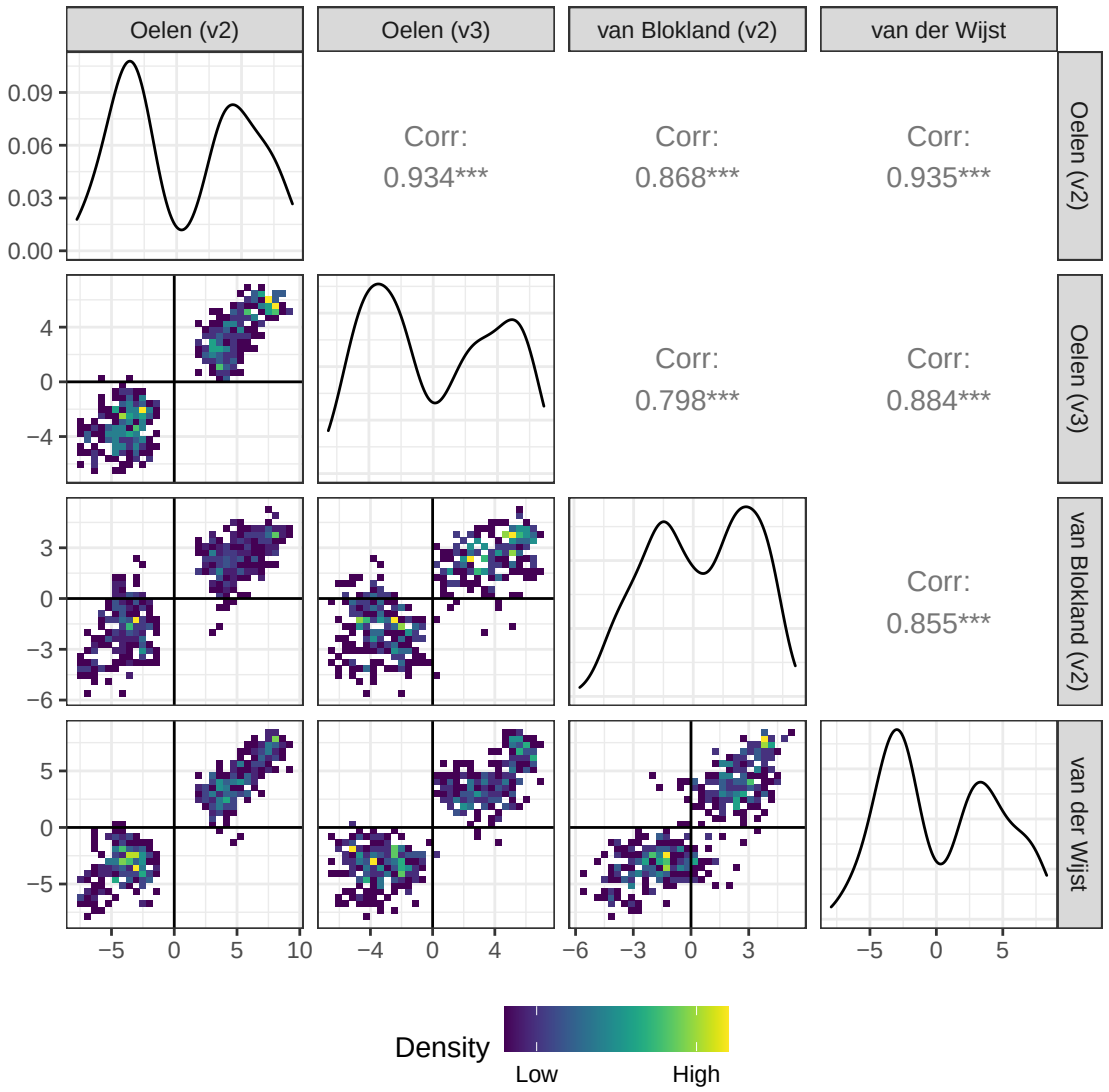


Figure A.30.: **Comparison of Z-scores across datasets within the meta-analysis** Distribution of significant co-eQTL Z-scores per dataset, that was included in the meta-analysis, for the CD4+ T cells. The plot shows scatter density plots between datasets (lower triangle), distributions within each dataset (diagonal) and correlations between datasets (upper triangle). Figure and legend taken from [3].

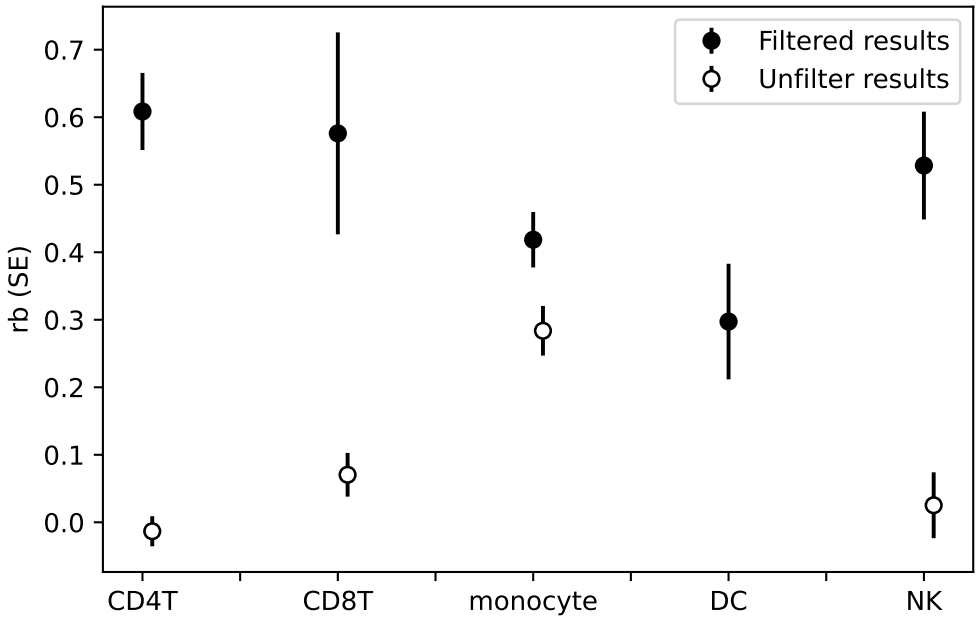


Figure A.31.: **Replication rates for co-eQTls from filtered and unfiltered strategy** Comparison of rb values from BIOS replication analysis between co-eQTLs identified with the filtering strategy and that without the filtering strategy (filled vs unfilled dots). Figure and legend taken from [3].

