DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Dissertation

Deciphering regulatory molecular mechanisms using graphical models

Johann Sebastian Hawe

July 2020

HelmholtzZentrum münchen

German Research Center for Environmental Health

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Deciphering regulatory molecular mechanisms using graphical models

Johann Sebastian Hawe

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r):	Prof. Dr. Stephan Günnemann
Prüfer der Dissertation:	1. TUM Junior Fellow Dr. Matthias Heinig 2. Prof. Dr. Julien Gagneur

Die Dissertation wurde am 06.07.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 09.11.2020 angenommen.

Acknowledgments

This work, and indeed my entire work during my time as a Ph.D. student, could not have been done without the help and continuous support of a select group of some very important people.

First, I would like to thank my Ph.D. advisor, Dr. Matthias Heinig, for giving me the excellent opportunity to perform my Ph.D. work at his lab. He always supplied me with interesting work and fruitful discussions from which I learned much. Thank you for your input and your support!

I would also like to thank all the great people in the Heinig Lab and the ICB. My time as a Ph.D. student would only have been half as much fun without my colleagues and office mates, and I'm glad to have had the opportunity to work with and learn from you. Valerio, we shared an office from the start, we wracked our brains together and had fun together. Thank you all for productive table tennis sessions and exhausting discussions (or was it the other way around?).

I am grateful to Alexis Battle for enabling a lab exchange with her lab at Johns Hopkins, which also started a fruitful collaboration. My time in the U.S. was a fantastic experience and really allowed me to grow, both personally and scientifically. Thank you!

Sometimes, it was good to have some distractions from 'the Ph.D. life' in more casual settings, be it in the form of training sessions, sport or music events, or just sitting together for a few drinks. I am lucky to have such wonderful friends who I can always count on. Thanks, every one of my friends, you are awesome!

I also want to acknowledge my thesis advisory committee, including Prof. Fabian Theis and Dr. Christian Gieger, and all the other valuable people I may have missed to mention in this acknowledgment. Thank you all for your input and support, and thank you Fabian for the possibility to do research at the ICB!

Finally, I want to acknowledge and thank the people closest to me. So many "thank yous" to my parents, Maria and Johann, who raised me and paved the way for what I have been doing all these years. Thanks, my brothers Andreas and Matthias and my sister Monika! It is always fun and you strengthened me. Thank you so much, Lena, my wife! You have been so very understanding and always supportive.

Really, to all of you: thanks!

Johann S. Hawe, July 2020

Preface

During the work leading to this dissertation, I contributed to several collaborative projects. To provide a complete context and better understanding of the work, I describe my contributions including the work of my collaborators in this thesis. In the following paragraphs, I give an overview of all works and highlight my contributions as well as acknowledge the important work of my collaborators.

meQTL project [1]

The meQTL project was a large collaborative effort between the Imperial College of London (ICL), the Research Unit of Molecular Epidemiology (AME) and the Institute of Computational Biology (ICB) at the HelmholtzZentrum München. This project is presented in Chapter 4. The corresponding manuscript is currently (July 2020) under review in *Nature Genetics* and parts of it including (adapted) figures have been used in this thesis. Prof. Chambers conceived the initial meQTL study and our collaboration partners from the ICL and AME provided the LOLIPOP and KORA cohort data and performed additional experiments to validate the computational results. Basic meQTL results, including the list of pruned meQTL, were provided by Dr. Lehne from the ICL and additional computational support provided by Dr. Loh (ICL) and Dr. Wilson (AME). I contributed to the project by performing functional follow up analyses for the mechanistic interpretation of the results. This included the integration of gene expression, Hi-C, chromHMM, and annotations of epigenetic regulators. I was the leading data analyst during the revisions of the manuscript. I performed computation and replication of genome-wide meQTL associations in KORA EPIC array data and additional enrichment analyses to corroborate our initial findings. Here, I also meta-analyzed and replicated genome-wide significant eQTM together with Katharina Schmid (ICB), who provided the important associations from the individual cohorts. Dr. Heinig conceived the random walk framework and I implemented analyses and interpretation of the network findings, by integrating the functional cohort data and providing follow up analyses, including *trans*-eQTM and eQTL calculations and visualization of networks for the manuscript. I further implemented the *QTLdb* website which provides easy and public access to the association results generated in this and other projects. Dr. Heinig and Prof. Chambers provided me with the opportunity to contribute significantly to this project which earned me a first authorship for this important work and for which I am very grateful. I would also like to thank all the collaboration partners whose extensive efforts went into the manuscript and in this thesis and who provided intensive discussions from which I learned much.

Network inference project [2]

This project is described in Chapter 5 of the thesis and is currently (July 2020) published on BioRxiv [2] and under review in *Genome Medicine*. Parts of the manuscript

have been used in this thesis including (adapted) figures. The study was conceived by Dr. Heinig and I performed all computational analyses, including the implementation of a fully reproducible workflow. I am grateful to have had access to the KORA and LOLIPOP data which was provided by our important collaborators Prof. Chambers from the ICL (LOLIPOP) and Dr. Gieger and Dr. Waldenberger from the AME (KORA) and their teams and without which this work would not have been possible. I would like to thank Prof. Battle and Ashis Saha from the Johns Hopkins University who assisted with the use of current GTEx v8 data. The valuable discussions with Prof. Battle significantly advanced this project. Also, I thank Prof. Theis, who, together with Prof. Battle, contributed to the design of the data analysis strategy.

Network inference review [3]

I wrote a review on network inference in multi-omics data together with my supervisor Dr. Heinig and Prof. Theis. It has been peer-reviewed and published in Frontiers in Genetics in 2019 [3]. The review provides a current and detailed view on multi-omics network inference, which went into this thesis in the introductory Chapter 1 and served as additional background for the network inference project (Chapter 5).

GR ChIP-seq analysis [4]

In addition, I have contributed to work that is not presented as part of the thesis, as it focuses on different aspects of gene regulation. This work led to a publication in Molecular Cell in 2019 [4]. Briefly, the main goal was to investigate circadian rhythmicity and diet-dependent activity of the glucocorticoid receptor (GR) nuclear transcription factor by analyzing its global DNA interaction profile in different conditions. To this end, 48 ChIP-seq experiments were conducted in which GR binding was profiled in mouse liver tissue. Mice were separated into two groups, group A (N=24) was harvested 5 days and group B (N=24) was harvested 12 weeks after the start of the experiment. These groups were further separated into mice nurtured with a low-fat diet (N=12) and with a high-fat diet (N=12). On the day of harvesting two mice were sacrificed starting at 7 am every 4 hours until 3 am the next day, yielding two biological replicates for each of the six time points, and liver tissue was extracted and subjected to GR ChIP-sequencing and RNA-sequencing. The study was designed and the experiments performed by our collaboration partners from the Institute of Diabetes and Obesity (IDO) of the HelmholtzZentrum München, including Prof. Uhlenhaut and Dr. Quagliarini who provided the numerous ChIP-seq samples. I implemented the initial analysis pipeline, including pre-processing, quality control, differential analysis, and visualization of results for the generated ChIP-seq data and performed follow up analyses together with Dr. Mir (IDO). Dr. Mir finalized the ChIP-seq analyses and analysis of RNA-seq data was performed by Kinga Balazs (IDO). I would like to thank all collaboration partners who made this work possible and Dr. Heinig for initiating my participation in this project, allowing me to investigate and learn much about these interesting data.

Abstract

Genetics studies seek to unravel the molecular mechanisms behind complex traits, which are influenced by both genetic and environmental factors. A majority of trait-related genetic variants are not easy to interpret as they lie outside of genes in regulatory regions. On the molecular level, genetic variants in regulatory elements might cause differences in gene expression and might ultimately lead to phenotypic differences.

In this thesis, we investigated regulatory networks to improve our understanding of trait-related variants and of how complex traits arise from genetic and epigenetic factors. To this end, we leveraged statistical associations between genetic variants and DNA methylation (meQTL) and gene expression (eQTL) derived from large-scale human population cohorts.

We identified genome-wide meQTL from data of 6,994 individuals and showed, that meQTL are enriched in active chromatin regions, in chromatin contact regions (TADs and Hi-C contacts) and for association with gene expression. *Trans*-QTL with numerous *trans* associated traits (QTL hotspots) are of particular interest as they are enriched in disease contexts. Integration with functional genomics data and network analyses allowed us to gain new insights into the mechanisms underlying *trans* hotspots, some of which were validated experimentally. We established candidate networks using a random walk based approach on *trans*-acting loci and enriched these networks by integrating multi-omics data. This approach allowed us to propose a novel regulatory network as the underlying mechanism of the rheumatoid arthritis associated *NFKBIE* locus.

Associations derived from multi-omics data aligned well with our random-walk networks, which prompted us to investigate novel methods for deciphering *trans* hotspot networks. To this end, we devised a unified strategy for the integration of *trans*-QTL hotspots with human multi-omics data and comprehensive biological prior knowledge. We analyzed these data and priors using state-of-the-art network inference algorithms in a large-scale simulation and replication study. We showed that methods utilizing prior knowledge outperform prior-agnostic methods and are robust to noise in priors. Detailed investigation of networks constructed from population-scale cohort data highlighted two novel molecular networks linked to schizophrenia and lean body mass. Both networks recovered known trait-associated genes and implicated novel genes.

Our studies advanced previous works by providing an extensive functional characterization of genome-wide QTL effects with a focus on *trans*-acting hotspots. We demonstrated, that existing biological knowledge can be used with multi-omics data to improve our understanding of genomic master regulators and to propose novel candidate genes linking genetic variants to human traits.

Kurzfassung

Genetische Studien haben zum Ziel, die molekularen Mechanismen hinter komplexen Merkmalen aufzudecken, welche sowohl von genetischen als auch von Umweltfaktoren beeinflusst werden. Ein Großteil der mit Merkmalen assoziierten genetischen Varianten ist nicht einfach zu interpretieren, da sie außerhalb von Genen in regulatorischen DNA Regionen liegen. Auf molekularer Ebene können genetische Varianten in regulatorischen Elementen Unterschiede in Genexpression verursachen und letztendlich zu phänotypischen Unterschieden führen.

In dieser Arbeit untersuchten wir regulatorische Netzwerke, um unser Verständnis von merkmalsbezogenen Varianten und der Entstehung komplexer Merkmale aus genetischen und epigenetischen Faktoren zu verbessern. Zu diesem Zweck nutzten wir statistische Assoziationen zwischen genetischen Varianten und DNA-Methylierung (meQTL) und Genexpression (eQTL), die aus großen menschlichen Bevölkerungskohorten abgeleitet wurden. Wir identifizierten genomweite meQTL aus Daten von 6.994 Individuen und zeigten, dass meQTL in aktiven Chromatinregionen, in Chromatinkontaktregionen (TADs und Hi-C-Kontakte) und für Assoziation mit Genexpression angereichert sind. *Trans*-QTL mit zahlreichen *trans*-assoziierten Merkmalen (QTL-Hotspots) sind von besonderem Interesse, da sie im Krankheitskontext vermehrt auftreten. Die Integration funktionaler genomischer Daten und zusätzliche Netzwerkanalysen ermöglichten es uns, neue Einblicke in die QTL-Hotspots zugrunde liegenden Mechanismen zu gewinnen, von denen einige experimentell validiert wurden. Mithilfe eines Random-Walk-basierten Ansatzes angewandt auf genomweit wirkende Loci rekonstruierten wir molekulare Netzwerke, welche wir durch die Integration von Multi-Omics-Daten bekräftigen konnten. Dieser Ansatz ermöglichte es uns ein regulatorisches Netzwerk als den zugrunde liegenden Mechanismus für den mit rheumatoider Arthritis assoziierten *NFKBIE*-Lokus vorzuschlagen.

Aus Multi-Omics-Daten abgeleitete Assoziationen stimmten gut mit unseren Random-Walk-Netzwerken überein, was uns dazu veranlasste, neuartige Methoden zur Entschlüsselung von QTL-Hotspot-Netzwerken zu untersuchen. Zu diesem Zweck entwickelten wir eine Strategie für die simultane Integration von *Trans*-QTL-Hotspots mit menschlichen Multi-Omics-Daten und umfassendem biologischem Vorwissen. Wir analysierten diese Daten mithilfe modernster Netzwerkinferenzalgorithmen. In einer groß angelegten Simulations- und Replikationsstudie konnten wir zeigen, dass Methoden, welche A-Priori-Informationen verwenden, andere Methoden übertreffen und robust gegenüber Fehlern in diesen Informationen sind. Eine detaillierte Untersuchung von aus Kohortendaten abgeleiteten Netzwerken ergab zwei neuartige molekulare Netzwerke, welche mit Schizophrenie und schlanker Körpermasse assoziiert sind. Beide Netzwerke

identifizierten bekannte und implizierten neue merkmalsassoziierte Gene.

Unsere Studien erweiterten frühere Arbeiten durch eine umfassende funktionelle Charakterisierung genomweiter QTL-Effekte mit Schwerpunkt auf *trans* Hotspots. Wir konnten zeigen, dass vorhandenes biologisches Wissen mit Multi-Omics-Daten integriert werden kann, um unser Verständnis genomweit wirkender Varianten zu verbessern und neue Gene vorzuschlagen, welche diese mit menschlichen Merkmalen verbinden.

Contents

Acknowledgments i

Preface iii

Abstract v

Kurzfassung vii

1. Introduction 1
- 1.1. Thesis aims and structure . . . 4
  - 1.1.1. Aims . . . 4
  - 1.1.2. Structure . . . 4
- 1.2. Genome-wide association studies . . . 6
- 1.3. The central dogma of molecular biology and gene regulation . . . 7
  - 1.3.1. The central dogma . . . 7
  - 1.3.2. Mechanisms of gene regulation . . . 8
- 1.4. Systems biology and biological interactions . . . 11
  - 1.4.1. Experimental detection of interacting molecules . . . 11
  - 1.4.2. Quantitative trait locus studies . . . 13
  - 1.4.3. Biological network inference in systems biology . . . 15
  - 1.4.4. Leveraging biological prior knowledge for network reconstruction 17
- 1.5. Reproducible research . . . 20
  - 1.5.1. Reproducibility in computational biology . . . 21
  - 1.5.2. Dedicated workflow systems . . . 22
  - 1.5.3. Distributing workflows and software environments . . . 22

2. Materials 25
- 2.1. Microarrays: a cost effective way of obtaining multi-level functional data from large population cohorts . . . 25
  - 2.1.1. Determining the genetic make up of individuals through genotyping arrays . . . 26
  - 2.1.2. Quantitative DNA methylation analysis using microarrays . . . 28
  - 2.1.3. Using microarrays to quantify gene expression . . . 29
- 2.2. Population cohort data used in this thesis . . . 30
  - 2.2.1. The Cooperative Health Research in the Region of Augsburg . . . 31
  - 2.2.2. The London Life Sciences Prospective Population Study . . . 31
  - 2.2.3. Northern Finland Birth Cohorts . . . 32

2.2.4. The Saguenay Youth Study . . . 33
2.3. Public data . . . 33
2.3.1. The Encyclopedia of DNA Elements . . . 33
2.3.2. The Genotype Tissue Expression consortium . . . 34
2.3.3. The ARCHS4 database . . . 34
2.3.4. The Roadmap Epigenomics project . . . 35
2.3.5. STRING and BioGRID . . . 35
2.3.6. The ReMap resource . . . 36

3. Methods **37**
3.1. Statistical background . . . 37
3.1.1. Conditional independence and correlation of random variables . . . 37
3.1.2. Linear models . . . 38
3.1.3. Regularization in linear models . . . 43
3.1.4. Meta analysis . . . 44
3.1.5. Enrichment testing . . . 45
3.1.6. Multiple testing . . . 46
3.1.7. Additional background . . . 47
3.2. Processing of cohort data . . . 49
3.2.1. Genotyping data. . . . 49
3.2.2. Methylation data. . . . 49
3.2.3. Gene expression data. . . . 50
3.3. Processing of replication and validation data . . . 50
3.3.1. Data used for meQTL replication . . . 50
3.3.2. IP-MS data used for experimental validation of the *ZNF333* locus 52
3.4. Reproducible cloud enabled workflows . . . 52
3.4.1. The Snakemake workflow system . . . 52
3.4.2. Reproducible software environments . . . 53

4. Exploring the genetic architecture of DNA methylation **55**
4.1. Epigenetic gene regulation through DNA methylation . . . 56
4.2. Methods for meQTL investigation . . . 58
4.2.1. Identification and pruning of global meQTL . . . 58
4.2.2. Replication of meQTL in independent data . . . 61
4.2.3. Enrichment of *cis*-meQTL in functional chromatin states . . . 63
4.2.4. Enrichment of *longrange*- and *trans*-meQTL within chromatin contacts . . . 63
4.2.5. Enrichment of meQTL pairs for association with gene expression . 65
4.2.6. Selection of candidate genes for SNPs affecting CpGs in *trans* . . . 66
4.2.7. Enrichment of regulatory genes at meQTL loci . . . 66
4.2.8. TFBS enrichment at *trans* associated CpG sites . . . 66
4.2.9. Random walk analysis on locus graphs . . . 68
4.2.10. Experimental validation of novel regulators . . . 72

4.3. Genome-wide analysis of genetic effects on DNA methylation 74
4.3.1. Identification of a cosmopolitan set of meQTL 75
4.3.2. Replication in isolated leukocytes, adipocytes and adipose tissue . 76
4.3.3. Functional enrichment analyses 78
4.3.4. *Trans*-meQTL reveal novel regulatory patterns 83
4.3.5. Experimental validation confirms novel regulators 86
4.3.6. Effect of increased resolution and coverage on meQTL results . . . 88
4.4. Project summary 90

5. Prior based network inference **93**
5.1. Inferring multi-omics networks from functional data 94
5.2. Methods for regulatory network inference on *trans* hotspots 96
5.2.1. Generation of locus sets 96
5.2.2. Sources and formulation of biological priors 98
5.2.3. Simulation study design and replication analysis 101
5.2.4. Estimation of transcription factor activities 102
5.2.5. Graphical model based network inference 103
5.2.6. Network prioritization and final network creation 110
5.2.7. Colocalization analysis to corroborate networks 111
5.3. Multi-omics integration for *trans* hotspot regulatory network inference . . 113
5.3.1. Leveraging *trans*-QTL hotspots to reduce complexity 114
5.3.2. Collection of prior information 115
5.3.3. Method comparison by simulation and replication study 116
5.3.4. Application to real-world population data 122
5.4. Project summary 127

6. Discussion **129**
6.1. Systematic assessment of the genetic effects influencing DNA methylation 129
6.2. Identification of *trans*-acting regulatory mechanisms underlying DNA methylation 131
6.3. Biologically informed priors improve network inference 132
6.4. Prior based network inference yields novel insights into disease loci . . . 133
6.5. Future perspective of single-cell data in systems biology 134
6.6. Conclusions 135

Appendices **137**

A. List of Figures **139**

B. List of Tables **141**

C. Supplementary Information **143**
A. Data used for meQTL replication 143
A.1. Isolated white blood cell studies 143

A.2. Isolated adipocyte studies 144
A.3. DNA methylation in adipose tissue 144
B. Experimental validation of the ZNF333 *trans* locus 145
B.1. ChIP-seq experiment to determine *ZNF333* binding sites 145
B.2. Pull-down assay to identify ZNF333 binding partners 146
C. A public browser for quantitative trait loci 146
D. Workflow and code availability 147
E. Supplementary Tables 149
F. Supplementary Figures 153

Bibliography **157**

1. Introduction

Most common diseases in humans are complex traits, driven by multiple genetic and environmental factors [5]. Genetics studies seek to identify and quantify the genetic contribution in causing complex traits. DNA is the carrier of genetic information in living organisms, and differences in the DNA sequence, i.e. genetic variants, are the drivers of phenotypic variation between individuals and are central to understanding complex traits. Each diploid organism inherits two full copies of DNA in the form of individual chromosomes, one set of chromosomes from each parent. For instance, humans have 46 chromosomes occurring in pairs of two copies. Thus, each gene, i.e. a specific section on the DNA, is present in two copies, one on each of the parental chromosomes. The inherited genes can be identical or differ between the parental chromosomes and a gene is therefore said to be present in the same or different 'alleles' (types), where the combination of the two alleles is known as an individual's genotype for that gene. During germ cell generation, mixing of genes located on different chromosomes takes place, and only one randomly chosen allele for each gene is passed on to the new cell, leading to a mix of genetic information. Alleles on the same chromosome typically stay together. However, due to recombination events taking place between genetic loci, genes can be exchanged between the two parental chromosome copies, bringing together the paternal and maternal alleles on the same recombined chromosome. Genes further apart on the same chromosome have a relatively high probability of a recombination event occurring between them, in contrast to genes in close proximity. Such genes or genetic loci are said to be in linkage disequilibrium (LD), i.e. they are more often inherited together as one might expect in case of independent inheritance, and LD can make the interpretation of trait-associated genetic variants identified through genome-wide association studies (GWAS) difficult.

Nowadays, genetic studies typically investigate genetic markers such as single-nucleotide polymorphisms (SNPs), i.e. single base pair differences, rather than individual genes. In GWAS, millions of genetic markers of individuals in large population cohorts are tested for statistical association with complex traits, such as a specific disease, which enabled the discovery of many disease-associated (positively tested) genetic loci. But although GWAS have been very successful in identifying trait-associated loci [6], for most disease variants a direct causal explanation, i.e. how they specifically affect the studied trait, is not straight forward [7]. For instance, GWAS rely on genetic markers flagging numerous individual variants in close LD to each other and determining the actual disease driving ('causal') variant is difficult, because causal variants cannot be distinguished statistically from other correlated variants in LD [8].

Genetic variants could directly impact protein function by changing a protein-coding DNA sequence, i.e. DNA regions that are translated into proteins by cellular mechanisms to execute specific functional or structural roles [7]. This has been the expected mechanism of action for trait-associated variants and has been observed before for rare genetic diseases. However, for most trait-associated variants identified in GWAS (approx. 93%) this does not hold as they lie outside of protein-coding regions and only indirectly affect proteins [9]. The two main steps of processing protein-coding genes are 1) gene transcription from DNA into RNA and 2) translation of RNA into proteins, which can be summarized as *gene expression* and which is controlled through *regulatory elements* lying in the non-protein-coding part of the genome, sometimes far away from any protein-coding regions [9]. Disease associated variants are enriched for location in such regulatory elements [9]. It is, therefore, necessary to determine the genes affected by these variants, i.e. the genes which expression changes (either enhanced or repressed) due to changes of the genetic variant, and thereby explain how the variant affects a cell on a molecular level.

Due to LD and location of GWAS variants in non-coding regions, research focus has shifted in recent years from the discovery of trait-associated variants to *explanation and mechanism*, seeking to dissect the molecular consequences of trait-associated genetic loci [6, 10] and to pinpoint the causal variants. Recent molecular studies seek to understand, how and which genetic variants exert control over gene expression, and these studies have been made possible through technological breakthroughs in profiling genome-wide molecular data for large numbers of individuals. For instance, it is now possible to quantify the level of gene expression of all genes in a cell, yielding good readouts of global gene activity and enabling the identification of genetic variants, e.g. through associating genotypes with expression levels, which exert a direct impact on genes (quantitative trait loci, QTL) [11].