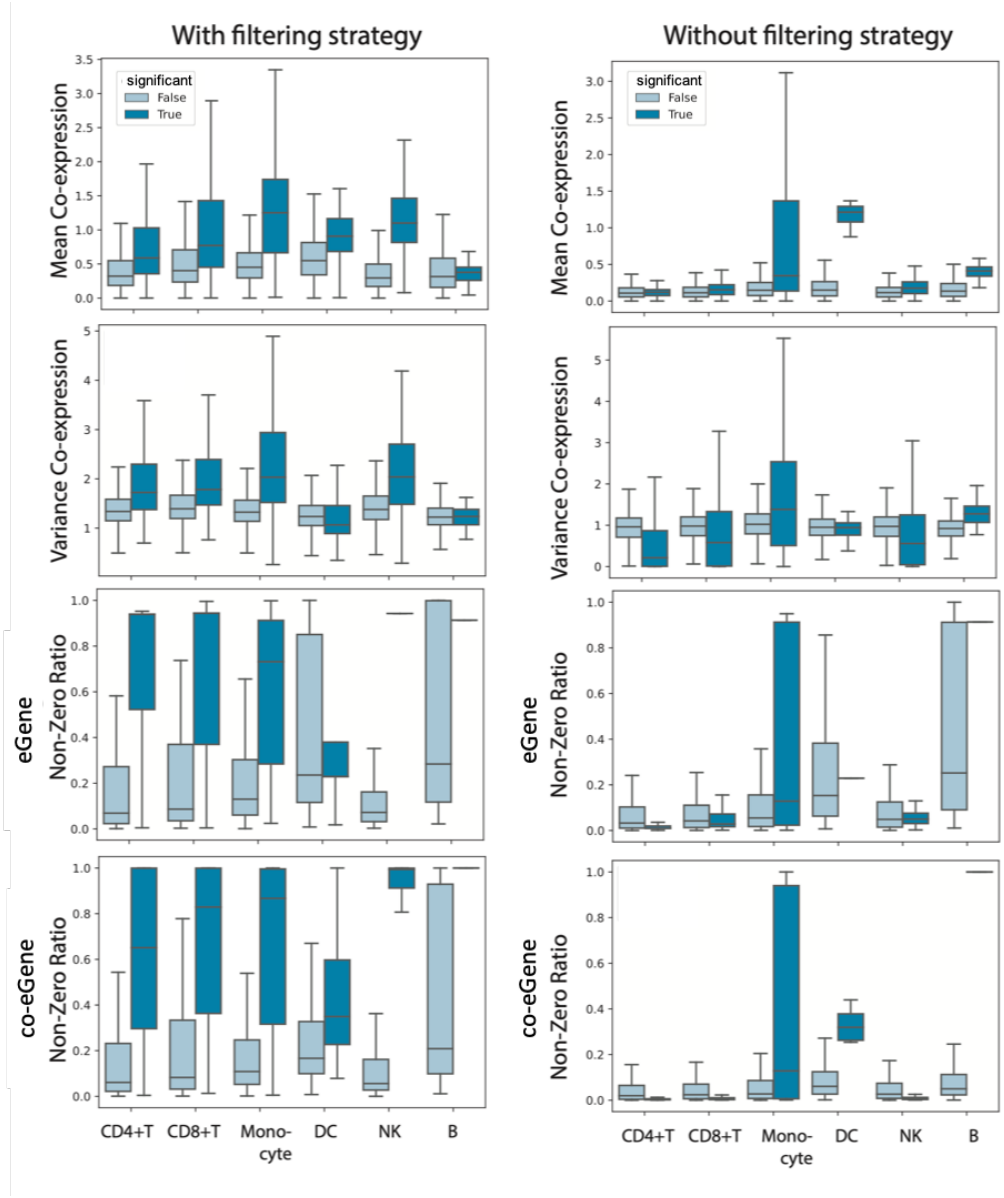


Figure A.32.: **Correlation distribution of co-eQTLs and non-significant triplets** Comparing the correlation distributions of co-eQTLs and non-significant triplets, obtained with approach including the filtering step (**a,c,e,g**) and without the filtering step (**b,d,f,g**). Comparison of co-expression mean (**a,b**), co-expression variance (**c,d**), non-zero rate of the eGenes (**e,f** and of the co-eGenes (**g,h**) between co-eQTLs and non-significant triplets. All analyses done for Oelen v2 dataset and separately for each cell type. Figure and legend taken from [3].

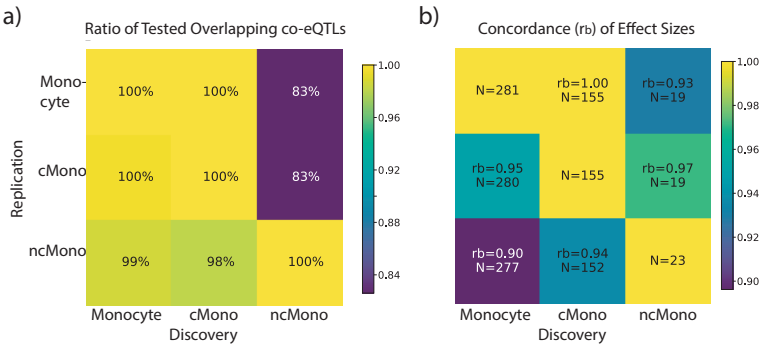


Figure A.33.: **Impact of sub-cell type composition on monocyte co-eQTLs**
Replication rates of co-eQTLs identified in monocytes, classical monocytes (cMono) and non-classical monocytes (ncMono) compared with each other, measuring the ratio of tested co-eQTLs in both (sub)cell types (**a**) and the r_b values for each replication (**b**). Figure and legend taken from [3].

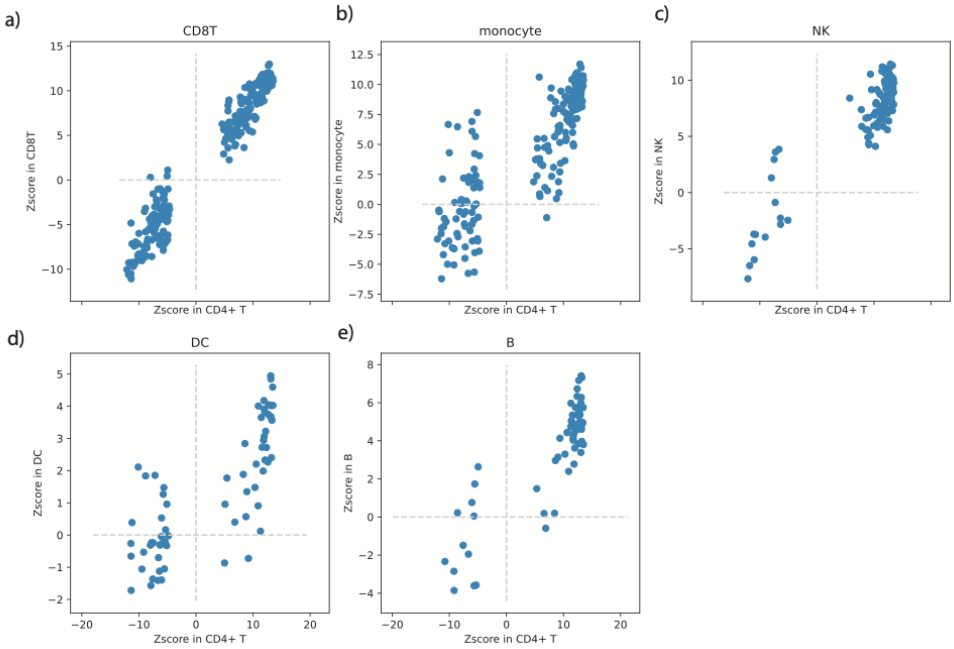


Figure A.34.: **Replication of co-eQTLs associated rs1131017–*RPS26* in CD4+ T cells in other cell types**
Effect sizes from rs1131017–*RPS26* co-eQTLs that were significant in CD4+ T cells were compared across cell types. Each panel shows the replication performance in the corresponding cell type as indicated in the panel titles. Figure and legend taken from [3].

List of Figures

List of Tables

Bibliography

[1] K. T. Schmid, B. Höllbacher, C. Cruceanu, et al. "scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies". In: *Nature Communications* 12.1 (2021), pp. 1–18. ISSN: 20411723. DOI: 10.1038/s41467-021-26779-7.

[2] J. S. Hawe, R. Wilson, K. T. Schmid, et al. "Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function". In: *Nature Genetics* (2022). ISSN: 1061-4036. DOI: 10.1038/s41588-021-00969-x.

[3] S. Li, K. T. Schmid, D. D. Vries, et al. "Identification of genetic variants that impact gene co- expression relationships using large-scale single-cell data". In: *bioRxiv* (2022).

[4] M. Loh, W. Zhang, H. K. Ng, et al. "Identification of genetic effects underlying type 2 diabetes in South Asian and European populations". In: *Communications Biology* 5.1 (2022), pp. 1–10. DOI: 10.1038/s42003-022-03248-5.

[5] M. Claussnitzer, J. H. Cho, R. Collins, et al. "A brief history of human disease genetics". In: *Nature* 577.7789 (2020), pp. 179–189. ISSN: 14764687. DOI: 10.1038/s41586-019-1879-7.

[6] A. Buniello, J. A. Macarthur, M. Cerezo, et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic Acids Research* 47.D1 (2019), pp. D1005–D1012. ISSN: 13624962. DOI: 10.1093/nar/gky1120.

[7] T. Lappalainen and D. G. MacArthur. "From variant to function in human disease genetics". In: *Science* 373.6562 (2021), pp. 1464–1468. ISSN: 10959203. DOI: 10.1126/science.abi8207.

[8] Y. Hasin, M. Seldin, and A. Lusis. "Multi-omics approaches to disease". In: *Genome Biology* 18.1 (2017), pp. 1–15. ISSN: 1474760X. DOI: 10.1186/s13059-017-1215-1.

[9] F. Abascal, R. Acosta, N. J. Addleman, et al. "Expanded encyclopaedias of DNA elements in the human and mouse genomes". In: *Nature* 583.7818 (2020), pp. 699–710. ISSN: 14764687. DOI: 10.1038/s41586-020-2493-4.

[10] B. Kremer, P. Goldberg, S. E. Andrew, et al. "A Worldwide Study of the Huntington's Disease Mutation: The Sensitivity and Specificity of Measuring CAG Repeats". In: *New England Journal of Medicine* 330.20 (May 1994), pp. 1401–1406. ISSN: 0028-4793. DOI: 10.1056/NEJM199405193302001. URL: http://www.nejm.org/doi/abs/10.1056/NEJM199405193302001.

[11] J. E. Posey, A. H. O'Donnell-Luria, J. X. Chong, et al. "Insights into genetics, human biology and disease gleaned from family based genomic studies". In: *Genetics in Medicine* 21.4 (2019), pp. 798–812. ISSN: 15300366. DOI: 10.1038/s41436-018-0408-7.

[12] D. Altshuler, M. J. Daly, and E. S. Lander. "Genetic mapping in human disease". In: *Science* 322.5903 (2008), pp. 881–888. ISSN: 00368075. DOI: 10.1126/science.1156409.

[13] L. Yengo, J. Sidorenko, K. E. Kemper, et al. "Meta-analysis of genome-wide association studies for height and body mass index in 700 000 individuals of European ancestry". In: *Human Molecular Genetics* 27.20 (2018), pp. 3641–3649. ISSN: 14602083. DOI: 10.1093/hmg/ddy271.

[14] J. J. Lee, R. Wedow, A. Okbay, et al. "Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals". In: *Nature Genetics* 50.8 (2018), pp. 1112–1121. ISSN: 15461718. DOI: 10.1038/s41588-018-0147-3.

[15] J. Zeng, A. Xue, L. Jiang, et al. "Widespread signatures of natural selection across human complex traits and functional genomic categories". In: *Nature Communications* 12.1 (2021), pp. 1–12. ISSN: 20411723. DOI: 10.1038/s41467-021-21446-3. URL: http://dx.doi.org/10.1038/s41467-021-21446-3.