Importantly, the expression of genes is also controlled through *epigenetic modifications* of the DNA, which do not change the underlying DNA sequence and which are variable even between cells of the same organism, in contrast to the DNA sequence. These modifications can be influenced by environmental factors [12–14] and they can, for instance, alter the accessibility of DNA for specific proteins. Those proteins are in turn responsible for activating or deactivating gene expression, and genetic and epigenetic mechanisms together can form complex regulatory interactions to exert their control on gene expression [11, 15]. It is now also possible to profile epigenetic marks, such as DNA methylation, genome-wide, allowing additional insights into the regulation of gene expression [16–18]. For example, causal variants could affect methylation of DNA at regulatory elements, leading to altered binding of transcriptional regulator proteins, which are in turn involved in modulating gene expression of a particular target gene [15, 19].

Current biological datasets can provide measurements of diverse molecular layers (e.g. genotype, expression or methylation layers) for the same set of individuals and are

hence often termed *multi-omics* data[1]. Recent studies seek to unravel disease mechanisms by integrating GWAS variants with QTL information from multi-omics data [3, 11], for example in expression [e.g. 20–22] or epigenetic [e.g. 1, 18, 23, 24] contexts. However, control of gene expression is more complex than a single variant influencing a single neighboring gene, similar to many diseases not being caused by a single gene or variant [10]. For example, studies seek to identify the likely causal variants for a trait and which regulatory elements they affect, the transcriptional regulators affected by changes in the regulatory elements as well as the respective regulatory target genes on the chromosome [23, 25]. In addition, another layer of complexity is added by regulatory elements being able to influence genes on different chromosomes (i.e. in *trans* of the regulatory element) than their own, and studies also aim to pinpoint these *trans* effects and how they are established [23, 25]. Importantly, *trans*-acting variants, specifically the ones statistically associated with numerous molecular traits, are enriched for disease associations, and therefore represent central subjects of investigation in genetics research [22, 23, 26]. Moreover, the 'omnigenic' model predicts that approx. 70% of heritability originates from *trans* effects [27] further emphasizing the need to systematically study the mechanisms underlying these loci. Generally, individual genes and loci interact in complex regulatory networks [11], involving numerous types of interacting molecules, e.g. protein-protein and protein-DNA interactions, which define the genome-wide regulatory processes driving complex traits. Identifying and understanding these networks thus is essential to comprehend disease mechanisms [11]. The recent advances in obtaining multi-omics profiles allow to investigate these relationships in detail and to recover regulatory networks, for instance by testing and combining statistical associations over multiple molecular layers, thereby detailing the molecular effects of *trans*-acting, trait-associated genetic variants [11, 21, 23, 28, 29]. Moreover, nowadays the wealth of biological data available to researchers through public databases can provide new angles for computational research, for instance by utilizing established interaction and multi-omics data to alleviate the reconstruction of genome-wide regulatory relationships [30–33].

What has been missing so far, is to take advantage of emerging population-scale multi-omics data in humans to investigate and understand the regulatory mechanisms underlying *trans*-acting variants, and thus advance our understanding of genetic and epigenetic gene regulation and complex traits.

[1] *omics* refer to the study of *omes*, the entirety of a certain subject. For instance, gen*omics* provides information about the gen*ome*

1.1. Thesis aims and structure

1.1.1. Aims

In this thesis, we seek to understand, how trait-associated genetic variants affect regulatory mechanisms throughout the genome by uncovering the regulatory networks underlying *trans*-acting variants. We provide a systematic assessment of important *trans*-QTL hotspots together with novel strategies for computationally reconstructing their underlying networks in the context of methylation and gene expression in humans. To this end, we developed and applied computational frameworks for biological high-throughput datasets, including genotype, gene expression and DNA methylation array data, to unravel the complex biological mechanisms underlying genetic and epigenetic gene regulation. These frameworks were applied to functional data in diverse contexts to understand the mechanistic consequences of *trans*-QTL hotspots, i.e. genomic master regulators. Specifically, we identified regulatory networks and candidate genes for *trans* hotspots, which have frequently been associated with diseases, by integrating curated prior knowledge with population scale multi-omics data and applying state-of-the-art algorithms for network gene prioritization and network inference.

1.1.2. Structure

In the remainder of this first chapter, we will give a general introduction to the field of systems biology and lay the foundations for the biological background needed to follow through with the rest of this work (Sections 1.2 and 1.3). Following this, we will introduce the concept of molecular interactions and quantitative trait loci (Section 1.4), and how investigating these can help shed light on results from genome-wide association studies. Following this, we give an overview on the computational procedures which can be employed to establish molecular interaction networks in diverse cellular contexts (Sections 1.4.3 and 1.4.4). We conclude the introductory chapter in Section 1.5, by briefly discussing the subject of reproducible research in computational biology and highlighting its importance in all modern research projects.

To give the reader the necessary overview on the type of data used throughout this thesis, Chapter 2 describes general experimental techniques to extract molecular level information from tissue samples such as genotypes (DNA sequence), gene expression and DNA methylation measurements. Here, we also introduce the large population cohorts from which such data were gathered and which we utilized in the described projects. In addition, we give a brief description of the public data repositories from which we obtained genome annotation, molecular interaction and high-throughput data for our analyses.

In Chapter 3, we will discuss the most important statistical aspects and methods forming the background for the computational approaches employed in the later chapters,

such as linear models and statistical enrichment tests. We further describe the general processing steps for the used cohort data and describe data processing for the data used for experimental validation of our results of the *trans*-meQTL project described in Chapter 4). In Section 3.4, we will take a closer look at reproducible research in computational biology and the specific tools we applied to achieve reproducibility in our studies.

After the methods background, we will describe the two most important projects which led up to the creation of this thesis, entailing the 'meQTL project' (Chapter 4) and the 'prior based network inference project' (Chapter 5). At the beginning of each project chapter, we show a brief glossary summarizing the most important acronyms and expressions used throughout the respective chapter.

In the first project (Chapter 4), we investigated global patterns of DNA methylation in the light of genetic variation, generating novel insights in genetic and epigenetic gene regulation and which we utilized for the interpretation of disease loci. After briefly revisiting the fundamentals of gene regulation through DNA methylation (Section 4.1), we detail the methods used to derive and functionally characterize meQTL identified across ethnicities ('cosmopolitan' meQTL) from two large population cohorts (Section 4.2.1). We then proceed to describing the results of our large-scale meQTL analysis, including replication in independent data (Section 4.3.2), the functional enrichment analyses (Section 4.3.3) and the two-step random walk based regulatory network generation for genetic hotspot variants (Section 4.3.4). In this project, we highlighted candidate genes and novel regulatory pathways involving DNA methylation systematically for identified *trans*-meQTL hotspots. For instance, we highlight a genetic locus around the *NFKBIE* gene. Our findings indicated a likely regulatory mechanism linking genetic variation at that locus with rheumatoid arthritis, an autoimmune disorder, mediated through DNA methylation regulation at CpG sites located in *cis* to genes important for the regulation of *IL-6* biosynthesis, a gene central to immune response. For another locus involving the *ZNF333* gene, we highlighted a hitherto undescribed regulatory pattern which we could validate experimentally.

Moving to Chapter 5, we detail the prior based network inference project which is in part based on the meQTL results obtained in the previous chapter. Motivated by our observations in the two-step network analysis in Chapter 4, we formulate a unified approach for direct multi-omics data integration, involving both discrete and quantitative a-priori information about molecular interactions. We first motivate our idea of using data-driven priors in network reconstruction (Section 5.1) and then describe the methodology employed to compute prior guided networks (Section 5.2.5). Ultimately, we show how we applied our strategy to recover regulatory networks from genome-wide trans-quantitative trait loci, including the meQTL from the previous chapter (Section 5.3). We highlighted novel, complex trait associated loci for which we

generated new regulatory hypotheses. For example, we generated a network around a schizophrenia susceptibility locus, involving known genes related to neurological disorders and schizophrenia (e.g. *PBX2*, *RNF5*) and implicating new, potentially disease relevant genes (e.g. *TCF12*, *CD6*). Similarly, we applied our framework to a hotspot identified in Skeletal Muscle tissue, and provided new insights for the lean body mass associated locus.

Finally, in the last chapter of this thesis we discuss our results from the two projects and put our findings into perspective with the current literature. We further provide an outlook in interesting topics for future studies in this area.

1.2. Genome-wide association studies

DNA provides the molecular blueprint of living organisms and individual genotypes, such as single nucleotide polymorphisms (SNPs, single changes in the DNA sequence), are the molecular determinants of inter-individual phenotypic variation [34]. Genetic variants such as SNPs can impact cellular mechanisms and structures in numerous ways, for instance by directly altering protein-coding regions of the genome. In recent years, tremendous efforts have been put into identifying the genetic drivers (e.g. SNPs) behind complex human traits such as diseases. Genome-wide association studies (GWAS) are a popular way of investigating associations between genotypes and complex traits, and thousands of GWAS investigating thousands of traits have been conducted in the last decade [6, 8]. A GWAS boils down to association testing of individual genotypes with the trait under investigation (BMI or a disease, for instance) under consideration of potential confounding variables such as population structure, age, or sex. Typically, in a single GWAS, millions of genotypes for a population are independently tested for association with the trait of interest using for example linear or logistic regression models. GWAS hits lying in protein-coding regions of the genome (see Section 1.3) can highlight specific genes as e.g. potential drug targets. However, a significant proportion (approx. 93%) of disease-associated SNPs is found in non-coding parts of the genome and far away from any protein-coding regions [9]. For instance, such non-coding genetic variants could have an impact on the binding of proteins to the DNA which are required to initiate gene transcription [19]. Generally, interpretation for these genetic loci concerning their functional impact is not straight forward. In addition, GWAS determine trait-associations for SNPs in linkage-disequilibrium (LD) blocks, containing numerous highly correlated SNPs rather than pinpointing individual 'causal' SNPs with a direct effect on the disease, hindering further insights in disease pathophysiology. Therefore, in recent years the focus in genomic research has shifted from identifying disease loci to obtaining mechanistic explanations, seeking to understand the molecular and cellular consequences of disease-related DNA variants. These efforts, which we also pursued in this thesis, have been made possible through advances in measuring genomics data, and a specific focus is now on understanding *gene regulation*, for which we will provide the

necessary background in the next section.

1.3. The central dogma of molecular biology and gene regulation

1.3.1. The central dogma

In the heart of modern molecular biology lies its decade-old central dogma, first proposed by Francis Crick [35] and in a different formulation by James Watson [36] in the 1950s. While Watson describes a relatively simple two-step processing model, where the carrier of genetic information in all living cells, i.e. DNA (desoxyribonucleic acid), is transcribed into RNA (ribonucleic acid) and RNA is further translated into proteins, Crick's model makes a statement about how information flows from DNA to RNA and protein, but never from protein to DNA or RNA, although information flow from RNA to DNA is stated as a possibility (see Figure 1.1). Information, in this context, is defined as the 'precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein' [35]. For instance, the DNA sequence, composed of the four nucleotides adenine (**A**), cytosine (**C**), thymine (**T**), and guanine (**G**), determines the RNA sequence (made of the same nucleotides, except thymines are replaced with uracils, **U**). The latter model by Crick still holds today and mechanisms that reverse transcribe RNA to DNA have been identified [37], however, more insights have been generated in how the expression of genes (i.e. their transcription to RNA) is controlled. Genes are defined sections on the DNA that can encode specific transcripts (RNAs) and if the gene is 'protein-coding', these RNAs form templates for proteins. The genes in the human genome (though this holds true for other organisms as well) are not all actively transcribed all the time, but rather are carefully orchestrated by the cell to be expressed (transcribed) only when needed [38]. For example, different genes might be active depending on the cell cycle stage or, in case of cell differentiation, different genes are active and inactive in more differentiated cells as compared to their progenitors (see e.g. Y.-H. Zhang, Y. Hu, Y. Zhang, et al. [39] for an example in blood cell differentiation). These differences are driven by epigenetic control, meaning that no changes in the DNA sequence are involved, but rather the way the sequence is interpreted is adjusted for the distinct contexts [15, 38].

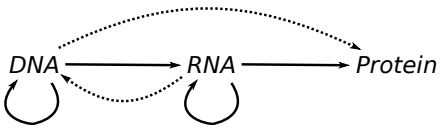


Figure 1.1.: The first draft of the 'Central Dogma of Biology' as outlined by Francis Crick [35]. As Crick stated, information flow between nucleic acids and from nucleic acids to protein is possible, however, it is never possible from protein to nucleic acids.

1.3.2. Mechanisms of gene regulation

In the projects described in this thesis, we aim to improve our understanding of the complex molecular mechanisms underlying *gene regulation*. At a specific point in time (e.g. cell-cycle stage) for specific cells (e.g. pancreatic cells, muscle cells, etc.) only the required subset of genes is actively transcribed and, if protein-coding, translated [38]. Genes not required on the other hand are turned off. The control of which genes are active to which degree is mostly referred to as *gene regulation* and is often viewed on the transcriptional level, i.e. whether and which kind of RNA is transcribed. However, this is a somewhat simplified view, as there are mechanisms of gene regulation beyond the transcriptional regulation. For example, even if a gene is transcribed, other RNAs can lead to degradation or inhibit translation of the freshly formed RNA by interacting with it, a mechanism known as RNA interference [40, 41]. Another example of gene regulation is alternative splicing. A gene is composed of exons (coding) and introns (non-coding) parts, and, typically, intronic regions of the gene are spliced out prior to it being translated into a protein. Hence, alternative splicing is an additional processing mechanism for transcripts, which allows a single gene to be processed into multiple, potentially protein-coding, transcripts [42] (so-called 'isoforms'), e.g. by selectively including or excluding specific exons or by forming alternative transcription start sites. Moreover, finished gene products (proteins) can also be modified through post-translational modifications, achieving more protein diversity as well as control of protein activity [43].

While the above are important mechanisms in gene regulation, in this thesis, we will focus on the epigenetic control of genes at the level of DNA transcription. Here, two types of epigenetic marks come into play: histone modifications and DNA methylation (see Figure 1.2). Mechanistically, both marks are tightly linked in regulating gene expression. Histone modifications are post-translational modifications of histone proteins at nucleosomes. Nucleosomes are structural components of chromosomes and each nucleosome consists of DNA that is wrapped around a complex of eight histone proteins or histones [44]. The complex consists of two copies each of four distinct proteins, H2A, H2B, H3, and H4. Different post-translational modifications of histone tails, i.e. the N-terminal part of the histone proteins, can lead to more or less accessible DNA to the transcriptional machinery (e.g. the RNA polymerase or other proteins needed for transcription, such as transcription factors). For example, acetylation of the 27th lysine residue (one letter code 'K') of H3, abbreviated as 'H3K27ac', is a mark generally associated with active regulatory regions and found e.g. in the promoter region (i.e. a specific area upstream and downstream of a genes transcription start site), of actively transcribed genes [44, 45]. On the other hand, high levels of H3K27me3 (tri-methylation of the 27th lysine residue of H3) have been associated with silent promoters and hence are seen as a sign of repressed gene transcription (compare Figure 1.2) [44, 46]. Moreover, histone modifications act in a combinatorial manner, such that different combinations of histone modifications at nucleosomes can implicate distinct consequences (forming the 'histone code') [47]. Large-scale projects measuring histone modifications in numerous

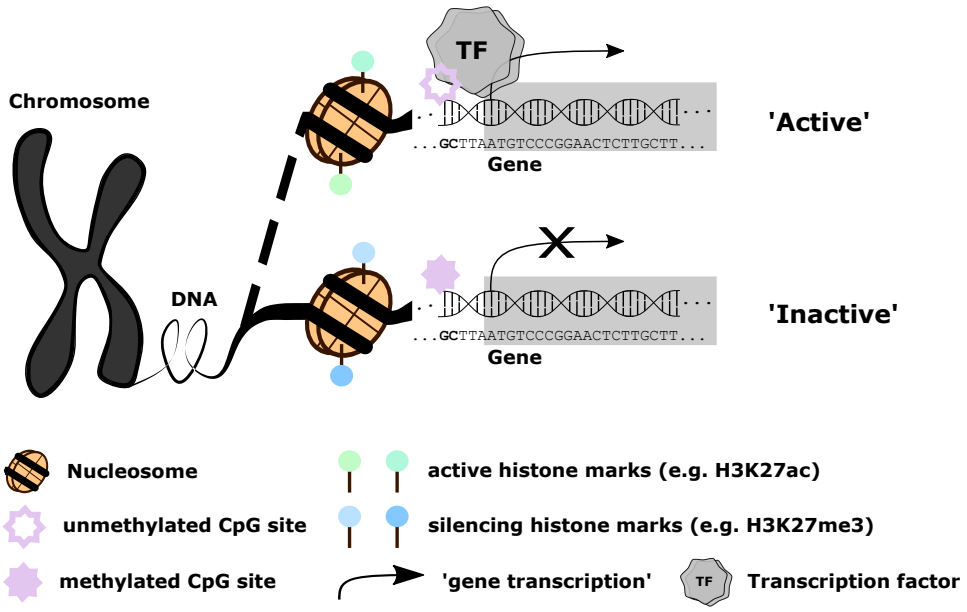


Figure 1.2.: Schematic showing the two main marks of epigenetic gene regulation, including DNA modifications via methylation of cytosines at CpG sites (purple marks) and post-translational modifications of histone tails (green and blue marks).

cell types (e.g. Roadmap Epigenomics [48]) set out to decipher this histone code, and tools such as chromHMM [49] have been successfully applied to data measuring histone modifications (e.g. ChIP-seq, see also Section 1.4.1) to define regulatory chromatin states (e.g. active TSS, enhancer, quiescent; compare Table 1.1). We make heavy use of established knowledge around histone modifications in both projects discussed in this thesis. In Chapter 4, we use information derived from histone modifications to assess the functional relevance of associations between genetic variants and DNA methylation, and in Chapter 5, we utilize established knowledge about the functional implications of histone modifications to generate informative priors for the inference of regulatory networks for explaining disease-associated variants.

Another level of epigenetic regulation, which plays the most important role in this thesis, is the methylation of cytosines, mostly at CG dinucleotides on the DNA (also termed 'CpG sites'), and referred to, simply, as *DNA methylation*. The methylation of DNA has been described as a crucial cellular mechanism, driving functional and structural properties of the genome, and is involved in the regulation of cellular differentiation and gene expression [12, 38]. Moreover, aberrations of DNA methylation patterns have been implicated to play a role in several complex diseases, such as neuropsychiatric disorders, atherosclerosis, cancer and type 2 diabetes [50–54] and DNA methylation patterns have been found to differ drastically between groups of individuals, for instance between smokers and non-smokers [14, 55]. These studies showed that environmental

state number	abbreviation/mnemonic	description
1	**TssA**	**Active TSS***
2	**TssAFlnk**	**Flanking Active TSS***
3	TxFlnk	Transcr. at gene 5′ and 3′
4	Tx	Strong transcription
5	TxWk	Weak transcription
6	**EnhG**	**Genic enhancers$^+$**
7	**Enh**	**Enhancers$^+$**
8	ZNF/Rpts	ZNF genes & repeats
9	Het	Heterochromatin
10	TssBiv	Bivalent/Poised TSS
11	BivFlnk	Flanking Bivalent TSS/Enh
12	**EnhBiv**	**Bivalent Enhancer$^+$**
13	ReprPC	Repressed PolyComb
14	ReprPCWk	Weak Repressed PolyComb
15	Quies	Quiescent/Low

Table 1.1.: ChromHMM states obtained from ChromHMM using the 15 state model model. States representing 'active regulatory states' and those used latter in this thesis are indicated in bold. The asterisks (*) indicate *promoter* related states and the pluses ($^+$) *enhancer* related states.

and genetic factors can influence DNA methylation and hence that it could provide a mechanistic explanation of how these exposures impact gene regulation and molecular phenotypes [13, 56–58]. Moreover, DNA methylation has previously been associated with genetic variation in multiple studies [18, 23, 50, 59], further hinting at a complex interplay of genetic (DNA variants) and environmental (e.g. smoking) factors in the regulation of gene expression and complex traits.

Biologically, depending on the relative position of the methylation site to a gene, changes in methylation can either enhance or repress transcription [60]. For example, methylation at gene promoters has widely been associated with gene silencing [61]. A possibility of how the effect might be established is by DNA methylation preventing transcriptional proteins to bind to the promoter and hence also preventing the respective gene to be transcribed. On the other hand, it has been shown that highly active genes often exhibit high levels of methylation within the gene body [62]. The mechanisms of how methylation of gene bodies might favor transcription are not yet well understood, but a reason for gene body methylation could be an effect on splicing of the gene transcript [63] or to avoid spurious transcription of the gene body [61].

In Chapter 4 of this thesis, we aim to understand the effects of genetic variants on DNA methylation, including the effects of disease-associated variants. We identify the regulatory mechanisms involved in mediating genetic effects on DNA methylation, including the effects of methylation on gene expression, to further our understanding of complex traits.

1.4. Systems biology and biological interactions

Systems biology seeks to model complex biological systems by generating a holistic view of all underlying cellular processes [64], including regulatory processes. By including genetic variants in the system, it is possible to observe their effects on a studied phenotype, i.e. how a change in sequence propagates through the cell via interaction networks to produce a specific trait [11]. Thereby, systems biology can help in providing a mechanistic explanation of the effect of disease-associated variants, an approach we adopt in this thesis for disease variants with genome-wide regulatory effects.

At the center of systems biology lies the central dogma of biology, i.e. the information encoded in the DNA is processed to RNAs, which are ultimately translated into proteins, and this process is tightly regulated (e.g. through epigenetic mechanisms such as DNA methylation or histone modifications) [35]. By looking at the entire DNA sequence (genome), the complete set of RNAs (transcriptome) and proteins (proteome), the entirety of DNA methylation (methylome) and combinations of those 'omic' layers, systems biology aims to understand how exactly information processing in cells is achieved. Specifically, systems biology postulates that by understanding the regulatory networks formed by molecular interactions within and between omic layers, for instance, which transcription factor proteins affect which genes, it is also possible to understand the molecular basis of diseases and other system-level phenotypes [64]. Generally, molecular interactions can be classified as either physical/direct or functional/indirect associations. Physical interactions involve two molecules directly interacting with each other, such as proteins in a protein complex. For functional interactions, on the other hand, the molecules involved are associated for example by exhibiting a common function or being involved in the same molecular pathway.

We will make extensive use of established molecular interactions later in this thesis for dissecting the effects of genetic hotspots and to construct regulatory networks from high-throughput molecular data. In this section, we will give a general overview of experimental techniques and computational strategies to recover physical and functional (direct and indirect) molecular interactions. In Section 1.4.1, we will highlight some experimental protocols to establish physical and functional interactions. Starting from Section 1.4.2, we discuss computational approaches to investigate high-throughput functional data for molecular interactions. For the computational part, we will focus specifically on the integration of multi-omics data, i.e. measurements of different omics on the same set of samples, and interaction profiling within these data, as this represents the main idea behind the projects discussed later in this thesis.