[16] E. A. Boyle, Y. I. Li, and J. K. Pritchard. "An Expanded View of Complex Traits: From Polygenic to Omnigenic". In: *Cell* 169.7 (2017), pp. 1177–1186. ISSN: 10974172. DOI: 10.1016/j.cell.2017.05.038. URL: http://dx.doi.org/10.1016/j.cell.2017.05.038.

[17] R. A. Fisher. "The Correlation between Relatives on the Supposition of Mendelian Inheritance." In: *Transactions of the Royal Society of Edinburgh* 52.2 (July 1919), pp. 399–433. ISSN: 0080-4568. DOI: 10.1017/S0080456800012163. URL: https://www.cambridge.org/core/product/identifier/S0080456800012163/type/journal_article.

[18] J. W. Belmont, A. Boudreau, S. M. Leal, et al. "A haplotype map of the human genome". In: *Nature* 437.7063 (2005), pp. 1299–1320. ISSN: 00280836. DOI: 10.1038/nature04226.

[19] D. J. Schaid, W. Chen, and N. B. Larson. "From genome-wide associations to candidate causal variants by statistical fine-mapping". In: *Nature Reviews Genetics* 19.8 (2018), pp. 491–504. ISSN: 14710064. DOI: 10.1038/s41576-018-0016-z. URL: http://dx.doi.org/10.1038/s41576-018-0016-z.

[20] EMBL-EBI. *GWAS Catalog*. 2022. URL: https://www.ebi.ac.uk/gwas/home.

[21] M. T. Maurano, R. Humbert, E. Rynes, et al. "Systematic localization of common disease-associated variation in regulatory DNA". In: *Science* 337.6099 (2012), pp. 1190–1195. ISSN: 10959203. DOI: 10.1126/science.1222794.

[22] K. K. H. Farh, A. Marson, J. Zhu, et al. "Genetic and epigenetic fine mapping of causal autoimmune disease variants". In: *Nature* 518.7539 (2015), pp. 337–343. ISSN: 14764687. DOI: 10.1038/nature13835.

[23] M. Tsompana and M. J. Buck. "Chromatin accessibility: A window into the genome". In: *Epigenetics and Chromatin* 7.1 (2014), pp. 1–16. ISSN: 17568935. DOI: 10.1186/1756-8935-7-33.

[24] Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (2015), pp. 317–329. ISSN: 14764687. DOI: 10.1038/nature14248.

[25] V. K. Rakyan, T. A. Down, D. J. Balding, et al. "Epigenome-wide association studies for common human diseases". In: *Nature Reviews Genetics* 12.8 (2011), pp. 529–541. ISSN: 14710056. DOI: 10.1038/nrg3000.

[26] G. Millán-Zambrano, A. Burton, A. J. Bannister, et al. "Histone post-translational modifications — cause and consequence of genome function". In: *Nature Reviews Genetics* (2022). ISSN: 14710064. DOI: 10.1038/s41576-022-00468-7.

[27] J. Mill and B. T. Heijmans. "From promises to practical strategies in epigenetic epidemiology". In: *Nature Reviews Genetics* 14.8 (2013), pp. 585–594. ISSN: 14710056. DOI: 10.1038/nrg3405. URL: http://dx.doi.org/10.1038/nrg3405.

[28] C. G. Bell, R. Lowe, P. D. Adams, et al. "DNA methylation aging clocks: Challenges and recommendations". In: *Genome Biology* 20.1 (2019), pp. 1–24. ISSN: 1474760X. DOI: 10.1186/s13059-019-1824-y.

[29] T. Battram, P. Yousefi, G. Crawford, et al. "The EWAS Catalog: a database of epigenome-wide association studies". In: *Wellcome Open Research* 7 (2022), p. 41. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.17598.1.

[30] T. Lappalainen, M. Sammeth, M. R. Friedländer, et al. "Transcriptome and genome sequencing uncovers functional variation in humans". In: *Nature* 501.7468 (2013), pp. 506–511. ISSN: 00280836. DOI: 10.1038/nature12531.

[31] R. Joehanes, X. Zhang, T. Huan, et al. "Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies". In: *Genome Biology* 18.1 (2017), pp. 1–24. ISSN: 1474760X. DOI: 10.1186/s13059-016-1142-6. URL: http://dx.doi.org/10.1186/s13059-016-1142-6.

[32] F. Aguet, A. N. Barbeira, R. Bonazzola, et al. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". In: *Science* 369.6509 (Sept. 2020), pp. 1318–1330. ISSN: 0036-8075. DOI: 10.1126/science.aaz1776. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.aaz1776.

[33] U. Võsa, A. Claringbould, H. J. Westra, et al. "Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression". In: *Nature genetics* 53.9 (2021), pp. 1300–1310. ISSN: 15461718. DOI: 10.1038/s41588-021-00913-z.

[34] J. F. Degner, A. A. Pai, R. Pique-Regi, et al. "DNase I sensitivity QTLs are a major determinant of human expression variation". In: *Nature* 482.7385 (Feb. 2012), pp. 390–394. ISSN: 0028-0836. DOI: 10.1038/nature10808. URL: http://www.nature.com/articles/nature10808.

[35] I. Dunham, A. Kundaje, S. F. Aldred, et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. ISSN: 0028-0836. DOI: 10.1038/nature11247. URL: http://www.nature.com/articles/nature11247.

[36] M. Uhlén, L. Fagerberg, B. M. Hallström, et al. "Tissue-based map of the human proteome". In: *Science* 347.6220 (2015). ISSN: 10959203. DOI: 10.1126/science.1260419.

[37] I. Assum, J. Krause, M. O. Scheinhardt, et al. "Tissue-specific multi-omics analysis of atrial fibrillation". In: *Nature Communications* 13.1 (2022), pp. 1–15. ISSN: 20411723. DOI: 10.1038/s41467-022-27953-1.

[38] M. J. Bonder, R. Luijk, D. V. Zhernakova, et al. "Disease variants alter transcription factor levels and methylation of their binding sites". In: *Nature Genetics* 49.1 (2017), pp. 131–138. ISSN: 15461718. DOI: 10.1038/ng.3721.

[39] G. Nicholson, M. Rantalainen, J. V. Li, et al. "A genome-wide metabolic QTL analysis in europeans implicates two Loci shaped by recent positive selection". In: *PLoS Genetics* 7.9 (2011). ISSN: 15537390. DOI: 10.1371/journal.pgen.1002270.

[40] S. E. Pierce, A. Booms, J. Prahl, et al. "Post-GWAS knowledge gap: the how, where, and when". In: *npj Parkinson's Disease* 6.1 (2020), pp. 1–5. ISSN: 23738057. DOI: 10.1038/s41531-020-00125-y. URL: http://dx.doi.org/10.1038/s41531-020-00125-y.

[41] S. Kim-Hellmuth, F. Aguet, M. Oliva, et al. "Cell type-specific genetic regulation of gene expression across human tissues". In: *Science (New York, N.Y.)* 369.6509 (2020). ISSN: 10959203. DOI: 10.1126/science.aaz8528.

[42] H. J. Westra, D. Arends, T. Esko, et al. "Cell Specific eQTL Analysis without Sorting Cells". In: *PLoS Genetics* 11.5 (2015), pp. 1–17. ISSN: 15537404. DOI: 10.1371/journal.pgen.1005223.

[43] L. Chen, B. Ge, F. P. Casale, et al. "Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells Resource Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells". In: *Cell* 167.5 (2016), pp. 1398–1414. DOI: 10.1016/j.cell.2016.10.026.

[44] M. Ota, Y. Nagafuchi, H. Hatano, et al. "Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases". In: *Cell* 184.11 (2021), pp. 3006–3021. ISSN: 10974172. DOI: 10.1016/j.cell.2021.03.056. URL: https://doi.org/10.1016/j.cell.2021.03.056.

[45] C. E. Romanoski, S. Lee, M. J. Kim, et al. "Systems Genetics Analysis of Gene-by-Environment Interactions in Human Cells". In: *American Journal of Human Genetics* 86.3 (2010), pp. 399–410. ISSN: 00029297. DOI: 10.1016/j.ajhg.2010.02.002. URL: http://dx.doi.org/10.1016/j.ajhg.2010.02.002.

[46] D. V. Zhernakova, P. Deelen, M. Vermaat, et al. "Identification of context-dependent expression quantitative trait loci in whole blood". In: *Nature Genetics* 49.1 (2017), pp. 139–145. ISSN: 15461718. DOI: 10.1038/ng.3737.

[47] B. J. Strober, R. Elorbany, K. Rhodes, et al. "Dynamic genetic regulation of gene expression during cellular differentiation". In: *Science* 364.6447 (2019), pp. 1287–1290. ISSN: 10959203. DOI: 10.1126/science.aaw0040.

[48] E. A. Houseman, W. P. Accomando, D. C. Koestler, et al. "DNA methylation arrays as surrogate measures of cell mixture distribution". In: *BMC Bioinformatics* 13.1 (2012). ISSN: 14712105. DOI: 10.1186/1471-2105-13-86.

[49] "Method of the Year 2013". In: *Nature Methods* 11.1 (Jan. 2014), pp. 1–1. ISSN: 1548-7091. DOI: 10.1038/nmeth.2801. URL: http://www.nature.com/articles/nmeth.2801.