1.4.1. Experimental detection of interacting molecules

We utilize published molecular interactions later in this thesis to help understand genetic loci underlying complex traits, both in the form of established interaction

networks and as prior information for network inference on functional omics data. Detection of interacting bio-molecules has gotten much attention in molecular biology and numerous experiments have been designed to achieve this [65–70]. A large-scale map of physical interactions can for example be constructed by systematically assessing direct interactions between molecules. For instance, in the case of protein complexes, protein-protein interactions (PPIs) can be screened by employing high-throughput yeast2-hybrid (Y2H) experiments or, as an alternative, affinity purification followed by mass-spectrometry (AP-MS) [66]. In another example, chromatin immunoprecipitation followed by sequencing (ChIP-seq) [67] has often been applied to construct global interaction maps of a protein (e.g. a specific transcription factor) with DNA sites (protein-DNA interactions). ChIP-seq has been applied in numerous contexts to study the activity of transcription factors or histone modifications [48, 60]. In our case, we make heavy use of the ReMap database of ChIP-seq derived transcription factor binding sites (TFBS) [71] and chromatin activity states derived from histone ChIP-seq data (chromHMM states) [48, 49]. An experiment similar to ChIP-seq, called cross-linking immunoprecipitation (CLIP-seq), can further be used to identify protein-RNA interactions [65, 68]. In addition, to probe direct interactions between DNA or RNA sites (DNA-DNA or RNA-RNA interactions, respectively), high-throughput Chromosome Conformation Capture (Hi-C) [69, 72, 73] or RAP-RNA sequencing [70] can be employed. We utilize published Hi-C data [74, 75] and derived associations between distinct DNA sites in this thesis to corroborate results from a large-scale functional association analysis of genetic variants with DNA methylation (Chapter 4). Finally, functional (indirect) interactions between molecules can be derived using experiments such as synthetic genetic array screens to obtain genetic interactions [76]. Another option is to apply computational approaches such as co-regulation (e.g. based on ChIP-seq) or co-evolution [77, 78].

One shortcoming of most of the experimental protocols, is, that they do not scale well to a broad range of biological contexts since they for example have to be performed in non-physiological conditions (Y2H) or are limited to a one-to-many interaction map (ChIP-seq, CLIP-seq, AP-MS). In contrast to these experiments, protocols to measure the global omics profiles in arbitrary biological contexts have been established, such as microarrays or high-throughput sequencing protocols (see also Section 2.1). As these can be applied to a large number of samples in arbitrary physiological conditions with relative ease and low cost, they enable the use of statistical methods to obtain molecular associations between individual molecules. Moreover, it is possible to combine readouts of different omic layers (e.g. the genome, methylome or transcriptome) if these are available for the same set of samples ('multi-omics' data). These data enable inference of interactions across omic boundaries, thereby generating a near to complete view on the studied system. This can be extended to inferring comprehensive regulatory networks from functional data as we will describe in the next sections.

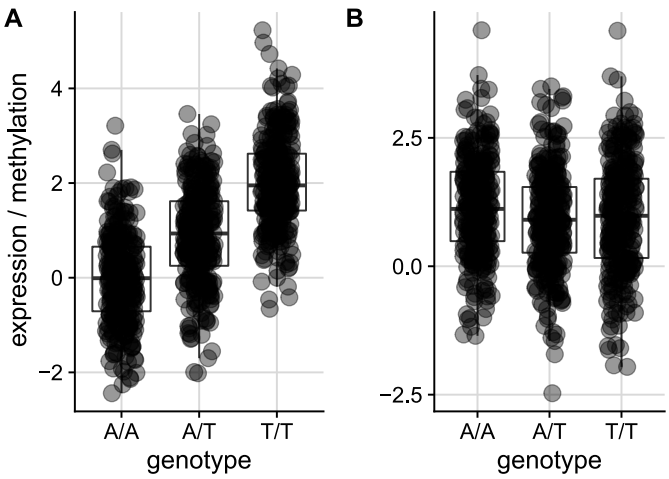


Figure 1.3.: Intuition behind expression/methylation QTL. Panel A: The genotype of a SNP (x-axis) determines the values on the y-axis such as the expression of a gene (eQTL) or methylation at CpG sites (meQTL) in a set of individuals. Panel B: The SNP genotype does not have a clear effect on the quantitative trait. Each dot represents a single simulated individual, boxplots show medians and lower/upper quartiles (horizontal lines) and 1.5 * inter-quartile range (vertical extensions).

1.4.2. Quantitative trait locus studies

By using intermediate molecular phenotypes, systems genetics studies seek to explain disease-associated, genetic variants in non-coding regions of the genome [11]. Throughout this thesis, we adopt this view and make extensive use of quantitative trait loci (QTL), which enable additional functional insights into the effects of GWAS variants. To this end, genetic variants are associated in a pairwise interaction approach with an intermediate molecular phenotype such as the expression level of genes (expression quantitative trait loci, eQTL) or the DNA methylation at CpG sites (methylation quantitative trait loci, meQTL) which are viewed as quantitative traits [compare e.g. 22, 23, 50, 79–82, as examples for meQTL and eQTL studies], rather than, or in addition to, being associated with a population level phenotype. Similarly to a GWAS, in a QTL study genotypes are assessed and molecular profiles generated for a large number of individuals and associations are determined by applying e.g. linear modeling, regressing the quantitative traits (dependent variables) against genotypes under consideration of available covariates (independent variables) (see Section 3.1.2).

Figure 1.3 shows the intuition behind QTL studies: The population is stratified by their respective genotype (for a single locus) and the resulting pattern of the quantitative trait (e.g. expression of gene X) is analyzed. If the extent of the trait is significantly different between the groups implied by the genotype (panel A), the latter is a QTL for the specific trait and it is not, otherwise (panel B).

While this approach necessitates the measurement of quantitative molecular data (see Section 2.1.3 and 2.1.2), it offers the opportunity to set trait-associated genetic variants into a functional context and, thereby, advance our understanding of how genetic effects propagate through the cell to form complex traits. Moreover, due to the drop in costs for measuring functional genomics data during the last decade, the additional integration of these data with GWAS results has become relatively easy to perform.

An important finding of QTL studies are *trans-QTL hotspots*, genetic variants on a

specific chromosome that influence multiple molecular phenotypes on other chromosomes [83]. A *trans*-QTL hotspot represents a coordinated genome-wide effect of a single genetic locus on numerous traits, such as transcript or protein levels, and are therefore of particular scientific interest as they represent genomic master regulators. To understand the changes the hotspot variant exerts in a cell, we need to investigate the underlying regulatory mechanisms by which the observed *trans* associations are implemented on a molecular level.

Promising first steps have been taken in this direction. For instance, the study by Bonder et al. investigated the effect of disease-associated genetic variants on gene expression and DNA methylation in whole blood [23]. After generating genome-wide eQTL and meQTL, the authors associated DNA methylation with gene expression to obtain targeted expression quantitative trait methylation (eQTM) results for their meQTL. Moreover, they integrated *trans*-meQTL with eQTM and additional TFBS from ChIP-seq data, thus establishing regulatory relationships spanning DNA, gene expression, protein expression, and DNA methylation. They found that disease-associated genetic loci lead to alterations in DNA binding of TFs (protein-DNA interactions) and DNA methylation changes, which subsequently mediate changes in gene expression networks. Bonder et al. generated novel disease-related hypotheses in the form of a gene regulatory network for a locus associated with ulcerative colitis. They highlighted a regulatory cascade, by which the genetic variant located in the first intron of *NFKB1* influences the gene's expression, in turn leading to altered methylation at distal DNA methylation sites which ultimately leads to change of expression of genes located in the vicinity of the CpG sites. Thus, the study established regulatory interaction maps for specific loci, such as the *NFKB1* locus, and generated new hypotheses of the underlying molecular mechanisms driving diseases and other phenotypes.

A more holistic approach is taken in a study by Suhre et al. [29]. Here, the authors first generated *trans*-pQTL (protein expression QTL) for a set of disease-associated genetic variants and then linked *trans* traits to the SNP by subsequently constructing a protein-protein interaction (PPI) network based on a targeted protein expression assay [29]. They integrated their pQTL for GWAS SNPs with the PPI network by including DNA-protein edges for all pQTL, thus establishing a disease context for their obtained networks. Using this technique, the authors generated novel disease insights for Alzheimer's Disease (AD). They inferred a thus far unknown relationship between a major AD risk variant (*rs4420638*) and splicing related proteins, hence elucidating molecular mechanisms underlying AD.

In another study by Võsa et al. [22], the authors established *trans*-eQTL for a total of 3,853 unique SNPs and 6,298 unique genes and used their findings, to investigate the functional consequences of GWAS SNPs on the expression patterns of genes. For their analysis, they integrated genotype and gene expression data to pinpoint local effects of variants on *cis* genes and subsequently analyzed TF-DNA binding sites at *trans*-eQTL genes using available ChIP-seq data. To corroborate the functional relevance of eQTL, they integrated their results with putative enhancer-promoter interactions

derived from Hi-C [75] to determine direct DNA-DNA contacts for distally related entities (SNPs and gene distance $> 100kb$). The authors estimate, that for approximately 17.4% of the identified *trans*-eQTL, the genetic effect could be explained by a direct interaction between a transcription factor (TF) encoded at the genetic locus and the respective *trans* gene. Importantly, their results indicate, that for the vast majority of loci a mechanistic explanation is still lacking.

In this thesis, we extend upon approaches such as the ones by Bonder et al., Suhre et al. and Võsa et al., which mostly looked at short paths through regulatory networks (e.g. direct explanation of *trans* effects through a TF encoded in *cis* of the variant) and tackle the issue of missing mechanistic explanations of *trans*-acting genetic loci. We address these issues by systematically reconstructing functional interaction networks for *trans*-QTL hotspots, specifically looking at regulatory cascades with paths longer than one, to further our understanding of complex traits. Moreover, we propose a fully integrated, genome-scale inference approach for multi-omics data, in contrast to the step-wise integration performed e.g. by Suhre and colleagues.

1.4.3. Biological network inference in systems biology

Systems biology seeks to generate holistic views on cellular systems [64], however, complex network models to explain, for instance, important *trans*-QTL hotspots are still lacking. In this section, we will describe general approaches to inferring interaction networks from biological high-throughput data, which we will tailor in the later parts of this thesis to fit an application on *trans* hotspots. Generally, such networks represent interactions between specific molecules of distinct cellular omic layers and are made of nodes (or vertices), which represent the individual bio-molecules, and links (or edges) between these nodes, which can reflect either direct physical or functional relationships.

In the previous section, we introduced approaches that perform step-wise integration of omics data in order to establish molecular interactions across multiple omics levels. While these approaches have been applied to great success [22, 23, 29] and are relatively straight forward to execute, simultaneous integration of multiple omics layers to construct heterogeneous (i.e. containing molecules from different omics layers) networks could exploit omics data to their full potential, e.g. by taking into account information from all available variables and omics data at the same time [32]. Hence, simultaneous integration approaches represent promising tools to further our understanding of how information is processed in a cell and, therefore, to explain trait-associated variants. A popular way to approach this idea is by employing so-called graphical models (see Section 5.2.5) and specifically Gaussian graphical models (GGMs). Gaussian graphical models are often preferred to pairwise integration approaches [29, 84, 85], and their extensions for applications to multi-omics data and integration of biological prior information to utilize established biological knowledge (e.g. PPI networks or public functional association data) are particularly promising [3].

For instance, Krumsiek et al. [84] applied GGMs to infer a metabolite reaction network

using a large-scale metabolite dataset. Although their analysis was based on a single omics layer (metabolome), they were nevertheless able to highlight the advantage of simultaneous network inference approaches over pairwise approaches by comparing their network to established metabolic reactions from the Kyoto Encyclopedia of Genes and Genomes database (KEGG, see Table 1.2). Their approach allowed the authors to propose additional direct associations between lipid metabolites, which, up until then, were only indirectly associated in the KEGG database.

For the inference of heterogeneous networks, methods that have previously been designed for single omic layers, such as the *graphical LASSO* [86, 87] or GENIE3 [88] for gene expression data, could be applied, but first need to be evaluated and benchmarked in multi-omics settings. Yet, *GENIE3* is the best performing method in two past DREAM network inference challenges (DREAM4/5, [89, 90]), and *GENIE3*, as well as other tree-based methods, hence are promising approaches for multi-omics network inference[2].

In a recent study by Saha et al. [85], the authors provide an example of heterogeneous network inference based on a *graphical LASSO* based algorithm [92], using GTEx (see Table 1.2 and Section 2.3) [82, 93] gene expression data to reconstruct transcriptome-wide (TWNs) and tissue-specific (TSN) networks. Although their work is based on a single omics level measure (i.e. gene expression), the authors quantify total expression (TE) and transcript isoform ratios (IR) for all genes based on the GTEx RNA-seq data and infer networks containing both TE and IR nodes effectively producing heterogeneous networks. Their approach allowed them to investigate splicing control mechanisms, for instance by analyzing TE-IR interactions which indicate likely splicing regulators. Furthermore, an interesting detail of their analysis is the construction of different LASSO penalties (see Methods 3.1.3) for distinct types of edges (TE-TE, TE-IR or IR-IR), by which they encode prior assumptions for observing these edges (see also Section 1.4.4 for an introduction to prior based inference). Application of their strategy allowed the authors to pinpoint specific splicing regulators across GTEx tissues, entailing known ones such as *RBM14* or *PP1R10* in addition to novel ones such as *TMEM160*. Further, they identified tissue specific regulators (e.g. *TTC36* in breast-mammary tissue), which might be crucial to understand disease-related regulatory pathways. This is further exemplified by their finding of *MAGHO* and *MAB21L1* as important hub genes, i.e. genes with a large number of edges, in brain-caudate and artery-aorta TSNs, as both genes have previously been found to be important for tissue-specific transcriptional regulation and to be essential in tissue development.

The idea of simultaneously integrating multi-omics data to infer regulatory networks is relatively novel, and hence specialized approaches such as mixed graphical models (MGMs, variations of GGMs allowing for other than Gaussian variable distributions) are often presented in the form of proof of concept studies on simulated data [94, 95].

[2] Also for application to large single-cell datasets, e.g. *GRNBoost* in the *SCENIC* workflow [91]

Although results on simulated data suggest that MGMs perform well, additional studies on real-world data are crucial to make use of their potential. Some works set out to make use of mixed models to infer regulatory networks, and particularly interesting in this context is the inclusion of phenotypes during network inference.

For example, Fellinghauer et al. [96] proposed a tree-based inference method they termed 'graphical random forests' (*GRaFo*), which was used by Zierer et al. [97] in a large multi-omics study to assess age-related disease comorbidities in the context of glycomics, metabolomics, epigenomics, and transcriptomics data. Here, the authors established a heterogeneous network and identified for example urate as a key factor linking metabolic syndrome phenotypes to renal function and body composition.

The integration of disease phenotypes together with genetic predisposition and other clinical data has also been investigated by Mohammadi, Abegaz, Heuvel, and Wit [98] using a Bayesian approach for graphical modeling. They focused on Dupuytren disease, a disease which affects finger contractures, and applied their Bayesian approach *BDgraph* to pinpoint disease indicators and assessments of disease severity in conjunction with 13 distinct possible risk factors. In their study, the authors did not have (multi-)omics data available, however, they nevertheless showcased the advantage of heterogeneous network inference in a clinical context to shed light on disease pathogenesis. Based on the family history of study participants, they corroborated a likely genetic risk to develop the disease and identified several key indicators with a direct effect on disease severity, including for instance alcohol consumption and age. Moreover, they were able to propose an improved therapy for individuals affected by this disease. Based on their finding that disease severity for individual fingers is correlated, they suggest executing surgical procedures simultaneously for affected fingers as compared to individual treatment as has been done before.

Genome-wide interaction networks, ideally heterogeneous ones spanning multiple omic layers, are crucial analysis tools for systems biology [11]. However, the inference of such networks is still challenging, especially in a large-scale context, and new methods are needed in order to do this successfully [99, 100]. In this thesis, we tailor established inference methods to the context of *trans*-QTL hotspots to alleviate network inference for complex trait associated genetic loci (Chapters 4 and 5).

1.4.4. Leveraging biological prior knowledge for network reconstruction

In Chapter 5, we aim to alleviate network inference from multi-omics data by using comprehensive prior knowledge about interactions. In the past, numerous large-scale studies generated functional data and annotation databases for human and other organisms (see Table 1.2 for a non-exhaustive overview), which can serve as prior information. These databases contain annotations of curated pathway (or network) information for diverse biological systems (e.g. KEGG, STRING, or BioGrid) on the one hand and rich functional omics data collected over a large number of samples (e.g. GTEx, Roadmap

Epigenomics, ENCODE) on the other hand. Pathway information are often deposited in a context-independent manner, i.e. they are often not tissue or cell type specific, whereas functional data inherently exhibit these properties. In general, these data, which are often publicly available, can be used to facilitate genomics analyses and already have been put to great use in genomics studies [33, 85, 101, 102].

resource	data type	organisms	reference
STRING	P-P[1]	$>$ 2000	[103, 104]
BioGrid	P-P	$>$ 60	[105]
inBio map	P-P	HS	[106]
GWAS catalog	D-PH	HS	[107]
GWAS atlas	D-PH	HS	[108]
PhenoScanner	D-PH	HS	[109]
KEGG	multiple	$>$ 5000	[110]
APID	P-P	$>$ 400	[111]
doRINA	P-R, miR-R	HS, MM, DM, CE	[112]
REMAP	P-D	HS	[71]
IntAct	P-P[2]	multiple	[113]
Pathway Commons	multiple	multiple	[114]
AGRIS	P-D	AT	[115]
ENCODE	G, T, E	HS	[60]
modENCODE	G, T, E	DM, CE	[116]
GTEx	G, T	HS	[82, 93, 101]
ROADMAP	E, T	HS	[48]
GEO	G, T, E	multiple	[117, 118]
ARCHS4	T	HS, MM	[119]
The Human Protein Atlas	T, P	HS	[120]
MetaboLights	M	multiple	[121]
TCGA	G, T, E	HS	[122]

Table 1.2.: Overview on selected resources for molecular interactions and omics datasets. Data type column depicts either the type of interactions (e.g. protein-protein interaction, P-P) or the type of omics data available in the data collection. **Interactions:** M=metabolite, P=protein, D=DNA, R=RNA, PH=phenotype; **Organisms:** HS=H. sapiens, AT=A. thaliana, MM=M. musculus, DM=D. melanogaster, CE=C. elegans; **Omics:** G=genomic, E=epigenomic, T=transcriptomic; Table adapted from Hawe, F. J. Theis, and Heinig [3].
[1] includes functional interactions
[2] focus on P-P, but arbitrary interactions possible

For instance, Saha et al. [85] used GTEx RNA-seq data to derive tissue context specific gene regulatory patterns (see Section 1.4.3). These data can also be used to pinpoint causal non-coding DNA variants derived from GWAS, e.g. by integrating GWAS results with interaction data to derive the causal mechanisms underlying certain phenotypes [29, 123]. In addition, tissue specific data collected in resources such as GTEx or ARCHS4 can also serve to unravel tissue specific effects of genetic variants [19, 93].

An interesting use of such data is through the derivation of biological prior information to alleviate the inference of regulatory networks from omics data and several methods have been proposed to achieve this. These include methods based on the graphical LASSO (e.g. *dwgLASSO* [31]), tree-based methods (e.g. *iRafNet*, *piMGM* [32, 33]) and Bayesian methods (e.g. *BDgraph* [98, 124]). By using and adapting these methods, as we set out to do in this thesis, large-scale datasets containing massive amounts of interactions or multi-omics data become important assets to infer disease-relevant regulatory interaction networks.

Given the large amount of interaction databases available, several studies set out to include some of these data to improve network reconstruction. For example, networks have been inferred by weighting in interactions derived from a given multi-omics dataset based on whether or not the interaction has previously been identified [31, 125].

In the study by [33], the authors proposed *piMGM* (prior incorporation Mixed Graphical Models), which they developed as an extension to *CausalMGM* [125, 126]. Briefly, the method independently applies *CausalMGM* for a predefined range of regularization parameters (see Section 5.2.5) on random subsets of the available samples and subsequently aggregates all generated models to construct a final 'stabilized' graph. Similar to LASSO based prior approaches (e.g. [30, 31]), *piMGM* incorporates priors derived from pathway knowledge to guide network inference. To evaluate their approach, they applied it to *TCGA* RNA-seq and cancer sub-type information to predict sub-types, utilizing priors curated from KEGG. *piMGM* successfully reconstructed known pathways (e.g. the *Notch* signaling pathway) and highlighted the most crucial pathway structures for breast cancer subtyping.

In another study, Zhu et al. [28] tackled the inference problem from a slightly different angle, employing a Bayesian Network approach based on Markov-Chain-Monte-Carlo sampling [127] to recover directed regulatory networks in yeast. Though applied only in a model system, the authors derived a stable causal network from a total of 1,000 sampled networks (edges present in $\geq$ 30% of the networks) and what is even more, they determined the direction of edges by including prior information based on curated PPI, TFBS, and eQTL. In order to demonstrate the validity of their network, they used gene knockout data to predict the downstream effects of controlled changes to the biological system. With their results, they supplemented existing yeast PPI resources with novel gene interactions and highlighted novel causal regulators of eQTL hotspots (see also Section 1.4.2).

Overall, although several studies investigated how to incorporate priors during network inference [e.g. 28, 30–32, 102, 124, 128, 129], most studies focus on synthetic datasets to show the general advantage of priors [31, 129, 130] or use model systems [28, 32, 129, 131]. Relatively fewer works describe the application of prior based network inference to functional omics data in humans [102, 124, 132, 133]. If human data are

considered, however, the inference is either limited to a specific pathway [133], only cell line data are used [132], or it is a case study in which no informative priors are applied [124]. One notable exception is the study by Zuo, Y. Cui, Yu, et al. [102], where priors derived from the STRING PPI database were used in conjunction with an inference model [31] to human cancer gene expression data, with a focus on analyzing differential gene expression.

Thus, existing methods need to be improved and ideas extended and brought to new biological contexts, in order to make full use of the potential of established biological knowledge in human contexts. In this thesis, we approach the challenge of curating comprehensive sets of biological priors from the massive amounts of available data to make them available to computational models for network inference. This entails adding complex information from functional databases, such as chromatin conformation data, gene expression or DNA accessibility data, to e.g. established PPI networks to formulate reliable prior beliefs. Here we would also like to note, that in recent years experimental techniques to e.g. directly determine protein-metabolite interactions [134] or to generate global protein-RNA interactions [68] have been introduced. These can be useful to increase the quality of existing interaction datasets, which, in turn, could alleviate prior-based network reconstruction efforts.

1.5. Reproducible research

A full discussion of all the aspects of reproducible research is beyond the scope of this thesis, however, all software used in its context has been implemented under the light of reproducibility, with the aim to create robust and reproducible computational workflows. In this section, we give a brief overview of what reproducibility in computational biology entails[3].