[50] A. Bastidas-Ponce, S. Tritschler, L. Dony, et al. "Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis". In: *Development (Cambridge)* 146.12 (2019). ISSN: 14779129. DOI: 10.1242/dev.173849.

[51] R. C. Tyser, E. Mahammadov, S. Nakanoh, et al. "Single-cell transcriptomic characterization of a gastrulating human embryo". In: *Nature* 600.7888 (2021), pp. 285–289. ISSN: 14764687. DOI: 10.1038/s41586-021-04158-y.

[52] A. P. Patel, I. Tirosh, J. J. Trombetta, et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma". In: *Science* 344.6190 (June 2014), pp. 1396–1401. ISSN: 0036-8075. DOI: 10.1126/science.1254257. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf%20https://www.science.org/doi/10.1126/science.1254257.

[53] T. Wu, E. Hu, S. Xu, et al. "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data". In: *The Innovation* 2.3 (2021), p. 100141. ISSN: 26666758. DOI: 10.1016/j.xinn.2021.100141. URL: http://dx.doi.org/10.1016/j.xinn.2021.100141.

[54] P. Bischoff, A. Trinks, B. Obermayer, et al. "Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma". In: *Oncogene* 40.50 (2021), pp. 6748–6758. ISSN: 14765594. DOI: 10.1038/s41388-021-02054-3.

[55] M. G. Van Der Wijst, H. Brugge, D. H. De Vries, et al. "Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs". In: *Nature Genetics* 50.4 (2018), pp. 493–497. ISSN: 15461718. DOI: 10.1038/s41588-018-0089-9. URL: http://dx.doi.org/10.1038/s41588-018-0089-9.

[56] A. S. Cuomo, D. D. Seaton, D. J. McCarthy, et al. "Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression". In: *Nature Communications* 11.1 (2020), pp. 1–14. ISSN: 20411723. DOI: 10.1038/s41467-020-14457-z.

[57] S. Yazar, J. Alquicira-Hernandez, K. Wing, et al. "Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease". In: *Science* 376.6589 (2022). ISSN: 0036-8075. DOI: 10.1126/science.abf3041.

[58] R. K. Perez, M. G. Gordon, M. Subramaniam, et al. "Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus." In: *Science (New York, N.Y.)* 376.6589 (2022), eabf1970. ISSN: 1095-9203. DOI: 10.1126/science.abf1970. URL: http://www.ncbi.nlm.nih.gov/pubmed/35389781.

[59] F. Tang, C. Barbacioru, Y. Wang, et al. "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nature Methods* 6.5 (2009), pp. 377–382. ISSN: 15487091. DOI: 10.1038/nmeth.1315.

[60] C. Ziegenhain, B. Vieth, S. Parekh, et al. "Comparative Analysis of Single-Cell RNA Sequencing Methods". In: *Molecular Cell* 65.4 (2017), pp. 631–643. ISSN: 10974164. DOI: 10.1016/j.molcel.2017.01.023.

[61] E. Mereu, A. Lafzi, C. Moutinho, et al. "Benchmarking single-cell RNA-sequencing protocols for cell atlas projects". In: *Nature Biotechnology* 38.6 (2020), pp. 747–755. ISSN: 15461696. DOI: 10.1038/s41587-020-0469-4. URL: http://dx.doi.org/10.1038/s41587-020-0469-4.

[62] A. Lafzi, C. Moutinho, S. Picelli, et al. "Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies". In: *Nature Protocols* 13.12 (2018), pp. 2742–2757. ISSN: 17502799. DOI: 10.1038/s41596-018-0073-y.

[63] D. Ramsköld, S. Luo, Y. C. Wang, et al. "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells". In: *Nature Biotechnology* 30.8 (2012), pp. 777–782. ISSN: 15461696. DOI: 10.1038/nbt.2282.

[64] S. Picelli, Å. K. Björklund, O. R. Faridani, et al. "Smart-seq2 for sensitive full-length transcriptome profiling in single cells". In: *Nature Methods* 10.11 (2013), pp. 1096–1100. ISSN: 15487091. DOI: 10.1038/nmeth.2639.

[65] M. Hagemann-Jensen, C. Ziegenhain, P. Chen, et al. "Single-cell RNA counting at allele and isoform resolution using Smart-seq3". In: *Nature Biotechnology* 38.6 (2020), pp. 708–714. ISSN: 15461696. DOI: 10.1038/s41587-020-0497-0. URL: http://dx.doi.org/10.1038/s41587-020-0497-0.

[66] T. Hashimshony, F. Wagner, N. Sher, et al. "CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification". In: *Cell Reports* 2.3 (2012), pp. 666–673. ISSN: 22111247. DOI: 10.1016/j.celrep.2012.08.003. URL: http://dx.doi.org/10.1016/j.celrep.2012.08.003.

[67] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, et al. "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types". In: *Science* 343.6172 (Feb. 2014), pp. 776–779. ISSN: 0036-8075. DOI: 10.1126/science.1247651. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.1247651.

[68] E. Z. Macosko, A. Basu, R. Satija, et al. "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets". In: *Cell* 161.5 (2015), pp. 1202–1214. ISSN: 10974172. DOI: 10.1016/j.cell.2015.05.002. URL: http://dx.doi.org/10.1016/j.cell.2015.05.002.

[69] A. M. Klein, L. Mazutis, I. Akartuna, et al. "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells". In: *Cell* 161.5 (2015), pp. 1187–1201. ISSN: 10974172. DOI: 10.1016/j.cell.2015.04.044. URL: http://dx.doi.org/10.1016/j.cell.2015.04.044.

[70] G. X. Zheng, J. M. Terry, P. Belgrader, et al. "Massively parallel digital transcriptional profiling of single cells". In: *Nature Communications* 8 (2017), pp. 1–12. ISSN: 20411723. DOI: 10.1038/ncomms14049. URL: http://dx.doi.org/10.1038/ncomms14049.

[71] A. Regev, S. Teichmann, E. Lander, et al. "Science Forum: The Human Cell Atlas". In: *eLife* (2017), pp. 1–30. ISSN: 2050-084X.

[72] F. A. Vieira Braga, G. Kar, M. Berg, et al. "A cellular census of human lungs identifies novel cell states in health and in asthma". In: *Nature Medicine* 25.7 (2019), pp. 1153–1163. ISSN: 1546170X. DOI: 10.1038/s41591-019-0468-5.

[73] M. Litviňuková, C. Talavera-López, H. Maatz, et al. "Cells of the adult human heart". In: *Nature* (2020), pp. 1–10. ISSN: 0028-0836. DOI: 10.1038/s41586-020-2797-4. URL: http://www.nature.com/articles/s41586-020-2797-4.

[74] V. Svensson, E. da Veiga Beltrame, and L. Pachter. "A curated database reveals trends in single-cell transcriptomics". In: *Database* 2020 (2020), pp. 1–7. ISSN: 17580463. DOI: 10.1093/DATABASE/BAAA073.

[75] L. Zappia and F. J. Theis. "Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape". In: *Genome Biology* 22.1 (2021), pp. 1–18. ISSN: 1474760X. DOI: 10.1186/s13059-021-02519-4.

[76] D. A. Cusanovich, R. Daza, A. Adey, et al. "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing". In: *Science* 348.6237 (2015), pp. 910–914. ISSN: 10959203. DOI: 10.1126/science.aab1601.

[77] J. D. Buenrostro, B. Wu, U. M. Litzenburger, et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". In: *Nature* 523.7561 (2015), pp. 486–490. ISSN: 14764687. DOI: 10.1038/nature14590.

[78] S. A. Smallwood, H. J. Lee, C. Angermueller, et al. "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity". In: *Nature Methods* 11.8 (2014), pp. 817–820. ISSN: 15487105. DOI: 10.1038/nmeth.3035.

[79] I. C. Macaulay, W. Haerty, P. Kumar, et al. "G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes". In: *Nature Methods* 12.6 (2015), pp. 519–522. ISSN: 15487105. DOI: 10.1038/nmeth.3370.

[80] Y. Hou, H. Guo, C. Cao, et al. "Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas". In: *Cell Research* 26.3 (2016), pp. 304–319. ISSN: 17487838. DOI: 10.1038/cr.2016.23.

[81] A. P. Frei, F. A. Bava, E. R. Zunder, et al. "Highly multiplexed simultaneous detection of RNAs and proteins in single cells". In: *Nature Methods* 13.3 (2016), pp. 269–275. ISSN: 15487105. DOI: 10.1038/nmeth.3742.

[82] S. Chen, B. B. Lake, and K. Zhang. "High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell". In: *Nature Biotechnology* 37.12 (2019), pp. 1452–1457. ISSN: 15461696. DOI: 10.1038/s41587-019-0290-0. URL: http://dx.doi.org/10.1038/s41587-019-0290-0.

[83] R. Oelen, D. H. de Vries, H. Brugge, et al. "Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure". In: *Nature Communications* 13.1 (Dec. 2022), p. 3267. ISSN: 2041-1723. DOI: 10.1038/s41467-022-30893-5. URL: https://www.nature.com/articles/s41467-022-30893-5.

[84] A. Nathan, S. Asgari, K. Ishigaki, et al. "Modeling memory T cell states at single-cell resolution identifies in vivo state-dependence of eQTLs influencing disease". In: *bioRxiv* (2021), p. 2021.07.29.454316. URL: https://www.biorxiv.org/content/10.1101/2021.07.29.454316v1%0Ahttps://www.biorxiv.org/content/10.1101/2021.07.29.454316v1.abstract.