Reproducible research (RR) is a topic that has been discussed frequently throughout the scientific community [135–140]. For instance, missing reproducibility of research has been shown to be a severe issue in cancer research: In a study by Begley and Ellis [141], the authors found that only the results of about 11% of cancer hallmark papers can be reproduced independently. This is an alarmingly low number, especially since thousands of papers are published in the cancer context each year [142]. In another study conducted by the Open Science Collaboration [143], the authors set out to reproduce results from 100 studies conducted in the field of psychology, spanning three distinct journals. The authors focused on reproducing statistically significant results using the original data where available. In their analyses, they achieved successful reproduction of the original results in only 39% of the tested cases. Finally, in a study by Prinz, Schlange, and Asadullah [144], the authors surveyed 67 published studies across the fields of oncology, cardiovascular disease, and women's health using in-house data. Here, they identified inconsistencies for 65% of the studies and found that in-house data are completely in line with published results in only about 21%.

[3]Note, that for full reproducibility other areas such as e.g. wet-lab biology need to be addressed, too.

One has to clarify, though, that non-reproducible studies do not necessarily imply that the original findings are false, but merely that they cannot be reproduced, implicating potentially unreliable conclusions. Therefore, reproducible results build up trust and confidence in published findings and can be built upon in future studies [140, 145].

The findings highlighted above show the importance of teaching and implementing reproducible research. Therefore, large collaborative and open-source efforts are currently in progress, which seek to educate about the importance and implementation of reproducible research (e.g. The Turing Way[4]). These efforts can help avoiding (or mitigating) a reproducibility crisis [146] and we aimed to add to these efforts by providing fully reproducible workflows.

1.5.1. Reproducibility in computational biology

Reproducible research in computational biology entails that a set of results published in an original publication can be fully reproduced, given only the original data as well as the description of computational steps executed to arrive at those results. While this definition might be somewhat simplified (compare e.g. *The Turing Way* for a detailed discussion about definitions of nomenclature related to RR), it brings with it two absolute requirements, namely

1. the availability of the original data and
2. minute documentation of all computational steps

Already, item (1) can be a major bottleneck for reproducible research, especially in case the data are gathered from human subjects and therefore are highly sensible and subject to strict data protection laws. In that case, researchers might not be allowed to deposit the collected data in an online, publicly accessible repository. Initial solutions to this problem, such as de-identifying the data, are to be regarded critically when it concerns sequencing data from which the genetic makeup of an individual can be derived, as these data typically contain enough information to allow identification of the donor even if the data are de-identified [147]. Moreover, this issue is especially critical since these data also give further insights on direct relatives of the original donor [148].

In contrast to the data, documentation of the computational steps performed to arrive at the published results, as stated in item (2), is always possible and should, ideally, automatically arise directly from a well written and organized *workflow*. This documentation or workflow specification should include a full compilation of the software environment under which computations have been run, including operating system, software (e.g. R or python), and package versions.

While this might seem trivial at first, often such documentation is either not written from the point of view of an outsider (e.g. someone who wants to follow up the project) or does not contain all the needed information. For instance, certain information like

[4] `https://the-turing-way.netlify.app/introduction/introduction.html`

package versions, used parameters, or a specific data filtering step might not explicitly be written down, as it is self-evident to the developer, and hence is missing to fully retrace the analysis. Moreover, package versions might change during project development, for example when updating packages to their newest versions.

While, in principle, reproducibility can be achieved manually by keeping a clear and tidy documentation of the project at all times, it also poses the risk of an inconsistent documentation to workflow relationship, e.g. that a subtle change in the workflow is not reflected in the documentation. To solve this issue, in recent years systems have been developed which, amongst other things, seek to unite the documentation process with the specific implementation of a workflow.

A strong focus of the work leading to this thesis was on implementing reproducible workflows in all projects, specifically by using a workflow management system called Snakemake [149] in conjunction with containerization solutions such as conda[5] and Charliecloud [150] (see Section 3.4).

1.5.2. Dedicated workflow systems

Several dedicated systems are available which aim to simplify the process of creating reproducible data analysis workflows. Some of the more popular systems include *Galaxy* [151], *KNIME* [152], *Nextflow* [153], and *Snakemake* [149], all of which can be used to construct and publish reproducible pipelines. While *KNIME* focuses on providing a graphical user interface to facilitate its usage by less technical educated users, *Nextflow* and *Snakemake* are mostly code-based, i.e. the workflow is formulated similar to *Makefiles* in Unix[6]. In brief, all these systems define rules, which specify how inputs are processed and output files produced using a set of commands and parameters recorded in the rule definition. These rules can then be chained together by the respective input and output definitions such that complex workflows can be constructed, starting at the raw data as input for the first rule and ending with the last rule producing the desired results (e.g. a summary table or plot), with as many rules as needed in between. Using this concept, arbitrary workflows can be defined and run, which automatically document each step taken to arrive at a specific set of outputs given the respective input data.

1.5.3. Distributing workflows and software environments

An important aspect of reproducibility is the documentation of software and package versions. To facilitate this, and to enable straight forward distribution of workflows, most workflow systems can use well defined (static) software environments (software 'containers' or 'images') with the purpose of making the complete workflow self-contained (i.e. independent of software installed globally on the system the workflow is run on).

[5] `https://docs.conda.io/en/latest/index.html`

[6] Some graphical user interfaces have been developed by the community, e.g. `https://github.com/UMMS-Biocore/dolphinnext` for *Nextflow*

Several containerization solutions have been proposed, such as Docker [7], Singularity [154], or Charliecloud [150]. Containerization enables the user to build an environment from scratch, defining e.g. an operating system and specific software and package versions, which can then be used by workflow systems to execute code. Software and scripts can hence be executed within a closed environment, which facilitates repeating or moving the analysis to a different physical computer. For instance, Snakemake supports the specification of a Docker or Singularity software container for a workflow, which is initialized for each processing step to perform calculations. This also means that the complete workflow can be packaged and transferred to a different system, yet the software and package versions are the same as on the original system. This is a huge step towards reproducible research which can be implemented without much effort.

[7] `https://www.docker.com/`

2. Materials

In this chapter, we give a general introduction to the experimental data types and molecular data used throughout the thesis, with the structure as follows: In Section 2.1, we'll describe the microarray technology, a well established technology which has been used in diverse settings and organisms to generate molecular high-throughput data. We will explain the specialized microarray technologies employed in the context of this thesis to obtain different levels of molecular data in humans, i.e. genotype, DNA methylation, and gene expression data. Next, in Section 2.2 we provide a description of the human population cohorts we utilized, the microarray data of which were used to obtain molecular multi-omics profiles from thousands of individuals. Finally, we will detail the public data repositories we used, including databases for both functional multi-omics data and bio-molecular interactions (Section 2.3).

2.1. Microarrays: a cost effective way of obtaining multi-level functional data from large population cohorts

Generally, to generate molecular data a bio-sample from the individual of interest needs to be collected. For the cohort data used in this thesis, for instance, molecular data was generated from whole-blood samples as blood is easy to access and therefore makes population-scale studies feasible. Its relevance to immune system-related traits makes blood an interesting target tissue in epidemiological studies, not least by providing a link between the environment and diseases such as allergies triggered by an environmental factor. To determine the genetic makeup of an individual, i.e. the DNA sequence, it suffices to take a sample from any tissue of the body as the information encoded in the DNA is the same throughout the body. Depending on the type of analysis, however, certain studies might necessitate samples from a specific tissue. For instance, DNA methylation and gene expression can vary drastically between different tissues and cell types and in fact between cells from the same tissue at different cell stages (see Introduction). Therefore, ideally one obtains samples from a tissue related to the investigated trait, such as a heart sample in case of e.g. cardiomyopathy or other heart-related traits. Naturally, this poses an obstacle in human studies as specific tissues (such as heart) might not be easily accessible. In our studies, we rely on data generated from whole-blood samples collected in large population cohorts using microarray technologies. Microarrays are a cost-effective and reliable way to generate molecular readouts from diverse molecular layers (e.g. expression of genes or DNA methylation) for a large number of individuals [155, 156].

2.1.1. Determining the genetic make up of individuals through genotyping arrays

There are two dominant strategies for assessing the genetic variants in DNA samples from individuals: genotyping microarrays, typically used to detect common variants, and whole-genome sequencing, which also enables the discovery of rare variants. While whole-genome sequencing (WGS) offers the possibility to determine the complete genome sequence with high accuracy (depending on the sequencing coverage), it is also relatively expensive and hence can be infeasible to be applied in large population cohorts[1]. Genotyping arrays, on the other hand, assess only a relatively small fraction of genotypes of an individual's genome (typically below 1% [157]). For instance, the Affy Axiom 6.0 array assesses about 900,000 variants covering most of the genetic variants with a minor allele frequency of above 0.1 [158], and the Illumina Infinium Omni5Exome array covers approximately 4.3 million variants[2] (0.1% of the human genome). This shortcoming, however, can be amended by exploiting LD structure and known haplotypes (see Introduction), typically by employing genotype imputation from reference panels with WGS information available, making genotyping arrays a cheaper and comparable alternative to WGS. Indeed, modern microarrays have been designed to yield good imputation quality and utilizing imputation can increase the power of genetic association studies by up to 10% [157]. In the next two sections, we will cover the fundamentals of genotyping arrays and outline the basics of genotype imputation.

Genotyping arrays

Genotyping arrays are a form of DNA microarrays (DNA ChIPs) which are used to determine the genetic makeup of a particular biological DNA specimen. An array contains hundreds of thousands of *probes* (microscopic short stretches of DNA oligonucleotides, *oligos*), designed such that they cover, for instance, SNPs with known disease or trait associations, known exonic variants, or high-frequency polymorphisms [158]. The *oligos* represent a section of DNA complementary to the DNA sequence in the studied organism. There are different ways of how microarrays can be designed. Here we will consider two general approaches: hybridization based (e.g. Affymetrix arrays) and single-base pair extension technologies (used e.g. by Illumina). For hybridization-based arrays, multiple *oligos* of the same stretch of DNA are gathered at a single spot on the array. These spots partly contain the reference allele and partly the alternative alleles expected to be discovered in the organism. In the experiment, DNA is extracted from a specific biosample and broken up (digested) into small oligonucleotides, which are then hybridized under near-optimal hybridization conditions to the probes on the microarray (complementary base pairing of sample *oligos* to probes) [159]. Using e.g. fluorescent marks, it is then possible to determine the fraction of *oligos* matching different alleles, enabling the determination of the genotype of the sample of interest. For instance, if an

[1] Although costs have dropped significantly during the last decade [157]

[2] `https://www.illumina.com/products/by-type/microarray-kits.html`, last accessed 1/24/2020

approximate 50:50 ratio of the signals for reference and alternative alleles is observed, the specific locus is deemed heterozygous for this allele. Microarrays using single-base pair extension work similarly, i.e. providing multiple *oligos* per assayed genotype, but *oligos* are designed to stop one base-pair before the variant of interest. After the hybridization of fragmented sample DNA to the primer *oligos*, single-base pair extension of the primer is performed using e.g. fluorescently labeled nucleotides and a DNA polymerase. The fluorescent signal emitted by the included nucleotides can then be detected and used to determine the base at the position of interest. In our studies, we utilized genotype data from diverse microarray platforms, including e.g. the Affymetrix Affy Axiom 6.0 and the Illumina OmniExonExpress, and further employed imputation (see below) to obtain a more complete coverage of the genome.

Imputation of unassessed genotypes

Arrays, while relatively cheap, usually yield genotype readouts for only a fraction of the SNPs present in the genome and hence additional methods such as imputation of the remaining variants need to be applied to increase the power of genotyping studies [157]. These methods can also be employed to infer genotypes at positions for which the array could not determine a clear signal. Imputation makes use of general properties of genomes such as linkage disequilibrium, haplotype structure, and knowledge about recombination hotspots. The basic idea is that, generally, two individuals will share short stretches of DNA (haplotypes) originally derived from a common ancestor even if the two individuals seem not to be related on a first glance. Therefore, (sparse) genotype profiles obtained from an array experiment can be matched against a set of sequenced (non-sparse) reference genomes typically of the same ethnicity (e.g. using European reference samples for European study samples). By obtaining matched segments of individual DNA stretches from the reference population genotypes can be inferred for the study sample using probabilistic modeling. The latter is necessary since a single study haplotype could be represented by a multitude of reference haplotype segments. To this end, most imputation methods employ a Hidden Markov Model (HMM) to obtain allele probabilities for study samples based on a specific reference panel such as the 1000 genomes project [160], TopMED [161] or HapMap [162]. Popular imputation methods include e.g. Beagle [163] and IMPUTE2 [164, 165], and most contemporary methods follow an HMM based imputation approach and are implemented to optimize computational efficiency. For instance, methods typically make use of pre-phasing, i.e. separate determination of haplotypes without direct imputation. Pre-phasing greatly facilitates computations and allows step-wise integration of more genotyping data, thereby enabling larger reference panels to be used to improve imputation accuracy [166]. Moreover, imputation servers have been made available, which can be used by researchers to perform imputations for study samples in a standardized and secure way [157].

2.1.2. Quantitative DNA methylation analysis using microarrays

Similar to genotyping approaches, protocols for DNA methylation assessment can both involve microarrays and next-generation sequencing. Here, we focus on DNA methylation profiling using microarrays as such data is utilized throughout this thesis. DNA methylation, however, cannot be assessed directly, i.e. it is not possible to exploit complementary base pairing in a direct manner, and hence DNA samples need to be pre-processed. Whole-genome bisulfite conversion of extracted DNA converts unprotected, i.e. unmethylated, cytosines into uracil ('U' nucleotide), and subsequent amplification of converted DNA yields thymidine ('T') bases in place of the original cytosine ('C') bases. This process can be exploited to determine at which positions in the genome methylation occurred by comparing the converted DNA to a reference. Similar to the genotyping arrays, DNA methylation microarrays such as the Infinium HumanMethylation450 BeadChip from Illumina contain thousands to hundred thousands of beads with attached oligonucleotides, designed to cover relevant regions of the genome such as gene promoters, exons or intergenic regions [167, 168]. To assess methylation at a particular locus, these oligos represent both unmethylated (i.e. 'T' nucleotides) and methylated loci (i.e. 'C' nucleotides) with respect to the bisulfite converted DNA. After bisulfite conversion, sample DNA is hybridized to the array and signals for methylated (M) versus unmethylated (U) CpG sites recorded. These signal values can be summarized in the form of β-values for each CpG site C, which are defined as follows [169]:

$$\beta_C = \frac{max(M_C, 0)}{max(M_C, 0) + max(U_C, 0) + 100} \tag{2.1}$$

Therefore, the β-value for each CpG reflects either no methylation ($\beta = 0$) or complete methylation ($\beta = 1$) for a specific CpG site C. The distribution of β-values typically shows two peaks, one around 0 and one around 1 indicating the unmethylated and methylated sites, respectively. Figure 2.1 shows the distribution of β-values in the KORA methylation data (see below) for 50 randomly picked individuals as an example.

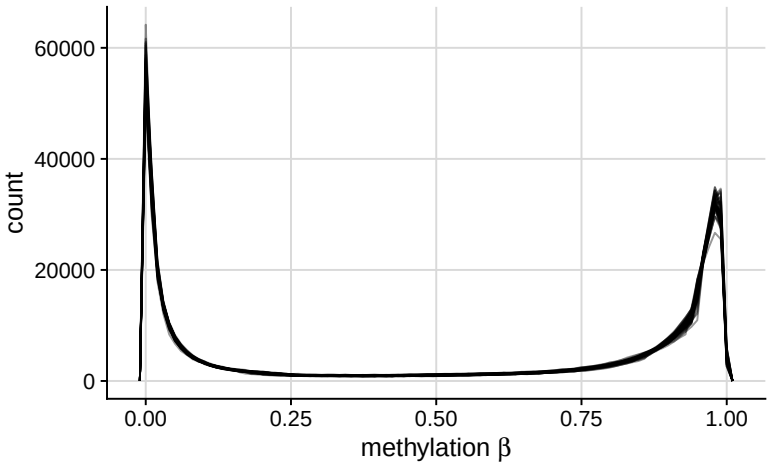


Figure 2.1.: Example distribution of β values obtained from the unmethylated and methylated probe signals of a methylation array. Data were obtained for all 485,512 available CpGs for a set of 50 individuals (individual lines) in the KORA cohort.

In addition to regular probes, which represent a specifically designed set of methylation loci of interest, control probes are included on arrays. These control probes can be used 1) to determine significant signals for individual probes ('probe calling') by estimating a background intensity distribution [e.g. 170] and 2) to perform background correction for CpG intensities [e.g. 171]. For instance, this can be achieved by regressing out principal components obtained from the intensities of negative control probes, e.g. probes designed not to match the human genome, from the CpG β-values [171], a strategy we also employed for the methylation data in the projects described in this thesis (see Section 3.2.2).

In this work, we mostly analyzed data obtained from peripheral whole-blood samples, which, though easily accessible, contain a mixture of diverse blood cell types where the ratio of cell types can differ between samples. The cell subset composition of each sample can confound statistical analyses as different cell types exhibit different methylation profiles [172]. The white cell subset composition of whole-blood samples can be estimated through methylation arrays, specifically through the application of the 'Houseman method' [172]. The Houseman method seeks to estimate the white blood cell proportions, entailing granulocytes, monocytes, NK cells, B-cells, and T-cells, utilizing a linear model based approach and assuming, that observed DNA methylation patterns are strongly correlated with the distribution of white blood cell types. We utilize Houseman white blood cell estimates obtained from our methylation data to correct for differences in cell-type composition between individuals in most of our analyses.

2.1.3. Using microarrays to quantify gene expression

DNA microarrays can also be used to assess transcriptional levels of expressed genes. Here, the individual probes on the arrays are designed to match only exonic regions of genes such that processed transcripts can hybridize to the probe. Experimentally, RNA is extracted from the sample of interest (in contrast to the DNA extracted for use with genotyping and methylation arrays) and reverse transcribed into complementary DNA (cDNA) which can then be hybridized to the probes on the array. Importantly, cDNA only contains information about the final transcription products after splicing. Prior to hybridization, cDNA is amplified and fluorescence labeled such that the amount of RNA in the sample can be estimated by detecting the fluorescent signal at individual probes of the array [173]. Similar to genotype or methylation arrays, gene expression arrays, such as the Illumina Human HT12 BeadChIPs used for the projects in this thesis, can measure the expression of thousands of genes simultaneously for a single sample. A shortcoming of microarrays is, that only RNAs that have been considered during their design can be detected. Next-generation sequencing based RNA assessment through e.g. RNA-seq would allow quantifying all sample RNA regardless of previous assumptions, similar to whole-genome sequencing, but is not yet used as the standard for large population cohorts. However, reduced costs, increased sensitivity, and other benefits make RNA-seq a sensible alternative to microarrays, which will potentially be replaced by this newer

technology [119].

The extraction of RNA from a biosample is a more intricate procedure compared to extracting DNA. For instance, RNA is less stable than DNA and degenerates relatively fast. Degeneration can happen partially during cell lysation, which introduces a bias when quantifying distinct transcripts [174]. Therefore, it is crucial to establish the quality concerning RNA degradation of the obtained RNA sample. The RNA integrity number (RIN) has specifically been designed for this [175]. It is based on an automatic assessment of features of an electropherogram contributing to RNA integrity such as the ratio of specific areas under the electropherogram curves against the total areas under the curve [175]. In our data, we used RINs to remove samples of low quality and as covariates for the association of gene expression with DNA methylation values (see Chapter 4).

2.2. Population cohort data used in this thesis

This section presents an overview of all cohorts, which data were used in light of this thesis. These data were collected independently of this thesis and made available to us by our collaboration partners at the Imperial College of London (Prof. J. Chambers) and the Research Unit of Molecular Epidemiology at the HelmholtzZentrum München (Dr. C. Gieger). A summary showing background for each cohort is given in Table 2.1.

For reference, we will cover the most important information concerning data collection which was provided by our collaboration partners and will refer to the respective original publications for specific details concerning e.g. experimental procedures for brevity. As indicated above, these cohorts are not targeted to a specific disease or trait and hence easily accessible whole-blood bio-samples were collected from the participating individuals. These samples were subsequently analyzed using microarrays to measure DNA methylation and gene expression and to determine genetic variants.

	EUR discovery	EUR replication				SA discovery	SA replication
Phenotypes	KORA F4	KORA F3	NFBC1966	NFBC1986	SYS	LOLIPOP disc.	LOLIPOP rep.
N	1,731	485	732	514	337	1,841	1,354
Sample	WBL	WBL	WBL	WBL	WBL	WBL	WBL
Country	Germany	Germany	Finland	Finland	Canada	UK	UK
Ethnicity	EUR	EUR	EUR	EUR	EUR	SA	SA
Design	Pop. based	Pop. based	Pop. based	Pop. based	Family	Pop. based	Pop. based
Age (yrs)	61 (8.9)	52.9 (9.65)	31 (0.3)	16.1 (0.4)	31.7 (32.5)	51.7 (10.1)	51.1 (10.1)
Sex (M)	49%	52%	44%	47%	48%	74.6%	45.7%

Table 2.1.: Overview over the cohorts used in the meQTL study. EUR = European; SA = South Asian; WBL = Whole Blood.

2.2.1. The Cooperative Health Research in the Region of Augsburg

The Cooperative Health Research in the Region of Augsburg (KORA) is conducted in the region of Augsburg in southern Germany and collects independent population-based health surveys together with subsequent follow-up examinations of subjects of German nationality. Surveys are cross-sectional and were assessed across all individuals in the population between 25 and 74 years of age. Specifically, four cross-sectional health surveys S1-S4 were performed at five year intervals, with individuals randomly sampled from the region of Augsburg in a two-stage approach [176]. All surveys assessed baseline information regarding socio-demographic variables, risk factors (such as smoking and alcohol consumption) as well as medical and family history of chronic diseases and use of medications, in addition to a standardized medical examination [176]. For the S3 (conducted in 1994/1995) and S4 (conducted between 1999-2001) surveys, consisting of 4,856 and 4,261 participants, respectively, follow-up examinations yielded the F3 and F4 KORA survey data entailing 2,974 and 3,080 participants. An additional follow-up study of the KORA S4 ('FF4') entailed a total of 2,279 participants. At all examinations, i.e. at the initial assessment and the follow ups, anthropometric and clinical variables were assessed. Both surveys are independent of each other and do not overlap with respect to individuals and no population stratification could be detected in previous publications [177, 178].

The work presented in this thesis is based on a sub-sample of 1,731 participants of KORA F4 and 485 participants of KORA F3 with methylation and genotyping data available. For genotyping individuals the Affymetrix Axiom platform and software were used. The Illumina Infinium HumanMethylation450K BeadChip was used to generate DNA methylation profiles in KORA F3 and F4. For KORA FF4 methylation data were measured using the Illumina EPIC array (N=1,848 individuals). In addition, gene expression data were measured using the IlluminaHT-12 v3 BeadChip in the F4 cohort and made available for a total of 1,091 individuals. For 681 individuals in KORA F4 all three data modalities were available. Specifics concerning the experimental procedures, including sample preparation, have been described elsewhere [81, 179, 180]. The ethics committee of the Bavarian Medical Association approved the studies and all study participants gave written informed consent.