[85] A. S. Cuomo, T. Heinen, D. Vagiaki, et al. "CellRegMap: A statistical framework for mapping context-specific regulatory variants using scRNA-seq". In: *bioRxiv* (2021), p. 2021.09.01.458524. URL: https://www.biorxiv.org/content/10.1101/2021.09.01.458524v1%0Ahttps://www.biorxiv.org/content/10.1101/2021.09.01.458524v1.abstract.

[86] H. Harikumar, T. P. Quinn, S. Rana, et al. "Personalized single-cell networks: a framework to predict the response of any gene to any drug for any patient". In: *BioData Mining* 14.1 (2021), pp. 1–15. ISSN: 17560381. DOI: 10.1186/s13040-021-00263-w.

[87] A. Pratapa, A. P. Jalihal, J. N. Law, et al. "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data". In: *Nature Methods* 17.2 (2020), pp. 147–154. ISSN: 15487105. DOI: 10.1038/s41592-019-0690-6. URL: http://dx.doi.org/10.1038/s41592-019-0690-6.

[88] M. G. van der Wijst, D. H. de Vries, H. E. Groot, et al. "The single-cell eQTLGen consortium". In: *eLife* 9 (2020), pp. 1–21. ISSN: 2050084X. DOI: 10.7554/eLife.52155.

[89] A. S. Cuomo, G. Alvari, C. B. Azodi, et al. "Optimizing expression quantitative trait locus mapping workflows for single-cell studies". In: *Genome Biology* 22.1 (2021), pp. 1–30. ISSN: 1474760X. DOI: 10.1186/s13059-021-02407-x.

[90] V. Svensson. "Droplet scRNA-seq is not zero-inflated". In: *Nature Biotechnology* 38.2 (Feb. 2020), pp. 147–150. ISSN: 1087-0156. DOI: 10.1038/s41587-019-0379-5. URL: http://www.nature.com/articles/s41587-019-0379-5.

[91] W. Chen, Y. Li, J. Easton, et al. "UMI-count modeling and differential expression analysis for single-cell RNA sequencing". In: *Genome Biology* 19.1 (2018), pp. 1–17. ISSN: 1474760X. DOI: 10.1186/s13059-018-1438-9.

[92] B. Servin and M. Stephens. "Imputation-based analysis of association studies: Candidate regions and quantitative traits". In: *PLoS Genetics* 3.7 (2007), pp. 1296–1308. ISSN: 15537390. DOI: 10.1371/journal.pgen.0030114.

[93] J. T. Leek and J. D. Storey. "Capturing heterogeneity in gene expression studies by surrogate variable analysis". In: *PLoS Genetics* 3.9 (2007), pp. 1724–1735. ISSN: 15537390. DOI: 10.1371/journal.pgen.0030161.

[94] R. Moore, F. P. Casale, M. Jan Bonder, et al. "A linear mixed-model approach to study multivariate gene–environment interactions". In: *Nature Genetics* 51.1 (2019), pp. 180–186. ISSN: 15461718. DOI: 10.1038/s41588-018-0271-0.

[95] A. A. Shabalin. "Matrix eQTL: Ultra fast eQTL analysis via large matrix operations". In: *Bioinformatics* 28.10 (2012), pp. 1353–1358. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts163.

[96] J. J. Faraway. *Linear Models with R*. Chapman and Hall/CRC, Aug. 2004, pp. 1–221. ISBN: 9781135437664. DOI: 10.4324/9780203507278. URL: https://www.taylorfrancis.com/books/9781135437664.

[97] T. Fawcett. "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers ROC Graphs : Notes and Practical Considerations for Data Mining Researchers". In: *HP Invent* (2003), p. 27. ISSN: 08997667. DOI: 10.1.1.10.9777.

[98] R. E. Peterson, K. Kuchenbaecker, R. K. Walters, et al. "Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations". In: *Cell* 179.3 (2019), pp. 589–603. ISSN: 10974172. DOI: 10.1016/j.cell.2019.08.051. URL: https://doi.org/10.1016/j.cell.2019.08.051.

[99] E. Zeggini and J. P. A. Ioannidis. "Europe PMC Funders Group Meta-analysis in genome-wide association studies". In: *Pharmacogenomics* 10.2 (2009), pp. 191–201. DOI: 10.2217/14622416.10.2.191.Meta-analysis.

[100] K. L. Lunetta. "Methods for meta-analysis of genetic data". In: *Current Protocols in Human Genetics* SUPPL.77 (2013), pp. 1–8. ISSN: 19348266. DOI: 10.1002/0471142905.hg0124s77.

[101] M. Civelek and A. J. Lusis. "Systems genetics approaches to understand complex traits". In: *Nature Reviews Genetics* 15.1 (2014), pp. 34–48. ISSN: 14710056. DOI: 10.1038/nrg3575.

[102] J. A. Nelder and R. W. M. Wedderburn. "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), p. 370. ISSN: 00359238. DOI: 10.2307/2344614. URL: https://www.jstor.org/stable/10.2307/2344614?origin=crossref.

[103] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2009), pp. 139–140. ISSN: 14602059. DOI: 10.1093/bioinformatics/btp616.

[104] S. Anders and W. Huber. "Differential expression analysis for sequence count data". In: *Genome Biology* 11.10 (Oct. 2010), R106. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106. URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106.

[105] M. I. Love, W. Huber, and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (2014), pp. 1–21. ISSN: 1474760X. DOI: 10.1186/s13059-014-0550-8.

[106] A. Farcomeni. "A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion". In: *Statistical Methods in Medical Research* 17.4 (2008), pp. 347–388. ISSN: 09622802. DOI: 10.1177/0962280206079046.

[107] Y. Benjamini and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (Jan. 1995), pp. 289–300. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1995.tb02031.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1995.tb02031.x.

[108] P. H. Westfall and S. S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, 1993.

[109] H. Ongen, A. Buil, A. A. Brown, et al. "Fast and efficient QTL mapper for thousands of molecular phenotypes". In: *Bioinformatics* 32.10 (2016), pp. 1479–1485. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv722.

[110] M. Jones. "Kumaraswamy's distribution: A beta-type distribution with some tractability advantages". In: *Statistical Methodology* 6.1 (Jan. 2009), pp. 70–81. ISSN: 15723127. DOI: 10.1016/j.stamet.2008.04.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S1572312708000282.

[111] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, May 2013. ISBN: 9781134742707. DOI: 10.4324/9780203771587. URL: https://www.taylorfrancis.com/books/9781134742707.

[112] B. Vieth, C. Ziegenhain, S. Parekh, et al. "powsimR: power analysis for bulk and single cell RNA-seq experiments". In: *Bioinformatics (Oxford, England)* 33.21 (2017), pp. 3486–3488. ISSN: 13674811. DOI: 10.1093/bioinformatics/btx435.

[113] H. L. Crowell, C. Soneson, P. L. Germain, et al. "Muscat Detects Subpopulation-Specific State Transitions From Multi-Sample Multi-Condition Single-Cell Transcriptomics Data". In: *Nature Communications* 11.1 (2020), pp. 1–12. ISSN: 20411723. DOI: 10.1038/s41467-020-19894-4. URL: http://dx.doi.org/10.1038/s41467-020-19894-4.

[114] J. W. Squair, M. Gautier, C. Kathe, et al. "Confronting false discoveries in single-cell differential expression". In: *Nature Communications* 12.1 (2021). ISSN: 20411723. DOI: 10.1038/s41467-021-25960-2. URL: http://dx.doi.org/10.1038/s41467-021-25960-2.

[115] X. Chen and H. Ishwaran. "Random forests for genomic data analysis". In: *Genomics* 99.6 (2012), pp. 323–329. ISSN: 08887543. DOI: 10.1016/j.ygeno.2012.04.003.

[116] G. Eraslan, Ž. Avsec, J. Gagneur, et al. "Deep learning: new computational modelling techniques for genomics". In: *Nature Reviews Genetics* 20.7 (2019), pp. 389–403. ISSN: 14710064. DOI: 10.1038/s41576-019-0122-6. URL: http://dx.doi.org/10.1038/s41576-019-0122-6.

[117] S. Tangirala. "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm". In: *International Journal of Advanced Computer Science and Applications* 11.2 (2020), pp. 612–619. ISSN: 21565570. DOI: 10.14569/ijacsa.2020.0110277.

[118] A. Liaw and M. Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22.

[119] J. C. Chambers, M. Loh, B. Lehne, et al. "Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: A nested case-control study". In: *The Lancet Diabetes and Endocrinology* 3.7 (2015), pp. 526–534. ISSN: 22138595. DOI: 10.1016/S2213-8587(15)00127-8.

[120] S. Wahl, A. Drong, B. Lehne, et al. "Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity". In: *Nature* 541.7635 (2017), pp. 81–86. ISSN: 14764687. DOI: 10.1038/nature20784.

[121] Y. Zhang, R. Wilson, J. Heiss, et al. "DNA methylation signatures in peripheral blood strongly predict all-cause mortality". In: *Nature Communications* 8 (2017), pp. 1–11. ISSN: 20411723. DOI: 10.1038/ncomms14617.

[122] J. Nicodemus-Johnson, R. A. Myers, N. J. Sakabe, et al. "DNA methylation in lung cells is associated with asthma endotypes and genetic risk". In: *JCI Insight* 1.20 (2016). DOI: 10.1172/jci.insight.90151.

[123] A. Nishiyama and M. Nakanishi. "Navigating the DNA methylation landscape of cancer". In: *Trends in Genetics* 37.11 (2021), pp. 1012–1027. ISSN: 13624555. DOI: 10.1016/j.tig.2021.05.002. URL: https://doi.org/10.1016/j.tig.2021.05.002.

[124] S. C. Maas, M. M. Mens, B. Kühnel, et al. "Smoking-related changes in DNA methylation and gene expression are associated with cardio-metabolic traits". In: *Clinical Epigenetics* 12.1 (2020), pp. 1–16. ISSN: 18687083. DOI: 10.1186/s13148-020-00951-0. URL: https://doi.org/10.1186/s13148-020-00951-0.

[125] J. Nicodemus-Johnson and R. A. Sinnott. "Fruit and juice epigenetic signatures are associated with independent immunoregulatory pathways". In: *Nutrients* 9.7 (2017). ISSN: 20726643. DOI: 10.3390/nu9070752.

[126] D. Schübeler. "Function and information content of DNA methylation". In: *Nature* 517.7534 (Jan. 2015), pp. 321–326. ISSN: 14764687. DOI: 10.1038/nature14192.

[127] C. Schmidl, M. Klug, T. J. Boeld, et al. "Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity". In: *Genome Research* 19.7 (2009), pp. 1165–1174. ISSN: 10889051. DOI: 10.1101/gr.091470.109.

[128] M. J. Ziller, H. Gu, F. Müller, et al. "Charting a dynamic DNA methylation landscape of the human genome Michael". In: 500.7463 (2014), pp. 477–481. DOI: 10.1038/nature12433.Charting.

[129] A. Hellman and A. Chess. "Gene body-specific methylation on the active X chromosome". In: *Science* 315.5815 (2007), pp. 1141–1143. ISSN: 00368075. DOI: 10.1126/science.1136352.

[130] P. A. Jones. "Functions of DNA methylation: Islands, start sites, gene bodies and beyond". In: *Nature Reviews Genetics* 13.7 (2012), pp. 484–492. ISSN: 14710056. DOI: 10.1038/nrg3230. URL: http://dx.doi.org/10.1038/nrg3230.

[131] R. Joehanes, A. C. Just, R. E. Marioni, et al. "Epigenetic Signatures of Cigarette Smoking". In: *Circulation: Cardiovascular Genetics* 9.5 (Oct. 2016), pp. 436–447. ISSN: 1942-325X. DOI: 10.1161/CIRCGENETICS.116.001506. URL: https://www.ahajournals.org/doi/10.1161/CIRCGENETICS.116.001506.

[132] W. Guan, B. T. Steffen, R. N. Lemaitre, et al. "Genome-Wide Association Study of Plasma N6 Polyunsaturated Fatty Acids Within the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium". In: *Circulation: Cardiovascular Genetics* 7.3 (June 2014), pp. 321–331. ISSN: 1942-325X. DOI: 10.1161/CIRCGENETICS.113.000208. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf%20https://www.ahajournals.org/doi/10.1161/CIRCGENETICS.113.000208.

[133] S. Y. Shin, E. B. Fauman, A. K. Petersen, et al. "An atlas of genetic influences on human blood metabolites". In: *Nature Genetics* 46.6 (2014), pp. 543–550. ISSN: 15461718. DOI: 10.1038/ng.2982.

[134] W. J. Astle, H. Elding, T. Jiang, et al. "The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease". In: *Cell* 167.5 (2016), pp. 1415–1429. ISSN: 10974172. DOI: 10.1016/j.cell.2016.10.042.

[135] M. A. Kamat, J. A. Blackshaw, R. Young, et al. "PhenoScanner V2: An expanded tool for searching human genotype-phenotype associations". In: *Bioinformatics* 35.22 (2019), pp. 4851–4853. ISSN: 14602059. DOI: 10.1093/bioinformatics/btz469.

[136] D. Leland Taylor, A. U. Jackson, N. Narisu, et al. "Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle". In: *Proceedings of the National Academy of Sciences of the United States of America* 166.22 (2019), pp. 10883–10888. ISSN: 10916490. DOI: 10.1073/pnas.1814263116.

[137] M. E. Lindholm, F. Marabita, D. Gomez-Cabrero, et al. "An integrative analysis reveals coordinated reprogramming of the epigenome and the transcriptome in human skeletal muscle after training". In: *Epigenetics* 9.12 (2014), pp. 1557–1569. ISSN: 15592308. DOI: 10.4161/15592294.2014.982445.

[138] R. Holle, M. Happich, H. Löwel, et al. "KORA - A Research Platform for Population Based Health Research". In: *Das Gesundheitswesen* 67.S 01 (Aug. 2005), pp. 19–25. ISSN: 0941-3790. DOI: 10.1055/s-2005-858235. URL: http://www.thieme-connect.de/DOI/DOI?10.1055/s-2005-858235.

[139] W. Rathmann, K. Strassburger, M. Heier, et al. "Incidence of Type 2 diabetes in the elderly German population and the effect of clinical and lifestyle risk factors: KORA S4/F4 cohort study". In: *Diabetic Medicine* 26.12 (Dec. 2009), pp. 1212–1219. ISSN: 07423071. DOI: 10.1111/j.1464-5491.2009.02863.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1464-5491.2009.02863.x.

[140] X. Zhou and M. Stephens. "Genome-wide efficient mixed-model analysis for association studies". In: *Nature Genetics* 44.7 (2012), pp. 821–824. ISSN: 10614036. DOI: 10.1038/ng.2310.

[141] A. Taylor-Weiner, F. Aguet, N. J. Haradhvala, et al. "Scaling computational genomics to millions of individuals with GPUs". In: *Genome Biology* 20.1 (Dec. 2019), p. 228. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1836-7. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1836-7.

[142] K. Schramm, C. Marzi, C. Schurmann, et al. "Mapping the genetic architecture of gene regulation in whole blood". In: *PLoS ONE* 9.4 (2014). ISSN: 19326203. DOI: 10.1371/journal.pone.0093844.

[143] R. Tishbirani. *Regression shrinkage and selection via the Lasso*. 1996. URL: https://statweb.stanford.edu/~tibs/lasso/lasso.pdf.

[144] J. Friedman, T. Hastie, and R. Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010). ISSN: 1548-7660. DOI: 10.18637/jss.v033.i01. URL: http://www.jstatsoft.org/v33/i01/.

[145] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. New York: Springer, 2002. ISBN: 0-387-95457-0. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

[146] R. Diaz-Uriarte. "GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest". In: *BMC Bioinformatics* 8.1 (Dec. 2007), p. 328. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-328. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-328.

[147] M. A. Busby, C. Stewart, C. A. Miller, et al. "Scotty: A web tool for designing RNA-Seq experiments to measure differential gene expression". In: *Bioinformatics* 29.5 (2013), pp. 656–657. ISSN: 13674803. DOI: 10.1093/bioinformatics/btt015.

[148] S. N. Hart, T. M. Therneau, Y. Zhang, et al. "Calculating sample size estimates for RNA sequencing data". In: *Journal of Computational Biology* 20.12 (2013), pp. 970–978. ISSN: 10665277. DOI: 10.1089/cmb.2012.0283.

[149] H. Wu, C. Wang, and Z. Wu. "PROPER: Comprehensive power evaluation for differential expression using RNA-seq". In: *Bioinformatics* 31.2 (2015), pp. 233–241. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu640.

[150] C. I. Li and Y. Shyr. "Sample size calculation based on generalized linear models for differential expression analysis in RNA-seq data". In: *Statistical Applications in Genetics and Molecular Biology* 15.6 (2016), pp. 491–505. ISSN: 15446115. DOI: 10.1515/sagmb-2016-0008.

[151] C. B. Azodi, L. Zappia, A. Oshlack, et al. "splatPop: simulating population scale single-cell RNA sequencing data". In: *Genome Biology* 22.1 (2021), pp. 1–16. ISSN: 1474760X. DOI: 10.1186/s13059-021-02546-1.

[152] C. W. Law, Y. Chen, W. Shi, et al. "Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". In: *Genome Biology* 15.2 (2014), pp. 1–17. ISSN: 1474760X. DOI: 10.1186/gb-2014-15-2-r29.

[153] G. Heimberg, R. Bhatnagar, H. El-Samad, et al. "Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing". In: *Cell Systems* 2.4 (2016), pp. 239–250. ISSN: 24054720. DOI: 10.1016/j.cels.2016.04.001. URL: http://dx.doi.org/10.1016/j.cels.2016.04.001.

[154] C. Hafemeister. *How Many Cells*. 2019. URL: https://satijalab.org/howmanycells/.

[155] Bio-Rad. *Bio-Rad. Cell frequencies in common samples - Flow Cytometry analysis*. URL: https://www.bio-rad-antibodies.com/flow-cytometry-cell-%20frequency.html.

[156] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. "Bayesian approach to single-cell differential expression analysis". In: *Nature Methods* 11.7 (2014), pp. 740–742. ISSN: 15487105. DOI: 10.1038/nmeth.2967.