2.2.2. The London Life Sciences Prospective Population Study

The London Life Sciences Prospective Population Study (LOLIPOP) is a prospective study comprising a cohort of 28k Indian Asian and European men and women who were recruited between 2003 and 2008 based on the registry of 58 General Practitioners in West London, UK [50]. All participants were examined with respect to cardiovascular and metabolic health at recruitment (including anthropometry) and blood samples collected to assess fasting glucose, lipid and insulin profiles as well as complete blood count and white cell counts. A total of 13,347 of all participants attended clinical follow-

up visits at which additional blood samples were taken and samples from both the initial assessment and the follow-up visits were stored for subsequent molecular assays (including genotyping and/or DNA methylation profiling) at $-80\,^{\circ}\mathrm{C}$.

In this thesis, we obtained DNA methylation profiles using the Illumina HumanMethylation450 array for a subset of these data of unrelated individuals comprising 1,841 South Asians with blood sample collected at enrolment and 1,354 South Asians with blood samples gathered at follow-up (SA discovery and SA replication cohorts, see Table 2.1). Individuals were further genotyped using a combination of genotyping arrays including the Illumina HumanHap300, Human-Hap610, OmniExpress and OmniExomeExpress arrays. In addition, gene expression data were generated using the Illumina HT-12 v4 BeadChIP (using the manufacturer's protocol) for a total of 975 participants including 816 South Asians and 159 Europeans. All participants gave written informed consent to be part of LOLIPOP and the study is approved by the National Research Ethics Service (07/H0712/150).

2.2.3. Northern Finland Birth Cohorts

We included both the Northern Finland Birth Cohort 1966 (NFBC66) and the Northern Finland Birth Cohort 1986 (NFBC86) in the meQTL project (see Chapter 4). The NFBC66 is a prospective follow-up study of children born in 1966 within the two northernmost provinces of Finland [181]. A total of N=8,463 individuals, who live in northern Finland or in the Helsinki area were invited for clinical examination. Of these, 6,007 individuals followed the invitation and attended the clinical examination at an age of 31 years. Blood samples were obtained and DNA extracted for a subset of 5,753 participants which are representative of the original cohort with respect to major environmental and social factors [182]. From the whole-blood samples DNA methylation was assessed using the Illumina HumanMethylation450 array and genotypes obtained using Illumina HumanCNV370DUO Analysis BeadChip for 807 participants who finished the study assessments.

In the NFBC86 study, individuals born between July 1st 1985 and June 30th 1986 in the provinces of Oulu and Lapland in Finland were included (N=9,203) [183]. Individuals living in the original target area or in the capital at the age of 16 were invited to a follow-up examination. A total of 7,344 subjects participated in the follow-up in 2001/2002 and of these 5,654 finished the questionnaire, the clinical assessment and gave a blood sample [184]. All samples were processed and genomic DNA extracted accordingly. For 566 individuals, DNA methylation profiles were obtained using the Illumina HumanMethylation450K array but 24 technical replicates were excluded from the resulting data sets. Individuals were genotyped using the Human OmniExomeExpress 8v1.2 BeadChip.

2.2.4. The Saguenay Youth Study

The Saguenay Youth Study (SYS) consists of a total of 1,991 individuals spanning two generations. It comprises N=1,029 adolescents and their N=962 parents and its overall goal is to investigate common cardiometabolic and brain diseases, specifically causal factors, early stages of development and trans-generational aspects [185]. Recruitment is based on a genetic founder population located in the Saguenay lac St Jean region in Quebec, Canada. All participants were extensively phenotyped, including recordings of blood pressure, serum lipidomic profiling and magnetic resonance imaging of brain and abdomen as well as detailed assessments of numerous other domains, involving cognition, mental health, diet, substance use, physical activity, sleep as well as family environment. For a total of 600 adolescents, genotypes were obtained using the Illumina Human610-Quad BeadChip, and for the remaining 424 adolescents and the 971 parents genotypes were generated using the Illumina HumanOmniExpress BeadChip. DNA methylation patterns were obtained for a subset of adolescents and their parents (N=132 and N=280, respectively) using the Infinium HumanMethylation450 array [186].

2.3. Public data

In this section, we give a brief introduction on the various public data used in this thesis. These entail large collections of molecular data sets or results derived from large-scale analysis and curation efforts and are established resources for computational biologists. Additional details on the data extracted from these databases will be provided in the individual project chapters where needed.

2.3.1. The Encyclopedia of DNA Elements

The Encyclopedia of DNA Elements is an established resource, which originally aimed to map all functional elements encoded in the human genome, defined as genomic regions encoding specific products (e.g. proteins) or exhibit specific biochemical patterns as for instance protein binding or defined chromatin structures [60]. To this end, the project initially analyzed 1,640 data sets in 147 cell types in order to annotate functional parts of all of the human genome, comprising ChIP-seq for diverse transcription factors and histone modifications, DNAse-seq, RNA-seq, and other experiment types across different cell-types and tissues and including experiments in model organisms (i.e. fly, worm and mouse). The current release contains data generated from 16,254 distinct experiments for all organisms and 10,475 human datasets, which are freely accessible via the ENCODE data portal[3].

[3] `https://www.encodeproject.org/`, last accessed 05/20/2020

2.3.2. The Genotype Tissue Expression consortium

The Genotype Tissue Expression (GTEx) consortium aims to generate a public resource for studying gene expression regulation in a tissue specific manner [93]. As of now, GTEx collected post-mortem samples from 54 tissues covering 948 healthy individual donors amounting to 17,382 sequencing datasets. For a total of 15,253 tissue-donor combinations, genotype data has been generated and expression quantitative trait loci (eQTL) in *cis* and *trans* have been calculated using a standardized pipeline for 49 distinct tissues [4] [82, 101]. Data from GTEx can be downloaded from their data portal at `https://www.gtexportal.org/` and an overview over the number of samples with genotype and gene expression data available is given in Figure 2.2. In our projects we mostly made use of the whole-blood data collected in GTEx, specifically the gene expression data and the calculated eQTL. We use expression data to filter protein-protein interactions from the STRING and BioGRID databases (see below and Chapters 4 and 5) and utilize eQTL to define prior information to guide network inference (Chapter 5). For the latter, we also use GTEx Skeletal Muscle data to derive mechanistic interpretations of a *trans*-acting genetic locus discovered in this tissue.

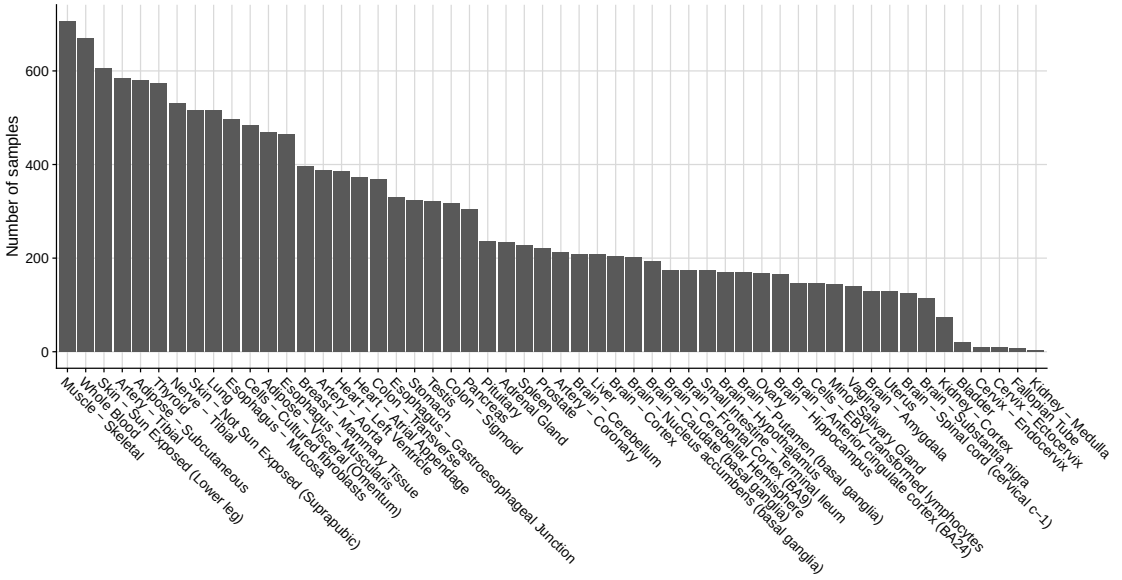


Figure 2.2.: Number of samples in GTEx which have genotype and gene expression data available for the distinct tissues.

2.3.3. The ARCHS4 database

ARCHS4 is a public data repository of uniformly processed RNA sequencing data from diverse human and mouse tissues [119]. Raw sequencing data were collected in a large

[4] https://www.gtexportal.org/home/tissueSummaryPage, last accessed 05/20/2020

curation effort from the Gene Expression Omnibus (GEO) [117] and the Sequence Read Archive (SRA) [187] and processed using a unified, cloud based workflow facilitating further use of these data in new experiments. A total of 187,946 samples have been made accessible in ARCHS4 upon publication (103,083 mouse and 84,863 human samples) [119]. In addition, the ARCHS4 web service[5] allows easy navigation, filtering (e.g. via collected meta data) and downloading of the available data. In our case, we used Skeletal Muscle gene expression data retrieved from ARCHS4 to formulate prior knowledge about gene co-expression in the network inference project (see Chapter 5).

2.3.4. The Roadmap Epigenomics project

The Roadmap Epigenomics project aims to provide a public resource of human epigenomics data, primarily focused around sequencing experiments to obtain, for instance, DNA methylation, histone modification, and chromatin accessibility information measured in primary human tissue and stem cells [48]. The current version of the resource comprises 2,804 genomic datasets, entailing a total of 1,821 histone modification, 360 DNAse-seq, 277 DNA methylation, and 166 RNA-Seq datasets from which a total of 150.21 billion sequencing reads have been mapped to the human genome[6]. Of the 2,804 datasets, 1,936 have been fully released and are divided in a total of 111 'reference epigenomes'. For each of those, a core set of five histone marks has been established on a genome-wide scale, including H3K4me3, H3K4me1, H3K27me3, H3K9me3, and H3K36me3. A widely used result derived from these histone marks are the cell-type specific chromHMM chromatin states [49]. ChromHMM states are derived from histone mark combinations using a Hidden Markov Model and provide a functional segmentation of the genome (by default in 200 base pair windows). We utilized the chromatin states obtained from the 15-state chromHMM model, which are available for each of the reference epigenomes on the Roadmap project's data portal[7]. An overview on these states including a brief description is given in Table 1.1.

2.3.5. STRING and BioGRID

The STRING database was established by the STRING consortium and makes known and predicted protein-protein interactions (PPI) publicly available [103, 104]. Interactions in STRING are derived from five main data sources including genomic context predictions, high-throughput experiments, co-expression analysis, automated text-mining and previous knowledge curated from other databases[8]. It contains over 2×10^9 PPI for over 5,000 organisms (mostly bacteria, including 477 eucaryotes), covering a total of 24.6×10^6 proteins. For the work described in Chapter 4, we curated STRING v9 PPI

[5] `https://amp.pharm.mssm.edu/archs4/`

[6] `http://www.roadmapepigenomics.org/data/`, last accessed 05/20/2020

[7] `https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html`

[8] `https://string-db.org/cgi/about.pl`

from human based on experimental and previous database evidence[9].

BioGRID [105, 188] is an international effort to establish comprehensive interaction networks for distinct organisms and follows a slightly different strategy to STRING, curating PPI only based on high-throughput experiment results and from individual studies (comprising over 70,000 publications[10]). At the moment of writing this thesis, BioGRID (v 3.5.185) contains a total of 1,871,024 genetic and physical protein interactions which are publicly available. We use PPI from BioGRID (v 3.5.166) to define locus sets and prior information in the project described in Chapter 4.

2.3.6. The ReMap resource

The (ReMap) database [71] provides transcription factor binding sites (TFBS) from numerous quality controlled and publicly available ChIP-seq data sets, uniformly processed to facilitate usage of TFBS across applications. Most data originated from ENCODE (see above), the gene expression omnibus (GEO) [117] or ArrayExpress [189], which represent some of the largest repositories for genomic data. Overall, the resource contains TFBS from 2,829 public ChIP-seq datasets comprising 485 distinct transcription factors measured over 346 cell types of diverse tissue origin[11]. The final set of regions in the genome that exhibit transcription factor occupancy amounts to a total of over 80 million peaks (TFBS) [71]. Of these, a total of 99.5% show a size of below 1.5kb with a mean size of 286 base pairs, in concordance with previous observations of overall narrow/sharp TF peaks [190]. For our analyses, we downloaded the set of merged (i.e. intersected) peaks for each transcription factor from ReMap. As we used this annotation together with whole-blood derived multi-omics data, we kept only results derived from experiments in a blood related cell-type/experiment including the terms listed in Table 2.2 below.

Table 2.2.: List of filters applied to the ReMap TFBS data set to obtain only blood related cell-type experiments.

amlpz12_leukemic	blood	bcell	bjab	bl41
aplpz74_leukemia	lcl	plasma	gm	hbp
lymphoblastoid	kasumi	k562	mm1s	p493
erythroid	sem	thp1	u937	

[9] obtained from `https://string-db.org/cgi/download.pl`

[10] `https://wiki.thebiogrid.org/doku.php/aboutus`, last accessed 06/24/2020

[11] `http://pedagogix-tagc.univ-mrs.fr/remap/`, last accessed 06/20/2020

3. Methods

In this chapter we set the methodological basis to discuss the work presented throughout the rest of the thesis. We will start of with a description of the statistical tools and concepts used, covering basic linear models and regularization, meta analysis, multiple testing and enrichment tests. This is followed by the description of the pre-processing of the available cohort and specific validation data used in our projects. Finally, we give a brief introduction to how we implemented reproducible and portable workflows using Snakemake and Charliecloud.

3.1. Statistical background

3.1.1. Conditional independence and correlation of random variables

Here we give a brief introduction to conditional independence and correlation between random variables as this is the basic concept behind the graphical models (also: conditional dependence graphs) applied in Chapter 5.

Independence and conditional probability of random variables

Let A and B be continuous random variables with $p(A)$ and $p(B)$ being their respective density functions. Also, let $p(A, B)$ be the joint probability density of A and B. The two variables are said to be *independent* ($A \perp\!\!\!\perp B$), if

$$p(A, B) = p(A)\ p(B)\ , \tag{3.1}$$

i.e. when their joint density is simply the product of the individual densities, and therefore, loosely speaking, knowing about one variable does not affect our knowledge of the other variable [191]. In addition, the *conditional* probability of A given C, i.e. the probability of A depending on C, is defined as

$$p(A|C) = \frac{p(A, C)}{p(C)},\ \forall\{C|p(C) > 0\}\ , \tag{3.2}$$

which can be extended to the conditional probability for the joint distribution of two variables (A and B) given C:

$$p(A, B|C) = \frac{p(A, B, C)}{p(C)},\ \forall\{C|p(C) > 0\} \tag{3.3}$$

We can expand the two concepts to form the idea of *conditional independence*, where we observe the dependency of two variables (e.g. A and B) given a third variable (C):

$$A \perp\!\!\!\perp B|C \Leftrightarrow p(A,B|C) = p(A|C)p(B|C),\ \forall\{C|p(C) > 0\} \tag{3.4}$$

It is straight forward to further extend conditional independence to more than one conditioning variable, as it is for instance necessary with the graphical models described in Chapter 5. Here, two variables are conditioned against all other variables present in a graph structure to estimate their conditional dependence.

Covariance and correlation of random variables

The covariance for two random variables A and B measures the linear relationship between the variables and is defined as

$$cov(A,B) = E((A - E(A))(B - E(B))) = E(AB) - E(A)E(B)\ , \tag{3.5}$$

where $E()$ is the *expected value* or *mean* of the respective variables. The correlation of two random variables represents a normalized measure of their relationship (normalized by the product of the variance $var()$ of the variables):

$$\rho = corr(A,B) = \frac{cov(A,B)}{\sqrt{var(A)var(B)}}. \tag{3.6}$$

The correlation ranges from -1 to 1 which indicate perfect anti-correlation or correlation, respectively. A correlation value of 0 indicates no (linear) relationship between the variables and the variables are said to be uncorrelated.

Interestingly, if A and B are independent (Equation 3.1) then both covariance and correlation are 0 and hence they are not correlated. In case A and B follow a multivariate Gaussian distribution, i.e. $(A,B) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (\mu_A, \mu_B)$ and $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} \sigma_{AA} & \sigma_{AB} \\ \sigma_{BA} & \sigma_{BB} \end{smallmatrix}\right)$, the converse is also true: If the variables show a correlation/covariance of 0 (e.g. $\sigma_{AB} = 0$) then they are independent [192]. This forms the basis for Gaussian graphical models, where the covariance structure indicates conditional dependencies between individual random variables (see Chapter 5). Briefly, in graphical models the inverse of the covariance matrix $\mathcal{P} = \Sigma^{-1}$, also called the *precision matrix*, can be used to determine *partial correlations* between all variables in the model, i.e. the correlation between variables conditioned on all other variables [193]. If an entry in the precision matrix is 0 the corresponding (normally distributed) variables are conditionally independent and vice versa, leading to absence or presence of edges in the graph structure, respectively. The concept of graphical models will be discussed in more detail in Chapter 5.

3.1.2. Linear models

In this section and throughout this thesis when we formulate a statistical model we will denote the number of samples with N and the number of variables going into the model

with P. We will indicate matrices with capital and column vectors with small letters in bold script (e.g. $\mathbf{X}$ or $\mathbf{y}$) and denote scalar values in normal script, typically indexed if they originate from a vector (e.g. y_i or β_1).

Linear models can be used to describe the (linear) effect of one or more predictor (or independent) variables $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P)$, with e.g. $\mathbf{x}_1 = (x_1, x_2, \ldots, x_N)$ being a column vector, on a dependent (or response) variable $\mathbf{y} = (y_1, y_2, \ldots, y_N)$. To put it into context, in our case N is the number of cases or samples for which measurements (e.g. of methylation at CpG sites) are available and P is the number of variables which have been measured (e.g. number of CpG sites), whereas $\mathbf{y}$ contains an outcome y_i for each sample (e.g. expression of a gene potentially influenced by DNA methylation). The linear model asserts that its output (dependent variables) is a linear function of the input (independent variables) and takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_P x_{iP} + \epsilon_i = \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} + \epsilon_i \ , \tag{3.7}$$

where the outcome $\mathbf{y}$ is modeled through an intercept term β_0 and through linear combinations of the independent variables in $\mathbf{X}$, and where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is the variable coefficient vector. It is often assumed that ϵ_i is an independent and identically distributed (*iid*) Gaussian error term with mean zero and unknown variance σ^2, i.e.

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \ . \tag{3.8}$$

The Gaussian error assumption also allows us to rephrase the model in the form

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2) \ . \tag{3.9}$$

The goal is then to determine the unknown variable coefficients (effect sizes, βs) and the σ.

Maximum likelihood estimation for linear models

We can estimate the parameters by applying maximum likelihood estimation (MLE) on the model described in Equation 3.9, also known as least squares regression.

For MLE, we typically formulate the log-likelihood of the model which is given via

$$\mathit{ll}(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} log \ p(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) \ . \tag{3.10}$$

In addition, we usually use and minimize, rather than maximize, the more convenient negative log-likelihood. Assuming a Gaussian model we now insert the definition of the Gaussian distribution in the log-likelihood, yielding

$$\mathit{ll}(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} log \ \left\{ (\frac{1}{2\pi\sigma^2})^{\frac{1}{2}} exp(-\frac{1}{2\sigma^2}(y_i - \boldsymbol{\beta}\mathbf{x}_i)^2) \right\} \ , \tag{3.11}$$

which boils down to

$$ll(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = -\frac{1}{2\sigma^2} RSS(\boldsymbol{\beta}) - \frac{N}{2} log(2\pi\sigma^2) \tag{3.12}$$

and where

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \boldsymbol{\beta}\mathbf{x}_i)^2 \tag{3.13}$$

is the residual sum of squares or the sum of squared errors. Therefore, the maximum likelihood estimate for β is also the one that minimizes the RSS which is why the MLE method is also known as least squares [191].

We then seek to minimize $RSS(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$:

$$argmin_\beta \sum_{i=1}^{N} (y_i - \sum_{j=0}^{P} x_{ij}\beta_j)^2 \ . \tag{3.14}$$

Hence, we minimize the error obtained when predicting $\mathbf{y}$ through the linear combinations involving the predictor variables and their coefficients (assuming a fixed σ in Equation 3.12). Switching to matrix notation and setting the derivative of $RSS()$ to zero, we obtain a closed form for an estimate of $\boldsymbol{\beta}$ (assuming full column rank of $\mathbf{X}$) as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \ . \tag{3.15}$$

Based on the estimate $\hat{\boldsymbol{\beta}}$, we obtain fitted values for our inputs in $\mathbf{X}$ as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \ . \tag{3.16}$$

Hypothesis testing

Finally, we seek to perform statistical inference with our model (Equation 3.9). Based on the assumption of Gaussian *iid* and zero mean error terms, the model allows for testing hypotheses about the values of the coefficients. Assume we want to evaluate a larger (i.e. more parameters) model Ψ against a smaller model ψ (i.e. some $\beta_i = 0$) with respect to whether or not the additional parameters in Ψ are necessary and where the variables in ψ are a subset of the variables in Ψ. Generally, we are interested in obtaining parsimonious models with low complexity, i.e. a small number of parameters. Therefore we set ψ to reflect our null hypothesis and Ψ to represent the alternative. How could we now compare the models? If we can formulate the likelihood of our data, we can employ the likelihood-ratio testing approach [194]. For instance, if the likelihood function of a model is $L(\boldsymbol{\beta}, \sigma|\mathbf{y}, \mathbf{X})$, then we can construct the likelihood ratio statistic as

$$LR = \frac{max_{\beta,\sigma\in\Psi} L(\boldsymbol{\beta}, \sigma|\mathbf{y}, \mathbf{X})}{max_{\beta,\sigma\in\psi} L(\boldsymbol{\beta}, \sigma|\mathbf{y}, \mathbf{X})} \tag{3.17}$$

Our test should reject the null if the difference (the ratio) is too large. Specifically, we find that we should reject the null if

$$\frac{\hat{\sigma}_{\psi}^2}{\hat{\sigma}_{\Psi}^2} > a \ , \tag{3.18}$$

where a is a constant and which is equivalent to (b being another constant):

$$\frac{RSS_{\psi} - RSS_{\Psi}}{RSS_{\Psi}} > b \ . \tag{3.19}$$

We can hence exploit the RSS (Equation 3.12) of the two models to identify whether the difference in their likelihoods is sufficiently small in order to favor the smaller model (ψ) (a model with more parameters will always fit the data better, yielding a smaller RSS). Formally, assume that the number of parameters are q and p for Ψ and ψ, respectively, then we obtain (if the null is true) the two independent quantities

$$\frac{RSS_{\psi} - RSS_{\Psi}}{q - p} \sim \sigma^2 \chi^2_{q-p} \tag{3.20}$$

and

$$\frac{RSS_{\Psi}}{N - q} \sim \sigma^2 \chi^2_{N-q} \ , \tag{3.21}$$

where the denominators are used for scaling purposes [194]. The ratio of the two independent χ^2 distributed quantities yields the F-statistic which follows a Fisher-distribution [192]:

$$F = \frac{(RSS_{\psi} - RSS_{\Psi})/(q - p)}{RSS_{\Psi}/(N - q)} \sim \mathcal{F}_{q-p,N-q} \tag{3.22}$$

Using this information, we can reject the null hypothesis if the probability of observing $\mathcal{F}_{q-p,N-q}(F) < \alpha$ for a significance threshold α, telling us that the difference in RSS is too large for the null model to be the better explanation.