[157] T. M. Brückl, V. I. Spoormaker, P. G. Sämann, et al. "The biological classification of mental disorders (BeCOME) study: A protocol for an observational deep-phenotyping study for the identification of biological subtypes". In: *BMC Psychiatry* 20.1 (2020), pp. 1–25. ISSN: 1471244X. DOI: 10.1186/s12888-020-02541-z.

[158] G. Monaco, B. Lee, W. Xu, et al. "RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types". In: *Cell Reports* 26.6 (2019), pp. 1627–1640. ISSN: 22111247. DOI: 10.1016/j.celrep.2019.01.041.

[159] H. M. Kang, M. Subramaniam, S. Targ, et al. "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation". In: *Nature Biotechnology* 36.1 (2018), pp. 89–94. ISSN: 15461696. DOI: 10.1038/nbt.4042.

[160] I. Mandric, T. Schwarz, A. Majumdar, et al. "Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis". In: *Nature Communications* 11.1 (2020), pp. 1–9. ISSN: 20411723. DOI: 10.1038/s41467-020-19365-w. URL: http://dx.doi.org/10.1038/s41467-020-19365-w.

[161] H. Zhu and H. Lakkis. "Sample size calculation for comparing two negative binomial rates". In: *Statistics in Medicine* 33.3 (Feb. 2014), pp. 376–387. ISSN: 02776715. DOI: 10.1002/sim.5947. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.5947.

[162] S. H. Jung. "Sample size for FDR-control in microarray data analysis". In: *Bioinformatics* 21.14 (2005), pp. 3097–3104. ISSN: 13674803. DOI: 10.1093/bioinformatics/bti456.

[163] A. F. Rendeiro, C. Schmidl, J. C. Strefford, et al. "Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks". In: *Nature Communications* 7.May (2016). ISSN: 20411723. DOI: 10.1038/ncomms11938.

[164] A. Moreno-Moral, M. Bagnati, S. Koturan, et al. "Changes in macrophage transcriptome associate with systemic sclerosis and mediate GSDMA contribution to disease risk". In: *Annals of the Rheumatic Diseases* 77.4 (2018), pp. 596–601. ISSN: 14682060. DOI: 10.1136/annrheumdis-2017-212454.

[165] 10X Genomics. *User Guides — 10x Genomics*. 2019. URL: https://www.10xgenomics.com/resources/user-guides.

[166] M. Enge, H. E. Arda, M. Mignardi, et al. "Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns". In: *Cell* 171.2 (2017), pp. 321–330. ISSN: 10974172. DOI: 10.1016/j.cell.2017.09.004. URL: https://doi.org/10.1016/j.cell.2017.09.004.

[167] H. E. Arda, L. Li, J. Tsai, et al. "Age-dependent pancreatic gene regulation reveals mechanisms governing human β cell function". In: *Cell Metabolism* 23.5 (2016), pp. 909–920. ISSN: 19327420. DOI: 10.1016/j.cmet.2016.04.002. URL: http://dx.doi.org/10.1016/j.cmet.2016.04.002.

[168] S. L. Wolock, R. Lopez, and A. M. Klein. "Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data". In: *Cell Systems* 8.4 (2019), pp. 281–291. ISSN: 24054720. DOI: 10.1016/j.cels.2018.11.005. URL: https://doi.org/10.1016/j.cels.2018.11.005.

[169] M. D. Luecken and F. J. Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular Systems Biology* 15.6 (2019), e8746. DOI: 10.15252/msb.20188746.

[170] F. A. Wolf, P. Angerer, and F. J. Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome Biology* 19.1 (2018), p. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1382-0.

[171] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. URL: https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008.

[172] D. C. Jones. *fastq-tools*. URL: https://help.rc.ufl.edu/doc/Fastq-tools.

[173] A. T. Lun, S. Riesenfeld, T. Andrews, et al. "EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data". In: *Genome Biology* 20.1 (2019), pp. 1–9. ISSN: 1474760X. DOI: 10.1186/s13059-019-1662-y.

[174] M. G. Van Der Wijst, D. H. De Vries, H. Brugge, et al. "An integrative approach for building personalized gene regulatory networks for precision medicine". In: *Genome Medicine* 10.1 (2018), pp. 1–15. ISSN: 1756994X. DOI: 10.1186/s13073-018-0608-4.

[175] H. Matsumoto, H. Kiryu, C. Furusawa, et al. "SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation". In: *Bioinformatics* 33.15 (2017), pp. 2314–2321. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx194.

[176] S. Aibar, C. B. González-Blas, T. Moerman, et al. "SCENIC: Single-cell regulatory network inference and clustering". In: *Nature Methods* 14.11 (2017), pp. 1083–1086. ISSN: 15487105. DOI: 10.1038/nmeth.4463.

[177] N. Papili Gao, S. M. Ud-Dean, O. Gandrillon, et al. "SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles". In: *Bioinformatics* 34.2 (2018), pp. 258–266. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx575.

[178] I. van Blokland, R. Oelen, H. de Groot, et al. "Single-cell dissection of the immune response after a myocardial infarction". In: *manuscript in preparation* ().

[179] M. A. Skinnider, J. W. Squair, and L. J. Foster. "Evaluating measures of association for single-cell transcriptomics". In: *Nature Methods* 16.5 (2019), pp. 381–386. ISSN: 15487105. DOI: 10.1038/s41592-019-0372-4. URL: http://dx.doi.org/10.1038/s41592-019-0372-4.

[180] T. P. Quinn, M. F. Richardson, D. Lovell, et al. "Propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis". In: *Scientific Reports* 7.1 (2017), pp. 1–9. ISSN: 20452322. DOI: 10.1038/s41598-017-16520-0. URL: http://dx.doi.org/10.1038/s41598-017-16520-0.

[181] T. Moerman, S. Aibar Santos, C. Bravo González-Blas, et al. "GRNBoost2 and Arboreto: Efficient and scalable inference of gene regulatory networks". In: *Bioinformatics* 35.12 (2019), pp. 2159–2161. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty916.

[182] V. Bergen, M. Lange, S. Peidli, et al. "Generalizing RNA velocity to transient cell states through dynamical modeling". In: *Nature Biotechnology* (2020). ISSN: 15461696. DOI: 10.1038/s41587-020-0591-3. URL: http://dx.doi.org/10.1038/s41587-020-0591-3.

[183] R. Cannoodt, W. Saelens, D. Sichien, et al. "SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development". In: *bioRxiv* (2016), pp. 1–15. DOI: 10.1101/079509. URL: https://www.biorxiv.org/content/10.1101/079509v2.

[184] V. Bergen, R. A. Soldatov, P. V. Kharchenko, et al. "RNA velocity—current challenges and future perspectives". In: *Molecular Systems Biology* 17.8 (2021), p. 2020. ISSN: 1744-4292. DOI: 10.15252/msb.202110282.

[185] Y. Baran, A. Sebe-Pedros, Y. Lubling, et al. "MetaCell: Analysis of single cell RNA-seq data using k-NN graph partitions". In: *bioRxiv* (2018), pp. 1–19. DOI: 10.1101/437665.

[186] V. A. Traag, L. Waltman, and N. J. van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific Reports* 9.1 (2019), pp. 1–12. ISSN: 20452322. DOI: 10.1038/s41598-019-41695-z.

[187] E. H. Simpson. "The Interpretation of Interaction in Contingency Tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 13.2 (July 1951), pp. 238–241. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1951.tb00088.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1951.tb00088.x.

[188] R. Gate, M. C. Kim, A. Lu, et al. "Mapping gene regulatory networks of primary CD4 + T cells using single-cell genomics and genome engineering". In: *bioRxiv* (2019), p. 678060. DOI: 10.1101/678060. URL: https://www.biorxiv.org/content/10.1101/678060v1%20https://doi.org/10.1101/678060.

[189] E. Papalexi, E. P. Mimitou, A. W. Butler, et al. "Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens". In: *Nature Genetics* 53.3 (2021), pp. 322–331. ISSN: 15461718. DOI: 10.1038/s41588-021-00778-2. URL: http://dx.doi.org/10.1038/s41588-021-00778-2.

[190] D. Szklarczyk, A. L. Gable, D. Lyon, et al. "STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets". In: *Nucleic Acids Research* 47.D1 (2019), pp. D607–D613. ISSN: 13624962. DOI: 10.1093/nar/gky1131.

[191] T. Qi, Y. Wu, J. Zeng, et al. "Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood". In: *Nature Communications* 9.1 (2018). ISSN: 20411723. DOI: 10.1038/s41467-018-04558-1.

[192] F. Hammal, P. De Langen, A. Bergon, et al. "ReMap 2022: A database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments". In: *Nucleic Acids Research* 50.D1 (2022), pp. D316–D325. ISSN: 13624962. DOI: 10.1093/nar/gkab996.

[193] C. A. de Leeuw, J. M. Mooij, T. Heskes, et al. "MAGMA: Generalized Gene-Set Analysis of GWAS Data". In: *PLoS Computational Biology* 11.4 (2015), pp. 1–19. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004219.

[194] S. Kasela, K. Kisand, L. Tserel, et al. "Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4 + versus CD8 + T cells". In: *PLoS Genetics* 13.3 (2017). ISSN: 15537404. DOI: 10.1371/journal.pgen.1006643.

[195] C. Polychronakos and Q. Li. "Understanding type 1 diabetes through genetics: Advances and prospects". In: *Nature Reviews Genetics* 12.11 (2011), pp. 781–792. ISSN: 14710056. DOI: 10.1038/nrg3069.

[196] D. Nickles, H. P. Chen, M. M. Li, et al. "Blood RNA profiling in a large cohort of multiple sclerosis patients and healthy controls". In: *Human Molecular Genetics* 22.20 (2013), pp. 4194–4205. ISSN: 09646906. DOI: 10.1093/hmg/ddt267.

[197] S. La Starza, M. Ferraldeschi, M. C. Buscarinu, et al. "Genome-wide multiple sclerosis association data and coagulation". In: *Frontiers in Neurology* 10.FEB (2019). ISSN: 16642295. DOI: 10.3389/fneur.2019.00095.

[198] R. Gonsky, P. Fleshner, R. L. Deem, et al. "Association of Ribonuclease T2 Gene Polymorphisms With Decreased Expression and Clinical Characteristics of Severity in Crohn's Disease". In: *Gastroenterology* 153.1 (2017), pp. 219–232. ISSN: 15280012. DOI: 10.1053/j.gastro.2017.04.002.

[199] I. Dotan, M. Allez, S. Danese, et al. "The role of integrins in the pathogenesis of inflammatory bowel disease: Approved and investigational anti-integrin therapies". In: *Medicinal Research Reviews* 40.1 (2020), pp. 245–262. ISSN: 10981128. DOI: 10.1002/med.21601. URL: http://dx.doi.org/10.1002/med.21601.

[200] H. Cai, J. Chen, J. Liu, et al. "CRIP1, a novel immune-related protein, activated by Enterococcus faecalis in porcine gastrointestinal epithelial cells". In: *Gene* 598 (Jan. 2017), pp. 84–96. ISSN: 03781119. DOI: 10.1016/j.gene.2016.11.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378111916308873.

[201] C. Hafemeister and R. Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome Biology* 20.1 (2019), pp. 1–15. ISSN: 1474760X. DOI: 10.1186/s13059-019-1874-1.

[202] Y. Hao, S. Hao, E. Andersen-Nissen, et al. "Integrated analysis of multimodal single-cell data". In: *Cell* 184.13 (2021), pp. 3573–3587. ISSN: 10974172. DOI: 10.1016/j.cell.2021.04.048. URL: https://doi.org/10.1016/j.cell.2021.04.048.

[203] G. La Manno, R. Soldatov, A. Zeisel, et al. "RNA velocity of single cells". In: *Nature* 560.7719 (2018), pp. 494–498. ISSN: 14764687. DOI: 10.1038/s41586-018-0414-6. URL: http://dx.doi.org/10.1038/s41586-018-0414-6.

[204] H. J. Westra, M. J. Peters, T. Esko, et al. "Systematic identification of trans eQTLs as putative drivers of known disease associations". In: *Nature Genetics* 45.10 (2013), pp. 1238–1243. ISSN: 10614036. DOI: 10.1038/ng.2756.

[205] S. Seabold and J. Perktold. "Statsmodels: Econometric and Statistical Modeling with Python". In: *Proceedings of the 9th Python in Science Conference* Scipy (2010), pp. 92–96. DOI: 10.25080/majora-92bf1922-011.

[206] M. Arnold, J. Raffler, A. Pfeufer, et al. "SNiPA: An interactive, genetic variant-centered annotation browser". In: *Bioinformatics* 31.8 (2015), pp. 1334–1336. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu779.

[207] S. H. Lin, D. W. Brown, and M. J. Machiela. "LDtrait: An online tool for identifying published phenotype associations in linkage disequilibrium". In: *Cancer Research* 80.16 (2020), pp. 3443–3446. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-20-0985.

[208] A. N. Barbeira, R. Bonazzola, E. R. Gamazon, et al. "Exploiting the GTEx resources to decipher the mechanisms at GWAS loci". In: *Genome Biology* 22.1 (2021), pp. 1–24. ISSN: 1474760X. DOI: 10.1186/s13059-020-02252-4.

[209] T. R. Gaunt, H. A. Shihab, G. Hemani, et al. "Systematic identification of genetic influences on methylation across the human life course". In: *Genome biology* 17 (2016), p. 61. ISSN: 1474760X. DOI: 10.1186/s13059-016-0926-z. URL: http://dx.doi.org/10.1186/s13059-016-0926-z.

[210] J. L. Min, G. Hemani, E. Hannon, et al. "Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation". In: *Nature Genetics* 53.9 (2021), pp. 1311–1321. ISSN: 15461718. DOI: 10.1038/s41588-021-00923-x.

[211] C. Do, C. F. Lang, J. Lin, et al. "Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation". In: *American Journal of Human Genetics* 98.5 (2016), pp. 934–955. ISSN: 15376605. DOI: 10.1016/j.ajhg.2016.03.027. URL: http://dx.doi.org/10.1016/j.ajhg.2016.03.027.

[212] D. Lin, J. Chen, N. Perrone-Bizzozero, et al. "Characterization of cross-tissue genetic-epigenetic effects and their patterns in schizophrenia". In: *Genome Medicine* 10.1 (2018), pp. 1–12. ISSN: 1756994X. DOI: 10.1186/s13073-018-0519-4.

[213] S. Heinz, C. E. Romanoski, C. Benner, et al. "The selection and function of cell type-specific enhancers". In: *Nature Reviews Molecular Cell Biology* 16.3 (2015), pp. 144–154. ISSN: 14710080. DOI: 10.1038/nrm3949.

[214] L. Bonaguro, J. Schulte-Schrepping, T. Ulas, et al. "A guide to systems-level immunomics". In: *Nature Immunology* 23.October (2022). ISSN: 15292916. DOI: 10.1038/s41590-022-01309-9.

[215] G. Gibson. "Perspectives on rigor and reproducibility in single cell genomics". In: *PLoS Genetics* 18.5 (2022), pp. 1–9. ISSN: 15537404. DOI: 10.1371/journal.pgen.1010210. URL: http://dx.doi.org/10.1371/journal.pgen.1010210.

[216] J. Wang, S. Xia, B. Arand, et al. "Single-Cell Co-expression Analysis Reveals Distinct Functional Modules, Co-regulation Mechanisms and Clinical Outcomes". In: *PLoS Computational Biology* 12.4 (2016), pp. 1–18. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004892.

[217] C. Chen, J. Peng, S. Ma, et al. "Ribosomal protein S26 serves as a checkpoint of T-cell survival and homeostasis in a p53-dependent manner". In: *Cellular and Molecular Immunology* 18.7 (2021), pp. 1844–1846. ISSN: 20420226. DOI: 10.1038/s41423-021-00699-4. URL: http://dx.doi.org/10.1038/s41423-021-00699-4.

[218] K. Kamimoto, C. M. Hoffmann, and S. A. Morris. "CellOracle: Dissecting cell identity via network inference and in silico gene perturbation". In: *bioRxiv* (2020). DOI: 10.1101/2020.02.17.947416.

[219] C. B. González-Blas, S. D. Winter, G. Hulselmans, et al. "SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks". In: *bioRxiv* (2022). DOI: 10.1101/2022.08.19.504505. URL: https://www.biorxiv.org/content/10.1101/2022.08.19.504505v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.08.19.504505v1.abstract.

[220] A. Griffon, Q. Barbier, J. Dalino, et al. "Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape". In: *Nucleic Acids Research* 43.4 (2015), pp. 1–14. ISSN: 13624962. DOI: 10.1093/nar/gku1280.

[221] L. Teng, B. He, J. Wang, et al. "4DGenome: A comprehensive database of chromatin interactions". In: *Bioinformatics* 31.15 (2015), pp. 2560–2564. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv158.

[222] D. Hnisz, B. J. Abraham, T. I. Lee, et al. "Super-Enhancers in the Control of Cell Identity and Disease". In: *Cell* 155.4 (Nov. 2013), pp. 934–947. ISSN: 00928674. DOI: 10.1016/j.cell.2013.09.053. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867413012270.

[223] W. J. Kent, C. W. Sugnet, T. S. Furey, et al. "The Human Genome Browser at UCSC". In: *Genome Research* 12.6 (June 2002), pp. 996–1006. ISSN: 1088-9051. DOI: 10.1101/gr.229102. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.229102.

[224] M. Lizio, J. Harshbarger, H. Shimoji, et al. "Gateways to the FANTOM5 promoter level mammalian expression atlas". In: *Genome Biology* 16.1 (2015), pp. 1–14. ISSN: 1474760X. DOI: 10.1186/s13059-014-0560-6.

[225] E. Eisenberg and E. Y. Levanon. "Human housekeeping genes, revisited". In: *Trends in Genetics* 29.10 (2013), pp. 569–574. ISSN: 01689525. DOI: 10.1016/j.tig.2013.05.010. URL: http://dx.doi.org/10.1016/j.tig.2013.05.010.

[226] A. D. Yates, P. Achuthan, W. Akanni, et al. "Ensembl 2020". In: *Nucleic Acids Research* 48.D1 (2020), pp. D682–D688. ISSN: 13624962. DOI: 10.1093/nar/gkz966.

[227] B. M. Javierre, S. Sewitz, J. Cairns, et al. "Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters". In: *Cell* 167.5 (2016), pp. 1369–1384. ISSN: 10974172. DOI: 10.1016/j.cell.2016.09.037.