The setup above also allows us to investigate whether a single parameter of the model can be dropped. Let's explicitly define the null hypothesis for this case, which would be given by

$$H_0 : \beta_j = 0 \ . \tag{3.23}$$

We could directly obtain the F-statistic to evaluate this hypothesis by specifying Ψ as the full model and ψ as the smaller model only missing β_j. As an alternative and equivalent test we can directly use the t-statistic

$$t_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{N-p-1} \ , \tag{3.24}$$

where $se(\hat{\beta}_j)$ is the standard error for $\hat{\beta}_j$ and $se(\hat{\beta}_j) = \hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}$ [194, 195], and where t_j follows a t-distribution with $N - P - 1$ degrees of freedom under H_0. We can use it similarly to the F-statistic to assess the significance for this test.

Relation of linear models and conditional probabilities

Interestingly, we can also form a relationship between linear models and the conditional probabilities introduced in Section 3.1.1. For instance, imagine we partition a set of multivariate normal RVs $\mathbf{Z}$ in $(\mathbf{z}_A, \mathbf{z}_B)$ and $\boldsymbol{\mu}$ in $(\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)$, then the conditional distribution $p(\mathbf{z}_A|\mathbf{z}_B)$ is given as (see [196] for details):

$$p(\mathbf{z}_A|\mathbf{z}_B) = \mathcal{N}(\mathbf{Z}|\boldsymbol{\mu}_{A|B}, \Lambda^{-1}_{AA}) \ , \tag{3.25}$$

where $\Lambda = \Sigma^{-1}$ is the *precision matrix* (see also Section 3.1.1) and where

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A - \Lambda^{-1}_{AA}\Lambda_{AB}(\mathbf{z}_B - \boldsymbol{\mu}_B) \ . \tag{3.26}$$

Now assume, that our $\mathbf{X}$ from the linear model also follows a normal distribution, i.e.

$$p(\mathbf{X}) \sim \mathcal{N}(\mathbf{X}|\mu_{\mathbf{X}}, \sigma^2_{\mathbf{X}}) \tag{3.27}$$

and combining this with the linear model formulation (Equation 3.9) we have

$$p(\mathbf{y}|\mathbf{X}) \sim \mathcal{N}(\mathbf{y}|\mathbf{C}\mathbf{X} + \mathbf{d}, \sigma^2) \ , \tag{3.28}$$

where $\mathbf{C}$ and $\mathbf{d}$ are parameters determining the mean [196]. This then directly corresponds to the conditional distribution for a multivariate normal.

Linear model residuals

The *residuals* for our linear model can be calculated using

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \mathbf{x_i}\hat{\boldsymbol{\beta}} \tag{3.29}$$

and reflect the deviation of our observed y_i around the fitted values. If the model assumptions hold then $\hat{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$ (this can also be used for diagnosing linear models, for instance by plotting residuals against the estimated/fitted $\hat{y}_i$ [197]). An interesting application of residuals is that they can be used to remove the effect of confounding variables or batch effects (e.g. sex or age) from measurement variables (e.g. DNA methylation or gene expression) which we employ throughout this thesis. The estimated β_js are the effect sizes for the individual variables $\mathbf{x}_j$ and $\hat{\mathbf{y}}$ effectively summarizes the effect of all variables given their estimated effect sizes. Therefore, by removing this summarized estimated effect from the observed effect, i.e. by obtaining the residuals $\hat{\boldsymbol{\epsilon}}$, only the effects not explained by the given variables remain.

3.1.3. Regularization in linear models

In the context of genomic data the $N << P$ problem is of great importance, where N represents the sample size (e.g. number of individuals) and P reflects the number of variables being analyzed. In typical genomics settings N will amount to several hundreds[1] whereas P is around tens of thousands (e.g. genes) or hundreds of thousands (e.g. CpG sites). In the case of $N << P$, linear regression systems are under determined and an exact solution cannot be obtained [3]. To amend this, different forms of regularization (or penalization) of the parameters (betas) of linear models have been proposed including for instance L_1 (LASSO) or L_2 (ridge) regression. Both L_1 and L_2 penalization yield relatively more non-zero β estimates as compared the standard linear regression [195].

Lasso regularization

Here, we focus on L_1 regression (LASSO, **L**east **A**bsolute **S**hrinkage and **S**election **O**perator), which has also been applied in the context of graphical models (see Chapter 5) and tends to yield sparser models compared to ridge regression [195]. For this, the regression equation for obtaining the parameter estimates is extended by an additional L_1 penalization term (L_1 norm, hence the name) which penalizes models including many non-zero variables [195, 198]:

$$\hat{\boldsymbol{\beta}}^{lasso} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 \,, \tag{3.30}$$

subject to

$$\sum_{j=1}^{P} |\beta_j| \leq t \,. \tag{3.31}$$

We solve the system by using the Lagrangian form of the above equation, i.e.

$$\hat{\boldsymbol{\beta}}^{lasso} = argmin_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\} \,, \tag{3.32}$$

where λ is a constant and $\sum_{j=1}^{P} |\beta_j|$ is the L_1 regularization term. While their is no closed form solution to this problem, efficient procedures to solve it such as path-wise coordinate optimization have been proposed [195, 199].

An important remaining task is to select the optimal shrinkage parameter λ. To this end, cross validation can be applied where, for example, the estimated prediction error of the model is minimized over a range of different λs [195]. We apply this approach for graphical model selection via the LASSO in Chapter 5.

[1] Recent single-cell assays can easily provide thousands of samples (cells) but are not used in this thesis

A Bayesian view on regularization

Interestingly, we can also adopt a Bayesian view on regularization. By assuming a prior distribution for the model parameters (βs), L_1 and L_2 regularization are equivalent to setting a Laplace or a Gaussian prior on the parameters, respectively [197].

Briefly, we consider a prior for the βs for the linear model formulation, i.e.

$$p(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) \propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathbf{X}) = p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})p(\boldsymbol{\beta}) \ . \tag{3.33}$$

So the posterior distribution for the linear model parameters is proportional to the likelihood times the prior and the equality in Equation 3.33 follows from the assumption that X is fixed [197]. Assume the linear model as in Equation 3.7 and independent, normally distributed errors. We can further assume that $p(\boldsymbol{\beta}) = \prod_{j=1}^{p} g(\boldsymbol{\beta})$ for some density function $g()$. Applying lasso (L_1) regularization, for instance, directly yields the posterior mode for $\boldsymbol{\beta}$ if $g()$ is set as a Laplace distribution with mean zero and an to λ (see Equation 3.32) proportional scale parameter [197]. In this thesis (Chapter 5) we make use of this Bayesian view in the graphical lasso [87] to apply prior knowledge on molecular interactions during network inference.

3.1.4. Meta analysis

Meta analysis is a tool to combine results from multiple (potentially low sample size) studies to increase detection power of associations by effectively also increasing the analyzed sample size for instance in genome- and epigenome-wide association studies [200]. Generally, one can distinguish between 1) fixed and 2) random effect meta analysis. In 1), the assumption is that the underlying true effect observed in the different data sets is the same and differences arise only due to sampling errors, whereas for 2), the assumption is that the effect size can differ between data sets in addition to random sampling errors being present [201]. In this work, we focus on fixed effect meta analysis, specifically inverse-variance weighted fixed effect meta analysis which we used in our studies around the LOLIPOP and KORA cohorts to establish (epi-)genome wide associations. In contrast to other fixed effect meta analyses, such as the Fisher method for combining p-values [202], the inverse-variance method takes into account sample sizes for individual studies by weighting regression estimates by the inverse of their variance [201]. For example, let β_S be the estimated regression coefficient for a study S and let

$$var(\beta_S) = se(\beta_S)^2 = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \tag{3.34}$$

be its variance and where the *residual standard error* σ is known and can be estimated from the data as $RSE = \sqrt{\frac{RSS}{N-2}}$ [197]. Then, weights are derived as $w_{\beta_S} = \frac{1}{var(\beta_S)}$ and

new (combined) regression estimates can be derived using

$$\beta_{\hat{S}} = \frac{\sum_{S=1}^{|S|} \beta_S w_{\beta_S}}{\sum_{S=1}^{|S|} w_{\beta_S}} \tag{3.35}$$

and

$$se(\beta_{\hat{S}}) = \frac{1}{\sum_{S=1}^{|S|} w_{\beta_S}} , \tag{3.36}$$

yielding a Z-score $Z_{\hat{S}} = \frac{\beta_{\hat{S}}}{se(\beta_{\hat{S}})}$ which follows a standard normal distribution under the null hypotheses of no effect [201]. We can hence compare $Z_{\hat{S}}$ to the quantiles of a standard normal to obtain significance estimates for our test, having more power to detect associations due to increased (combined) sample size [200].

3.1.5. Enrichment testing

Advances in high-throughput experiments and in annotating the human genome and its gene products with functionally relevant information, for example based on open or closed chromatin, DNA contacts or information about the regulatory function of genes, enable studies to assess new results in a functional context. For instance, in Chapter 4 we make use of these annotations to evaluate identified relationships between genotypes and DNA methylation with respect to their functional relevance.

Generally, enrichment tests can be performed for two distinct sets (or classes) of entities. One set (set $\mathcal{S}$) is derived from the experiment and entities can either be within that set ($\mathcal{S} = 1$) or not ($\mathcal{S} = 0$). Typically, one seeks to match the second set of entities considered for the analysis to the entities in $\mathcal{S}$ (according to some properties) to form a background representing a 'null' distribution. For instance, in Chapter 4 we define $\mathcal{S}$ as all CpGs associated with a genotype and the background is sampled from all non-associated CpGs, one for each associated CpG and matching mean and standard deviation of population methylation values.

In addition, we can obtain the functional annotation $\mathcal{A}$ of interest and compare, whether our entities in set $\mathcal{S}$ are more often annotated with $\mathcal{A}$ ($\mathcal{S} = 1$ and $\mathcal{A} = 1$) than the ones in the background set ($\mathcal{S} = 0$ and $\mathcal{A} = 1$). Here, $\mathcal{A}$ could for example be a functionally relevant region of the genome such as an 'enhancer' region.

Based on the above information we have four distinct scenarios based on the possible values for $\mathcal{S} = 0, 1$ and $\mathcal{A} = 0, 1$, and we can generate a representative contingency table:

	annotated with $\mathcal{A}$	not annotated with $\mathcal{A}$	**row totals**
entity in $\mathcal{S}$	n_{11}	n_{12}	$r_1 = n_{11} + n_{12}$
entity not in $\mathcal{S}$	n_{21}	n_{22}	$r_2 = n_{21} + n_{22}$
column totals	$c_1 = n_{11} + n_{21}$	$c_2 = n_{21} + n_{22}$	$RC = c_1 + c_2 + r_1 + r_2$

Here, each entry is the number of times we observe combinations of our classes, e.g. n_{11} is the number of times $\mathcal{A} = 1$ and $\mathcal{S} = 1$, n_12 is the number of times $\mathcal{A} = 0$ and $\mathcal{S} = 1$, etc..

Now it is of interest to assess whether the observed set derived from our experiment is significantly more (or less) often annotated with a functional annotation as compared to the matched background. For instance, we could compare the fraction of observed entities annotated with the annotation ($\frac{n_{11}}{r_1}$) with the fraction of background entities annotated with the annotation ($\frac{n_{21}}{r_2}$) and determine which of the fractions is larger. To formalize this comparison, we can employ for instance Fisher's exact test or Pearson's χ^2-test. Here, we'd like to assess whether the annotation definition $\mathcal{A}$ is independent of the set classification $\mathcal{S}$, so we investigate

$$H_0 : P(\mathcal{S} = i, \mathcal{A} = j) = P(\mathcal{S} = i)P(\mathcal{A} = j) \ , \tag{3.37}$$

where $i, j \in \{1, 2\}$ is the row or column index, respectively [192].

Fisher derived an exact way using a hypergeometric test to obtain the probability of getting the observed values assuming both classes are independent of each other:

$$\mathcal{P}(n_{11}) = \frac{\binom{r_1}{n_{11}}\binom{r_2}{n_{21}}}{\binom{RC}{c_1}} \tag{3.38}$$

As an alternative to Fisher's exact test, which should be applied to small sample cases only, one can use Pearson's χ^2 test for large samples sizes [192].

For a 2×2 contingency table as written above, with $m = 2$ columns and $n = 2$ rows, the test statistic for the χ^2 test is defined as

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(n_{ij} - r_i c_j/n)^2}{r_i c_j/n} \ . \tag{3.39}$$

The test statistic follows a χ^2 distribution with two degrees of freedom for the above example, from which we can then obtain the P-value under our H_0. More generally, the degrees of freedom are given by the number of categories for both classes, i.e. $df = (n-1)(m-1)$.

These tests provide us with a framework for determining for instance the functional relevance of our genetic variant-methylation association results in Chapter 4.

3.1.6. Multiple testing

In our analyses, we often perform large numbers of hypothesis tests, for instance to identify significant associations between genetic variants and CpG sites (several million tests). This results in a large *multiple testing burden* or *family-wise error rate* (*FWER*): The probability of obtaining at least one false positive test (type 1 error) at a specific

significance threshold α increases with each additional test [192]. Specifically, the probability increases to

$$\alpha^* = 1 - (1 - \alpha)^n \ , \tag{3.40}$$

where n is the number of performed and independent tests.

There are several ways to cope with this multiple testing problem by adjusting the significance level α, such that the probability of obtaining a false positive is kept low. For instance, the stringent Bonferroni correction controls the *FWER* (a family of tests is the set of all performed tests on the data) by setting a new significance threshold $\hat{\alpha} = \frac{\alpha}{n}$ to control for type 1 errors (the probability of obtaining a false positive result) [192]. Intuitively, Bonferroni adjustment makes use of the observation that the probability of at least one type 1 error occurring cannot be greater than the sum of the individual probabilities of all tests being false positives.

An alternative (and less conservative) way to the Bonferroni correction is the Benjamini-Hochberg false discovery rate (FDR) procedure [203]. While Bonferroni correction controls the probability of obtaining a single false positive, the FDR controls the fraction of false positive findings over the total number of positive test results [203]. Here, for a set of hypotheses $H_0, H_1, \ldots, H_m$ and their respective P-values $P_0, P_1, \ldots, P_m$, sorted such that $P_0 \leq P_1 \cdots \leq P_m$, the FDR is controlled at a level q^* by determining the largest i for which

$$P_i \leq \frac{i}{m} q^* \ . \tag{3.41}$$

One could either determine the index i at which to draw the line between FDR significant and non significant results or one could compute the FDR q value from the P-values of all tests. Then, one can determine the significant set of tests at $FDR < q^*$ for all tests where $q < q^*$. This has the advantage that one can adjust the threshold q^* arbitrarily without having to re-identify the index i. To calculate q values it is important to order the set of P-values for all tests decreasingly and then calculate the adjusted P-value for each $i \in [1, 2, \ldots, m]$ as

$$p_{i_{adj}} = q_i = min(p_i \frac{m}{i}, \ p_{i-1_{adj}}) \ . \tag{3.42}$$

We applied this method to estimate q-values for the associations between genes and CpGs in Chapter 4 as it was not possible to fit all approx. 5.2×10^9 calculated associations into memory.

3.1.7. Additional background

We refrain from giving a detailed discussion on some of the concepts related to methods applied in this thesis as this would go beyond the scope of this thesis. Here we briefly introduce some of these concepts, just highlighting the most important aspects and refer

the interested reader to some of the excellent and extensive literature for details [191, 195, 197].

Cross validation. Cross validation is a concept often applied in machine learning tasks to optimize model parameters and to avoid over fitting of a model to the given data. Generally, one chooses a subset of the available data for training a model and uses the rest of the data (not used for training) to validate the model. One instance of cross validation is k-fold cross validation where the data is split into k subsets and one iteratively gathers $k-1$ subsets for training and evaluates the model on the remaining kth subset. Performance can then be evaluated over all folds for each of the k validation sets. We apply this procedure in Chapter 5 to estimate the best graphical lasso model over a range of different λ values for our network inference using the Bayesian Information Criterion as a model selection criterion (see below).

Bayesian Information Criterion (BIC). The BIC can be used for model selection based on the likelihood of the model and the number of parameters and samples [191]. It is defined as

$$BIC = log\ p(\mathcal{D}, \theta) - \frac{d}{2} log(N)\ , \tag{3.43}$$

where $log\ p(\mathcal{D}, \theta)$ is the log-likelihood for some data $\mathcal{D}$ and a set of estimated parameters θ, N is the number of samples and d is the number of free parameters in the model. For instance, in the graphical lasso [87] the number of free parameters is defined via the number of non-zero off-diagonal entries in the precision matrix, i.e. the number of non-zero model coefficients (βs) [204].

Markov Chain Monte Carlo (MCMC). In Chapter 5, we utilize the BDgraph method [124] for regulatory network inference, which is based on a specifically designed MCMC algorithm employing the exchange algorithm [205]. Generally, MCMC is a way to obtain samples from high-dimensional distributions [191] such as the posterior distribution for the graphical model as defined for BDgraph, and numerous algorithms have been proposed to do this both effectively (time-wise) and accurately. The algorithm 'walks' through a set of sequential states, e.g. the possible graph configurations in BDgraph through sequential addition and removal of edges, such that the fraction of 'time' it spends in each state is proportional to the target density and each new state depends only on the previous state information. The chain ultimately converges to the target density for a large enough number of iterations ('steps') of the algorithm. We will give some additional information on the methodology behind BDgraph in Chapter 5.

3.2. Processing of cohort data

3.2.1. Genotyping data.

Genotyping and calling were performed using the Affymetrix Axiom platform and software for KORA F4/FF4 and Illumina Genome Studio for KORA F3 and LOLIPOP. To remove low quality samples and SNP, these were subjected to a 97% and 98% call rate threshold, respectively. SNPs with an minor allele frequency (MAF) of below 1% or with Hardy-Weinberg Equilibrium $P > 5 \times 10^{-6}$ were removed and imputation was performed using the IMPUTE v2.3.0 software package [164, 165] with the 1000 Genomes Project cosmopolitan reference panel integrated haplotypes produced using SHAPEIT2 [206]. For the SYS cohort the protocol proposed by the ENIGMA Working Group was employed and the IMPUTE software [164] used for genotype imputations as for KORA F4/3. SNPs were restricted to a subset of 13 million SNPs which have at least been observed twice in European populations and are polymorphic in Caucasians. In the NFBC cohorts the GenCall algorithm was used for genotype calling and IMPUTEv2 used as for the other cohorts. Finally, polymorphic SNPs with IMPUTE info value of > 0.45 were filtered from these data. For our analysis and specifically for association testing we derived allele dosages from genotype data, i.e. numbers of observed and alternative alleles. In diploid organisms, this approach gives rise to three numeric values, reflecting homozygous reference ('0'), heterozygous ('1') and homozygous alternative ('2') alleles.

3.2.2. Methylation data.

Pre-processing of DNA methylation data was performed as previously described for KORA [207] and LOLIPOP [50]. All methylation data were processed using the CPACOR pipeline described in Lehne, A. W. Drong, Loh, et al. [171] with specific details described in [52]. Briefly, separate analyses for autosomes and sex chromosomes are performed and samples with low call rates for probes removed by applying a detection P-value threshold of $P < 10^{-16}$. Probe signal intensities were then quantile normalised and beta-values obtained which were adjusted for control probe principal components, white cell subset estimates [172] and additional study specific covariates such as age and sex. Quantile normalization was performed using the *wateRmelon* R package. Beta values for KORA FF4 EPIC array data were processed as described above. For NFBC66, 67 samples were removed from downstream analysis due to low marker call rate (<95%), a total of 7 samples were removed due to gender inconsistency and one sample due to global outliers in methylation values (1st PC score of the DNA methylation values outside mean $\pm 4SD$). In the NFBC86 study 18 samples were removed based on call rate (<95%) and additional 7 samples due to gender inconsistency.

For association analysis, methylation residuals were obtained using linear regression of beta values as outcome and the collected technical and clinical covariates as predictor

in the following (simplified) model:

$$CpG_{resid} = resid(CpG_{beta} \sim age + sex + Houseman + PC1 + PC2 + ... + PC20) \quad (3.44)$$

These residuals were then used as targets for the genome-wide meQTL analysis. Residual calculation varied slightly for the different cohorts, a full list of models for the individual cohorts is given in Supplementary Table C.2.

Whenever methylation data were combined with expression data, e.g. in expression quantitative trait methylation analysis or during the integrated network inference, methylation data were residualized without adjustment for age and sex but leaving all other covariates in the model specification in order to not adjust for these two covariates twice (i.e. in the methylation and the expression data).

3.2.3. Gene expression data.

Probe level expression data were background corrected, quantile normalized and log2 transformed using the *lumi* R package (v2.8.0 from bioconductor). Probes were excluded from analysis if a known SNP (population MAF>1%) resided under the probe sequence or in case no RefSeq gene annotation could be determined for the respective probe. We further limited analyses to probes present both on HT12-v3 (KORA) and HT12-v4 (LOLIPOP) microarrays to enable meta-analysis of both cohorts.

For expression quantitative trait locus analysis, we obtained residuals for transcript expression values similarly as we did for the methylation betas. We used linear regression to derive residuals from a model specifying the log2 transformed expression values as outcome and SNP dosages as predictors, including age, sex, RNA integrity number (RIN), RNA amplification plate (KORA) / RNA conversion batch (LOLIPOP) and sample storage time (KORA) / RNA extraction batch (LOLIPOP) as covariates. Equation 3.45 shows the model specification in detail.

$$expr_{resid} = resid(expression \sim age + sex + RIN + batch1 + batch2) \quad (3.45)$$

3.3. Processing of replication and validation data

3.3.1. Data used for meQTL replication

In the meQTL study ([1] and Chapter 4) we used three independent datasets to replicate our cross-ethnic meQTL findings. These data were generated and provided by our collaboration partners at the AME and subsequent replication analyses performed by our partners from the ICL. Descriptions on the experimental procedures applied were provided by our collaboration partners and can be found in the appendix (Section A) for reference. Here, we give a short general description of the data and the performed pre-processing as provided by our collaboration partners.

Isolated white blood cell studies

White cell subset samples were collected from 60 individuals comprising 30 obese and 30 normal weight people (Body-Mass-Index $BMI > 35\frac{kg}{m^2}$ and $BMI < 25\frac{kg}{m^2}$, respectively). All participants gave written informed consent as to their inclusion in this study (reference for research ethics committees: 07/H0712/150, 13/LO/0477 and 09/H0715/65). All obese and normal weight individuals were matched for age ($\pm$5 years), ethnicity and sex. Alleles were assessed (Illumina OmniExpress) and DNA methylation quantified (Illumina MethylEpic array) in line with the manufacturer's proposed protocols. The raw methylation data were pre-processed using R v.2.15 and intensities for beads retrieved using the *minfi* R package. Prior to analysis, marker intensities were quantile normalized. Respective quality control criteria for the genotype and methylation data were as described above for the cohort discovery analysis and no samples were excluded after quality control. Finally, genotypes not assessed using the array were imputed using the 1000 Genomes project Phase 3 as a reference with the IMPUTEv2 software package.

Isolated adipocyte studies

We performed additional replication of our findings across tissues in adipocyte samples. For this, we obtained samples from subcutaneous and visceral adipose tissue from 24 healthy controls ($BMI < 30\frac{kg}{m^2}$) and from 24 morbidly obese individuals ($BMI > 40\frac{kg}{m^2}$). All individuals were unrelated from a multi-ethnic background, aged between 18-60 years and were not diagnosed with type 2 diabetes. Individuals of the control group were matched according to age, sex and ethnicity to the cases. Written informed consent was given by all participants (Ethics committee reference 13/LO/0477). Genomic data were generated for a total of 47 of the 48 samples. Genotyping, methylation quantification and data pre-processing including quality control was performed as described above for the white blood cell subset analysis.

DNA methylation in adipose tissue

We collected 603 adipose tissue samples from the MuTHER study for further replication. MuTHER contains 856 samples from female individuals of European descent recruited from the TwinsUK Adult Twin Registry. All procedures were performed according to the ethical standards of the St. Thomas' Research Ethics Committee (REC reference 07/H0802/84) at St. Thomas' Hospital in London. Written informed consent was given by all individuals participating in the study. DNA methylation was profiled using the Illumina Infinium HumanMethylation450 BeadChIP as described previously [208]. Arrays were subsequently scanned with IlluminaHiScan SQ and raw methylation data exported to GenomeStudio v.2010.3 (including methylation module 1.8.2 for image intensity extraction). A combination of Illumina arrays (HumanHap300, Human-Hap610Q, 1M-Duo, and 1.2MDuo 1M) was used for genotype assessment and genotypes called using the Illuminus algorithm. Imputation was performed using the IMPUTEv2 software and the

1000 Genomes phase 3 reference panel, and only SNPs with an IMPUTE info value of > 0.4 permitted for analysis.

3.3.2. IP-MS data used for experimental validation of the *ZNF333* locus

We set out to identify protein binding partners of the *ZNF333* protein for the meQTL project to corroborate our findings. For this, our collaboration partners generated immunoprecipitation mass-spectrometry (IP-MS) data to identify proteins binding to *ZNF333* (see Appendix B.2 for details). In the computational enrichment analyses (see Section 4.2.10) we used two lists of proteins derived from these data: First, the complete list of proteins which were identified and quantified comprising at least two unique peptides (set termed P_{ZNF333_long} and second, the 'shortlist' of quantified proteins exhibiting an IP over control enrichment of $>= 2$ for both immunoprecipitation runs (anti-FLAG mAb and anti-ZNF333, termed set P_{ZNF333}).

3.4. Reproducible cloud enabled workflows

A focus of the work done in context of this thesis was on implementing reproducible workflows. As introduced in Chapter 1, the two core aspects of reproducibility are 1) availability of the original data and 2) a thorough documentation of all (computational) steps performed to achieve the reported results. Here, we focus on the second aspect and briefly describe how we utilized reproducible workflow systems to obtain fully documented and easily reproducible and executable workflows.

3.4.1. The Snakemake workflow system

Important factors in achieving reproducible research are well defined workflows especially in large-scale computational projects. In our projects, we used Snakemake [149] to implement fully reproducible workflows. Specifically, we used Snakemake version 5.7.4, which allowed us to integrate shell, python and R scripts in a single workflow using a workflow definition language based on python. *Snakemake* is built in python[2] and follows a file based scheme in two regards: First, it needs a single text file which describes the *rules* of the workflow and second, each rule relies upon *input files*, which are transformed to the respective *output files* by the commands supplied within the rule definition. For example, a relatively simple rule to annotate which genes of a specific gene annotation exhibit transcription factor binding sites (TFBS) at their TSS could take two annotation files as input and specify a script performing the annotation, yielding an output file containing the annotated TSS as results (see Figure 3.1).

In the rule in Figure 3.1, *annotate_tss_with_tf.R* is an R-script which would take the specified input and output file paths and execute a set of instructions to generate the output (genes annotated with TFBS) from the input (TFBS and gene annotation). The

[2]https://www.python.org/

```
rule annotate_tss_with_tf:
        input:
                tfbs_remap = "<path/to/remap_tfbs">,
                gene_annot = "<path/to/gene_annotation>"
        output:
                tfbs_annot = "<path/to/annotation_output>"
        script:
                "annotate_tss_with_tf.R"
```

Figure 3.1.: Example of a Snakemake rule to annotate gene TSSs with transcription factor binding site information.

created output can then be used as an input to another rule. Once all rules are defined, the user can call *Snakemake* by specifying either a single output file or a *target rule*, which will nudge *Snakemake* to evaluate what has to be done to generate the desired output: It checks which inputs are needed to create the output, and, if those are not yet available, gathers the rules which can generate the required input files. For each encountered rule, it will recursively check the inputs, whether they are present or have to be created, gather the respective rules, check again the inputs, and so on. This process effectively leads to the construction of a directed acyclic graph (DAG), containing all rules and their dependencies which have to be executed in order to create the desired output.

Publishing the workflow in principle allows any user who has Snakemake installed and the respective raw data available to run the complete analysis pipeline using a single Snakemake command ('target'). In addition, Snakemake allows easy visualization of the implemented tasks to get an overview and potentially debug the workflow. For instance, Figure 3.2 shows the *rulegraph* from the simulation study of the network inference project (Chapter 5), displaying all rules which need to be considered when executing the study, created using the *snakemake –rulegraph* option.

3.4.2. Reproducible software environments

Workflow systems such as Snakemake define the individual steps which have to be executed to reproduce results. However, one needs to also define the software environment in which the computations should be executed to achieve full reproducibility, including operating system, script (e.g. R, python) versions and package versions, as different software versions can yield different results. Typically, workflow systems allow to define or use such environments, either by including so call *conda* environments or by specifying software containers. In our case, we made use of software container solutions popular representatives of which are Docker, singularity and Charliecloud (see also Chapter 1). Briefly, one can define in details the desired properties of the software environment, for instance in a single *Dockerfile*, instructing the system on which the container is built to 'install' a particular OS in a specific version, including software and packages of a

Figure 3.2.: Example 'rule graph' for the Snakemake pipeline when executing the network inference simulation study. Graph was created using the *snakemake –rulegraph* command. Individual nodes are rules and edges reflect dependencies between the rules. When the target rule (bold) is called, Snakemake identifies all dependencies and builds an instance of the workflow.

defined version. This container can be exported and used on arbitrary systems, always providing the OS and software specifications as intended by the developer and needed by the workflow.

For instance, we utilized a Charliecloud [150] system to construct and export an image[3] specifically built for our network inference project (see Chapter 5). We deposited our image on Dockerhub[4] which allows easy sharing of such containers and enabled us to use our container both at the ICB compute cluster and the Maryland Advanced Research Computing Center (MARCC)[5], thereby having the same operating system, software and package versions available on both compute systems without large efforts. Thus, workflow systems including software containers enable researchers to implement fully reproducible computational workflows, which can easily be transferred between compute clusters and facilitate reproducible research.

[3] can be found at `https://hub.docker.com/repository/docker/jhawe/r3.5.2_custom`
[4] `https://hub.docker.com`
[5] `https://www.marcc.jhu.edu/`

4. Exploring the genetic architecture of DNA methylation

Chapter Glossary

SNP	**S**ingle **N**ucleotide **P**olymorphism - A single nucleotide change in a DNA sequence
MAF	**M**inor **A**llele **F**requency - The frequency of the less prominent genetic variant at a genetic locus in a population.
CpG/CpG site	A cytosine-guanine ('CG') dinucleotide found in the DNA. DNA methylation typically occurs at the cytosines of CpG sites.
meQTL	**me**thylation **Q**uantitative **T**rait **L**ocus - A SNP linked to a CpG through genotype and DNA methylation association.
cis*, *longrange*, *trans	Categorization of QTL pairs (e.g. SNP-CpG pairs for meQTL). *cis*: same chromosome within 1Mbp; *longrange*: same chromosome, distance $> 1Mbp$; *trans*: different chromosomes
QTL hotspot	A genetic locus (SNP) statistically associated with numerous distinct quantitative traits (such as genes or CpG sites)
PPI	**P**rotein-**P**rotein **I**nteraction
TFBS	**T**ranscription **F**actor **B**inding **S**ite
ChIP-seq	Chromatin-immunoprecipitation followed by sequencing - An experimental procedure to determine protein-DNA interactions such as transcription factor (TF) binding sites.
(C)RE	(***C**is*) **R**egulatory **E**lement - A stretch on the DNA linked with regulation of gene expression (in *cis* of the gene)
TAD	**T**opologically **A**ssociating **D**omain - Chromosome region identified via chromatin conformation capture (e.g. Hi-C) to exhibit a high number of intra-chromosomal interactions.
Locus graph	Concept used to functionally explain QTL hotspots. Consists of a set of nodes (SNP, genes and CpGs) and edges connecting the nodes (e.g. TF-DNA binding or PPI)

DNA methylation is a crucial cellular mechanism and a main factor in the regulation of gene expression, for instance through altering transcription factor binding, which

drives functional and structural properties of the genome and cellular and organism level phenotypes [12, 38]. Understanding the impact genetic variation has on DNA methylation is therefore essential to further our knowledge about cellular systems and, ultimately, disease [209]. Aberrations of DNA methylation patterns have been implicated in several complex diseases such as atherosclerosis and type 2 diabetes [50–54]. Environmental and genetic factors can influence DNA methylation and it could hence provide a mechanistic explanation of how these exposures impact gene regulation and molecular phenotypes [13, 56–58]. Genome-wide methylation quantitative trait locus (meQTL) studies [23, 208, 210–213] further confirmed the influence of genetic variants on DNA methylation, both in *cis* and in *trans*, and showed that *trans*-acting meQTL could ease the identification of genomic master regulators.

This chapter is focused on the analyses performed in the context of a large collaborative project between the Department of Epidemiology and Biostatistics at the Imperial College of London (ICL) as well as the Research Unit of Molecular Epidemiology (AME) and the Institute of Computational Biology (ICB) at the HelmholtzZentrum München and describes parts of the manuscript submitted for publication in *Nature Genetics* [1]. Our goal was to unravel the complex processes underlying genomic master regulators and human traits. To this end, we investigated the effect genetic variants exhibit genome-wide on DNA methylation in a meta-analysis of several large population cohorts and analyzed the functional implications of the observed associations through curated molecular interactions. In this chapter, we will detail our analysis strategy to obtain ethnicity independent ('cosmopolitan') meQTL and describe the extensive functional evaluation and network analyses performed to 1) establish their functional relevance and 2) derive information about master regulators.

4.1. Epigenetic gene regulation through DNA methylation

In our work, we focused on dissecting the regulatory implications of three distinct types of methylation quantitative trait loci (meQTL) based on the genomic locations of the associated SNPs and CpGs, namely:

1. ***cis*-meQTL**: (SNP, CpG) on the same chromosome, at most 10^6 bp (1Mbp) apart
2. ***longrange*-meQTL**: (SNP, CpG) on the same chromosome, more than 1Mbp apart
3. ***trans*-meQTL**: (SNP, CpG) on different chromosomes

Using this classification scheme, it is possible to address diverse biological questions and multiple mechanisms of action can be thought of to explain the distinct types of associations (see Figure 4.1).

cis-meQTL. These meQTL represent genotypes that are close to the affected CpG. Specifically, the SNP and the CpG 1) reside on the same chromosome, and 2) are within

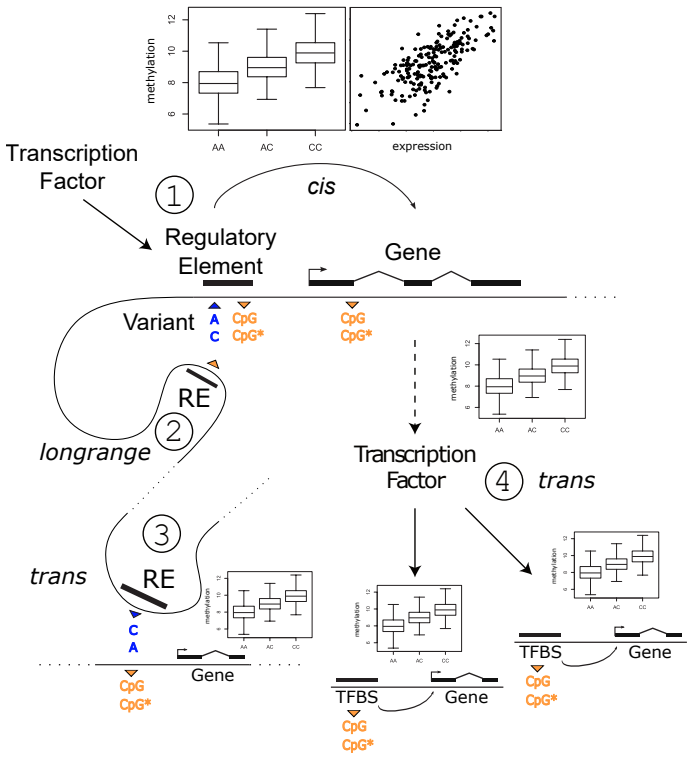


Figure 4.1.: Schematic of how SNPs could influence CpG methylation in *cis* , *longrange* and *trans* and how their combined effect can alter gene expression. **1)** *cis*-meQTL: SNP lies in same regulatory element (RE) as the CpG. It changes the RE sequence leading to altered TF binding and CpG methylation which in turn affects the expression of a nearby gene. **2)** *longrange*-meQTL: SNP and CpG reside in different REs but are brought together through local chromatin structures (e.g. chromatin loops). **3)** *trans*-meQTL: Similar to 2), but the contact of the two REs is established across chromosome boundaries. **4)** SNP affects the activity of a TF in *cis* ultimately leading to changes in DNA methylation of downstream targets of the TF.

1Mbp (Mega base pair) from each other. A possible explanation for this specific type of association is that the SNP and the CpG reside in the same regulatory element (RE) and that the SNP directly changes the sequence of the element. This would alter e.g. a potential TF binding site and lead to a direct change in the observed methylation pattern, ultimately affecting the expression of genes linked to the regulatory element (Figure 4.1, step **1**). We investigated whether or not the observed *cis*-meQTL indeed reside in the same regulatory element by performing an enrichment analysis in histone modification based chromatin states (*chromHMM* states [48, 49]). This analysis is detailed in Section 4.3.3.

longrange cis-meQTL. SNP-CpG pairs in this category reside on the same chromosome but have a relatively large distance (>1Mbp) to each other, bringing up the question of how these entities might realize their functional association. A possible explanation involves 3D chromatin structures, specifically chromatin loops as indicated in Figure 4.1, step **2**. Physical interactions of the genome bring together potential regulatory elements such as the ones the CpGs and SNPs reside in, which can then form regulatory relationships [75]. 3D chromatin structure can be assessed using next-generation sequencing techniques such as high-throughput chromatin conformation capture (Hi-C) [72] and promoter capture Hi-C (PCHi-C) [73]. To elucidate the potential role of DNA secondary structure in forming regulatory relationships we gathered public Hi-C and PCHi-C data [75] and evaluated possible enrichment of *longrange*-meQTLs in DNA contacts

(Section 4.3.3)

trans-meQTL. The final group of meQTL consists of SNP-CpG pairs that reside on different chromosomes. The mapping of *trans*-(me)QTL is generally a difficult problem and limited in its application by a low number of individuals in genomics studies (low power) and a large multiple testing burden (millions of SNPs are tested against hundreds of thousands of methylation sites). In this study, however, the meta-analysis of the large population cohorts enabled extensive analyses of *trans* associated meQTL pairs (see Section 3.1.4). Multiple distinct mechanisms could be thought of when considering *trans* associations, of which we consider two: First, 3D chromatin structures might connect the two entities similar to *longrange*-meQTL, but connecting different chromosomes rather than forming loops within a single chromosome (Figure 4.1 step **3**). Second, the observed *trans* effect could be mediated by a sequence of regulatory steps including transcription factors as indicated in step **4** of Figure 4.1. In this case, a possible sequence of actions might involve the meQTL SNP directly affecting the expression or function of a nearby *cis* gene, leading to a change in regulatory patterns involving associated proteins which alters transcription factor binding at the CpG sites and ultimately leads to a change in methylation in *trans*. For the first possibility (step 3 in Figure 4.1), we utilized Hi-C data to determine whether *trans*-meQTL are found to be more often located within the same inter-chromosomal chromatin contacts as we would expect by chance (Section 4.3.3). For the latter (step 4 in Figure 4.1) we devised an elaborate network analysis procedure based on established protein-protein interaction (PPI) networks and public chromatin immunoprecipitation sequencing (ChIP-seq) data to 1) pinpoint the most likely *cis* candidate genes mediating QTL effects, 2) generate hypotheses on how the genetic information flows through the protein-TF network and 3) identify the regulatory role of DNA methylation for individual loci. Using this approach we were able to dissect regulatory processes underlying important *trans*-QTL hotspots, including trait-associated genetic loci. The results of this analysis are detailed in Section 4.3.4.

4.2. Methods for meQTL investigation

4.2.1. Identification and pruning of global meQTL

To investigate the genetic influences on DNA methylation and to obtain new insights into the molecular pathways connecting genetic variants to phenotypes, we performed a global association analysis between imputed genotypes and DNA methylation at CpG dinucleotides (CpG sites), thereby establishing methylation quantitative trait loci (meQTL). Methylation quantitative trait loci are pairs of SNPs and CpGs which show a significant statistical association in their genotypes and DNA methylation pattern, respectively, and form the basis for all our analyses in this chapter. We set out to identify ethnicity independent ('cosmopolitan') meQTL pairs by first obtaining robust meQTL

within both ethnic groups and subsequently meta-analyzing them across ethnicities (see Figure 4.2). The computations to generate the full and pruned lists of cosmopolitan meQTL were performed by our collaboration partners at the ICL (Dr. Lehne).

Model description and probe filtering. Determining the statistical association between a SNP and a CpG boils down to performing a linear regression (see Section 3.1.2) between the measured genotypes (SNP allele dosages, see Section 3.2) and the obtained methylation beta values (for the CpG) for all available individuals, i.e.

$$CpG \sim \beta_0 + \beta_1 SNP.$$

Note, that typically batch effects and other covariates are added to the independent variables to account for confounding effects, however, in this study we removed potential confounders by residualizing the methylation data prior to association analysis (see Section 3.2 for details). While the association analysis is a relatively straight forward concept some issues need to be addressed before carrying out the regressions. First, the probes on the methylation array are prone to cover genomic regions containing genetic variants. This potential genetic influence of SNPs on the hybridization of the probes can lead to misinterpretation of the methylation at the respective CpG site and hence all probes containing SNPs with a minor allele frequency (MAF) of 1% (n=121,932 in Europeans; n=84,295 in South Asians) are removed. Secondly, CpG probe sequences can cross-hybridize, meaning that they show high sequence similarity to more than one region of the genome. These probes, too, can lead to erroneous methylation beta estimates and hence are excluded from the analysis (n=43,233). Another more technical issue is the sheer size of the data and the number of models that need to be calculated. For both populations, roughly 9 million SNPs had to be tested for genetic effects on approximately 350,000 CpG methylation sites (after filtering), amounting to a total of about 3.15×10^{12} individual association tests per population. We used QUICKTEST [214–216] to perform efficient meQTL calculation and to run this analysis in a feasible amount of time and further chunked the input methylation data to enable high-performance computing parallelization (i.e. distribution of calculations across different computing hosts). For the calculation of meQTL in KORA FF4 using the EPIC methylation data, we utilized matrixEQTL [217].

Meta-analysis of study populations. To generate cosmopolitan meQTL, we first performed separate discovery and replication analyses for both Europeans and South Asians based on the independent population-level data. For the discovery stage, we obtained meQTL results separately from KORA F4 (N=1,731, European ethnicity) and 1,841 LOLIPOP individuals (South Asian ethnicity), yielding 7.2 million and 11.4 million SNP-CpG pairs for Europeans and South Asians at $P < 10^{-14}$, corresponding to $P < 0.05$ after Bonferroni correction, respectively. The associations significant in the discovery phase were subsequently replicated in KORA F3/SYS/NFBC (N=2,068) and LOLIPOP (N=1,354) to obtain high confidence SNP-CpG associations for Europeans and South

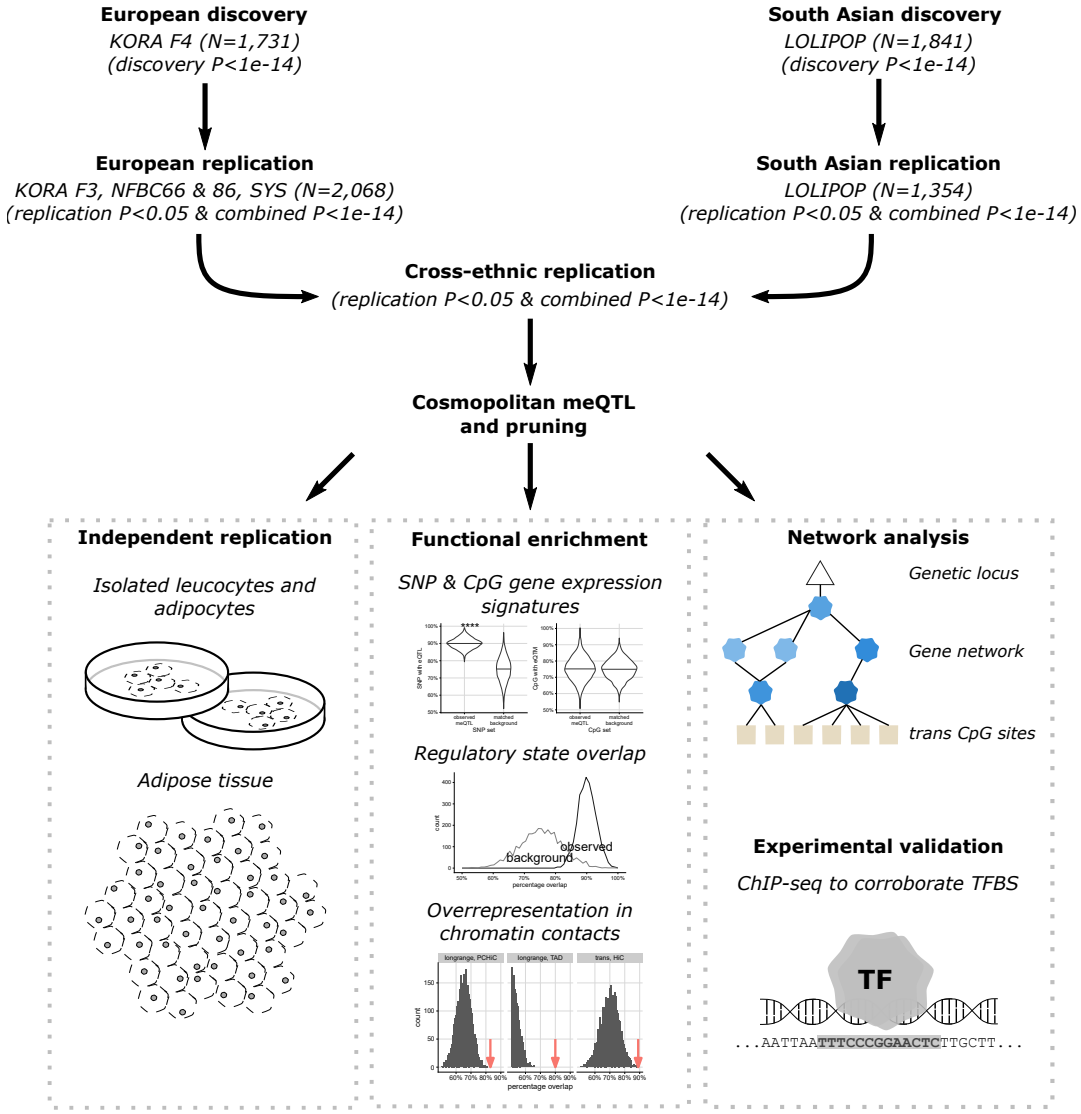


Figure 4.2.: The analysis plan followed in the meQTL project. After individual discovery and replication of meQTL in European and South Asian cohorts we kept only SNP-CpG pairs replicating across both populations. Cosmopolitan meQTL were subsequently pruned and independent meQTL loci identified. The final set of meQTL was then investigated for replication in independent data and diverse functional enrichments performed. *Trans*-meQTL hotspots were further subjected to detailed network analysis and followed up by experimental validation.

Asians, respectively. Combined replication P-values were calculated based on the effect sizes and standard errors from the individual association tests using fixed-effect meta-analysis (see Section 3.1.4) in the METAL software [218]. To pass the first replication stage and to be admitted to the cross-ethnic replication stage meQTL needed to 1) show a consistent direction of effect between discovery and replication, 2) have a replication $P < 0.05$ and 3) have a combined $P < 10^{-}14$. Finally, replicated meQTL derived for Europeans and South Asians entered the cross-ethnic replication stage in which we tested European associations against South Asian derived associations and vice versa. Associations replicating between the two ethnicities (same criteria as before, i.e. replication $P < 0.05$, combined $P < 10^{-14}$, same direction of effect) then formed the set of cosmopolitan meQTL (N=11,165,559 SNP-CpG pairs) which we focused on in all subsequent analyses. This strategy is also illustrated in the upper part of Figure 4.2.

Identifying independent meQTL and LD pruning. Due to local correlations between SNPs (linkage disequilibrium, see Section 2.1.1) and between neighboring CpG sites, the meQTL analysis led to redundant SNP-CpG pairs that effectively represent the same genetic and epigenetic loci. We employed a two-step approach to identify independent SNP and CpG loci over all cosmopolitan pairs which we termed 'sentinel pairs', or, with respect to the individual entities, sentinel SNPs and sentinel CpGs.

First, we performed an iterative conditional analysis using the same data as for the initial association analysis. For each CpG C, we selected the most strongly associated SNP S_L (i.e. the one with the lowest P) and repeated association tests for all other SNPs that were initially associated with C, while including S_L as an independent variable in the regression model. This effectively removes the genetic effect of S_L on the C and helps to determine independent genetic effects. We repeated this procedure for all remaining SNPs which still showed significant association to C ($P < 10^{-14}$) after accounting for the respective strongest SNP until no more SNPs remained. The resulting, independently associated SNPs were then linked to their respective CpG.

While the conditional analysis reduces the redundancy caused by linkage disequilibrium between SNPs, still a possibility remains, that the same genetic locus is represented by different SNPs. This can result from the fact that the top associated SNP for each genetic locus can be different for different CpGs. Therefore, in order to get rid of these indirect effects of local correlations, we reduced highly correlated SNPs and CpGs to independent SNP loci and methylation sites, respectively. To this end, we picked again the top associated entity (i.e. SNP or CpG with lowest P) as 'sentinel' SNPs and CpGs, and assigned all entities with $R^2 > 0.2$ and distance $< 1Mbp$ to the corresponding loci. This procedure was then repeated for the remaining markers until no more remained.

4.2.2. Replication of meQTL in independent data

We replicated cosmopolitan meQTL findings in three independent datasets: isolated leukocytes, isolated adipocytes and in adipose tissue from the MuTHER study (see also Section 3.3.1). Data were generated and provided by our collaboration partners together

with the initial results for this analysis, derived as described below.

We re-calculated associations for all 11,165,559 discovered cosmopolitan SNP-CpG pairs (see Chapter 4) individually in these three datasets. In the isolated adipocyte dataset, a total of 9,408,762 of the 11,165,559 cosmopolitan pairs were tested due to some SNPs and CpGs missing in the data as a result of the applied QC criteria or due to the different genotyping platforms. Associations between SNPs and CpGs were established using linear regression for which we used age, gender, ethnicity, and obesity case-control status as covariates to be in line with the initial association analysis. The MuTHER study contains samples from related individuals which needs to be considered during the association analysis in order to avoid confounding by sample kinship. Therefore, the GEMMA software was used to determine associations in these data [219]. GEMMA (**G**enome-wide **E**fficient **M**ixed **M**odel **A**ssociation) implements an efficient procedure to apply standard linear mixed models for genome-wide association study analyses. In essence, a univariate linear mixed model is fit to perform association tests for markers with a single phenotype, which allows GEMMA to account for population stratification and sample structure as well as to determine 'chip heritability', i.e. the proportion of variance in phenotypes explained by respective genotypes. In the final model for association testing we used as covariates the kinship matrix reported by GEMMA as well as age, gender, first 20 control probe, and first 5 genotype principal components.

For the replication across platforms, we utilized the MeDIP-seq based associations reported by Bell, F. Gao, Yuan, et al. [220]. The authors reported a total of 7,184 associations between genetic risk SNPs subdivided into LD blocks and "haplotype-specific DNA methylation (HSM) peaks", reflecting genomic regions of a minimum length of 500 base pairs (bp). Of the 7,184 associations, 328 involve at least one specific SNP and one CpG in a HSM peak, for each of which we collected the respective SNP-CpG pair from our cosmopolitan results. Next, we determined how many LD block-HSM peak associations can be replicated at various significance thresholds. In addition, to establish whether these numbers were more than expected by chance, we performed background sampling to generate a null distribution. To this end, we randomly selected 100 background pairs for each observed SNP-CpG pair with SNPs matched for MAF ($\pm$ 0.1%), and CpGs for standard deviation ($\pm$0.1%) as well as genomic distance ($\pm$ 10kbp). We further defined more relaxed criteria to obtain a background for non matched SNP-CpG pairs where we matched pairs for MAF ($\pm$0.1%), CpG standard deviation ($\pm$1%), and a genomic distance threshold of 50kbp.

For each of the 100 matched background sets we observed the total number of replicated LD block-HSM peak associations and for each association with more than one meQTL pair, we recorded the smallest P-value. Finally, we obtained an empirical P-value for the true number of replicated LD block-HSM peak associations according to our background distribution. In Bell, F. Gao, Yuan, et al. [220] the authors calculated a representative P-value from the mean P-value over all SNP-HSM peak pairs reflecting

the same HSM-LD block associations. We hence use the same procedure to calculate the results for all KORA pairs in this analysis and to determine the percent of replicated associations.

4.2.3. Enrichment of *cis*-meQTL in functional chromatin states

For *cis*-meQTL, where the involved SNPs and CpGs reside relatively close ($< 1Mbp$) to each other on the same chromosome, a possible explanation for a functional connection is that the observed polymorphisms change the sequence in a regulatory element (Figure 4.1). To investigate this, we assessed whether *cis*-meQTL are enriched in the same regulatory states, specifically in enhancer (potential distal regulatory element) or promoter (potential *cis* regulatory element) chromHMM states [48].

The chromHMM states were derived from five distinct histone marks using a multivariate hidden Markov model [49]. Combinations of certain histone modifications (see Section 1.3.2) indicate distinct chromatin states, for instance, whether a DNA stretch can be actively transcribed or is closed to the transcriptional machinery. A full overview of all defined states and the ones utilized in this analysis is given in Table 1.1.

We asserted the functional relationship of *cis*-meQTL pairs by analyzing, whether or not entity pairs reside more often within the same regulatory class than expected by chance, focusing on active regulatory marks relating to enhancer states ('enhancer' class) or promoter states ('promoter' class). To this end, for each *cis*-meQTL CpG in a promoter or enhancer class state we assessed whether it maps to the same regulatory class as at least one of its associated meQTL SNPs ('match') or in a different class ('ambiguous'). We then evaluated whether observed *cis*-meQTL loci are enriched to be in the same class by sampling a random but matched set of background pairs. For each CpG with N associated meQTL SNPs we obtained a random background CpG locus (i.e. including background SNPs) from all CpGs matching mean and standard deviation ($\pm$ 5%) and with at least N SNPs in similar proximity ($\pm$ 1kbp) and with matched minor allele frequency (MAF, $\pm$ 5%) as the SNPs from the meQTL CpG, making sure that we only obtain markers not part of the cosmopolitan meQTL pairs.

Finally, we annotated the sampled background loci with the same state information as we did for the observed meQTL. This procedure is repeated 100 times and for each iteration 1,000 *cis*-meQTL loci are sampled and 1,000 background pairs determined according to our criteria. To assess whether the distribution of the fraction of pairs in the same state for the observed pairs is significantly shifted above the background distribution we applied Wilcoxon's signed rank test.

4.2.4. Enrichment of *longrange*- and *trans*-meQTL within chromatin contacts

We used both PCHi-C contacts and Hi-C derived topologically associating domains (TADs) to investigate *longrange* associations and utilized inter-chromosomal chromatin contacts derived from Hi-C data published in Javierre, Sewitz, Cairns, et al. [75] to investigate *trans*-meQTL. TADs are derived from Hi-C data and reflect local regions

	same contact/TAD	**not in same contact/TAD**
observed pairs	A	B
background pairs	C	D

Table 4.1.: Example for a contingency table derived during the TAD/HiC analysis. Odds ratios indicating enrichment of observed pairs in contacts as compared to the background pairs can be computed as $OR = \frac{A/B}{C/D}$.

with a high number of intra-chromosomal DNA contacts (e.g. DNA loops) representing important regulatory features of eukaryotic cells [221]. While the underlying data for the *longrange* and *trans* analyses are slightly different, the general approach for both is similar: We seek to establish that observed meQTL pairs are more often located at chromatin contacts (*longrange* and *trans* pairs) or located within the same TAD (*longrange* pairs) as compared to a background set of meQTL SNP-CpG pairs, thereby providing additional evidence for functional associations through chromatin contacts. For the TAD and Hi-C enrichment we employ the same sampling procedure of background pairs as for the *cis*-meQTL chromHMM enrichment. However, for the PCHi-C enrichment we further constrained the sampling such that at least one of the entities resides in the promoter of a gene (2,000bp upstream and 1,000bp downstream of the transcription start site), which effectively mimics the inherent properties of promoter-capture Hi-C. Here, specifically designed oligonucleotides matching all known promoter regions are used to capture chromatin, thereby always capturing fragments for gene promoters together with their distally linked (i.e. , in contact) DNA fragments.

For each of the three analyses (*longrange* TAD/PCHi-C, *trans* Hi-C), we performed 150 iterations of sampling background pairs for all pruned meQTL in the respective category. By counting the number of overlaps of observed and background meQTL pairs with (PC)Hi-C contacts and the number of times a *longrange* SNP-CpG pair resides within the same TAD region, we assessed whether observed meQTL show functional evidence in the HiC data through enrichment in the derived contact regions. To investigate this formally, we generated contingency tables based on the counted overlaps which show the total number of observed and background pairs residing or not residing in the same TAD/chromatin contact. An example of such a table is given in table 4.1.

From these tables we obtained odds ratios (ORs), i.e. the fraction of the overlap fractions for the observed pairs vs the background pairs ($OR = \frac{A/B}{C/D}$). By setting $H_0 : OR \leq 1$, we obtained an empirical P-value using

$$P(\mathcal{D}|H_0) = \frac{\sum_{d \in \mathcal{D}} I(d \leq 1) + 1}{N + 1},$$

where N is the total number of iterations performed, $\mathcal{D}$ is the data (all ORs) and $I(.)$ is the indicator function, returning 1 if its argument evaluates as $TRUE$ and 0 otherwise. The empirical P-value reflects the probability of obtaining the same or a more extreme OR as we observed for the 'real' meQTL pairs given the null distribution based on the

matched background.

4.2.5. Enrichment of meQTL pairs for association with gene expression

To corroborate meQTL pairs we examined whether or not meQTL SNPs and CpGs are enriched for associations with gene expression. To this end, we calculated expression quantitative trait loci (eQTL) and expression quantitative trait methylation (eQTM) for all independent sentinel SNPs and CpGs, respectively. For eQTL, we first prepared the gene expression data, removing potential confounding effects as described in Section 3.2 (including Houseman estimates), and used sentinel SNPs as independent variables in a linear model to assess their effect on gene expression. For eQTM, we used the same expression data but adjusted CpG methylation data (beta values) only for the Houseman white blood cell proportions and the Illumina control probe principal components 1-20 (see Section 3.2), since we already adjusted for age and sex in the expression data. To determine eQTM we again regressed gene expression as the dependent variable against CpG methylation as the independent variable using a linear model. In addition, we removed the effect of *cis* regulatory SNPs which could potentially confound the eQTM signal by regressing out *cis*-eQTL and meQTL genotype information from the tested CpGs and genes. We employed a meta-analysis approach to increase the power of eQTL and eQTM detection by analyzing associations separately for KORA, LOLIPOP Indian Asians, and LOLIPOP Europeans and then combining individual results by inverse-variance meta-analysis (Section 3.1.4). Statistical significance was then inferred at $P < 4.04 \times 10^{-13}$ for eQTL and 8.7×10^{-12} for eQTM, each corresponding to $P < 0.05$ after Bonferroni correction.

We sampled a corresponding background for each of the sentinel SNPs and CpGs to quantify expectations under the null hypothesis, i.e. that observed meQTL are not enriched for gene expression associations. Background SNPs are sampled randomly from all available SNPs, constraint such that 1) they are not part of a significantly associated SNP-CpG pair, 2) SNPs are matched for minor allele frequency (MAF $\pm$ 5%) with the meQTL SNP and 3) they have a similar distance to their respective nearest gene ($\pm$ 10 kbp). We then assessed whether sentinel SNPs are over-represented for association with gene expression by building a contingency table reflecting 1) the two groups of SNPs, i.e. meQTL or background and 2) whether or not the SNP has at least one significant association with a gene, i.e. is an eQTL for at least one gene. We assessed statistical significance of expression enrichment by applying Fisher's exact test on the constructed tables and setting a P-value cutoff of $P < 0.05$. The same analysis was performed independently for meQTL SNPs and CpGs. For CpGs, we matched background CpGs using the same criteria, but instead of MAF we matched mean and variance of methylation betas ($\pm$5%) between meQTL and background CpGs. Finally, SNPs and CpGs where stratified according to their meQTL category (*cis*, *longrange*, *trans* and according to their observed effect sizes in the meQTL associations (low, medium, and high effect size groups). To obtain equal effect group sizes, we cut the distribution

of effect sizes at the 33% and 66% percentiles.

4.2.6. Selection of candidate genes for SNPs affecting CpGs in *trans*

For each locus we aimed to determine the most likely *cis* regulator mediating the observed effects by selecting 1) the nearest gene as a potential candidate and 2) the genes, for which the SNP is an eQTL in GTEx whole-blood gene expression data [82, 93] and in our study. Selecting by eQTL yielded a total of 507 SNP-gene pairs, 381 of which did not involve the gene nearest to the sentinel. Overall we selected 1,712 unique candidate genes over all sentinels. For the *trans*-acting sentinels where the random walk based approach (see below) could not be applied to determine the underlying regulatory network, e.g. due to *cis* genes not being present in STRING, we used genes selected as described above for downstream analyses (e.g. for the *ZNF333* locus).

4.2.7. Enrichment of regulatory genes at meQTL loci

We investigated the type of genes encoded in the regions around the identified sentinel SNPs to evaluate whether these are enriched for regulatory genes such as epigenetic modifiers and transcription factors. To this end, we curated three gene lists presented in Lemire, Zaidi, Ban, et al. [211] and which comprise 1) the curated list of epigenetic regulator genes in Supplementary Table 4 of [211], 2) the list of TFs curated from [222] (classes 'a' and 'b') and 3) the zinc-finger gene (ZNF) subset of these transcription factors (simple match of 'ZNF' in gene name).

We assessed enrichment of genes in these lists in the 1,847 identified *trans* SNP regions by sampling background SNPs for each sentinel with matched MAF ($\pm$ 5%) and then detecting the genes in a region around the matched SNP with the same size as the sentinel region. We then counted how often genes in the respective lists 1) are detected in the sentinel region and 2) are detected in the matched background region and calculated the fraction of detected genes versus the not detected genes. SNPs present in the table of cosmopolitan meQTL pairs were excluded from the list of background SNPs and the SNP sampling repeated 1,000 times. Significance of enrichment was then established by evaluating the fraction of odds ratios (observed over background, similar to Table 4.1) in the overlap tables less than or equal to 1 (using $H_0 : oddsratio \leq 1$ as in the HiC enrichment analyses) to obtain an empirical P-value, indicating the probability of obtaining the same or a more extreme OR under the null hypothesis.

4.2.8. TFBS enrichment at *trans* associated CpG sites

In order to show that *trans* associated CpGs are indeed functionally related to the identified *trans*-acting SNP loci, we sought to assess whether *trans* CpG regions are enriched for binding sites of DNA binding proteins/transcription factors. To achieve this, in a first step we collected all chromatin immunoprecipitation followed by sequencing (ChIP-seq) based transcription factor binding sites (TFBS) from the ENCODE [60] and the

ReMap resource ([71], see also Sections 2.3.1 and 2.3.6). After filtering for experiments performed in blood-related cell-lines (see Section 2.3.6), the final set of DNA binding proteins contained a total of 145 TFs from 246 distinct ChIP-seq experiments.

In the next step, we overlapped the curated TFBS for all transcription factors with the *trans* CpG regions for each sentinel and generated a list of TFs of interest for each genetic locus. An overlap between TFBS and CpG site was counted if the binding site overlaps the 100bp window around the CpG (i.e. 50bp upstream and downstream of the CpG).

Power analysis

Prior to over- and under-representation analysis of TFBS we first analyzed how many *trans* CpGs are needed at minimum to be able to detect such enrichments.

We seek to establish the enrichment using a Fisher test on the contingency table created from the observed TFBS overlaps and the overlap of a set of background TFs. Therefore the number of needed *trans* CpGs is dependent on whether the smallest achievable P-value in this Fisher test is below an adjusted significance threshold p_{adj}. In turn, the adjusted significance threshold directly depends on the number of tested loci $n_{loci}(minsize)$, where *minsize* is the minimal size threshold for the number of *trans* CpGs per locus. We hence performed the power calculation where we used the Bonferroni method to compute adjusted significance thresholds as $p_{adj}(minsize) = 0.05/(n_{TF} * n_{loci}(minsize))$, where $n_{TF} = 246$ is the total number of ChIP-seq experiments. Using this method we then proceeded to systematically construct contingency tables for different numbers of *minsize* assignments n_1 (i.e. minimal number of *trans* CpGs for the same sentinel) in the range $n_1 = 1..20$ and for the overall number of genetic loci with *trans* associated CpGs overlapping TFBS $n_2 = \{0..n_1\}$.

In order to define the background counts for the contingency tables we need to obtain the approximate binding frequency of TFs at CpG sites not associated with any sentinel SNP and the total number of CpGs in the background set. We estimated this background binding frequency of TFs by observing the mean binding frequency of all available TFs across all CpGs from the data yielding $p_{bg} \approx 0.05$. To get the number of CpGs in the background set we further assumed that for each enrichment test the background set of CpGs entails all CpGs available on the 450k array excluding the CpGs which are associated with the largest *trans* cluster ($maxsize \approx 250$ CpGs, hence $n_{bg} \approx 486,923$). Using these estimates we thus set the counts for the background in the contingency table to $p_{bg} \times n_{bg}$ and $(1 - p_{bg})xn_{bg}$ for overlapping and not overlapping sites, respectively, and determined for each cluster size n_1 the smallest Fisher test P-value over all n_2 values. Based on $p_{adj}(minsize)$ we ultimately obtained the minimal cluster size $n_{1_{min}}$ where an n_2 exists which yields a P-value $P < p_{adj}(minsize)$. In this analysis we found $n_{1_{min}} = 5$, yielding the smallest P-value $P = 3.6 \times 10^{-7}$, which falls below $p_{adj}(minsize) = 1.8 \times 10^{-6}$ for $minsize = 5$.

TFBS enrichment analysis on *trans* CpG signatures

We proceeded to systematically test the *trans* signatures for each of the 115 sentinel SNPs with ≥ 5 *trans*-CpGs for over-/under-representation of binding sites using the all 246 ChIP-seq TFBS datasets.

We retrieved conservative enrichment estimates by performing Fisher's exact test on the generated contingency tables for two different definitions of the CpG background and recording the highest P-value only. First, we defined the background as all CpGs available on the Illumina 450K methylation array. Second, we set the background based on randomly sampling sets of CpGs from the array which match in population mean and standard deviation of methylation betas to the associated *trans* CpGs for each genetic locus. Next, we obtained empirical P-values indicating the significance of the overlap between the observed *trans* sites and the TFBS by re-sampling 10,000 sets of matched CpGs of equal size for each sentinel. In order to adjust for multiple testing we applied the Benjamini-Hochberg method on the results for both backgrounds. Finally, we used a conservative criterion to define enriched or depleted transcription factor signatures requiring FDR less than 5% for both tests.

We observed $y = 45$ loci which showed significant enrichment of TFBS in the *trans* CpG signatures (see Supplementary Figure C.2). In order to check if our results represent enrichment beyond the null hypothesis, we estimated the number of genetic loci with *trans* CpGs overlapping TFBS for a random scenario. We can compute the probability of obtaining at least one random association, i.e. one false positive, per SNP using 1) the total number $n_{loci} = 115$ of loci, 2) the total number of $n_{TF} = 246$ ChIP-seq data sets and the fact that 3) we control the false discovery rate at 0.05, corresponding to a P-value threshold of $p_{Th} = 1 \times 10^{-3}$. The latter is the probability of getting a false positive in a given test and we can hence compute the probability of getting a false positive per SNP as $P(x > 0) = 1 - P(x = 0) = 1 - (1 - p_{Th})^{n_{TF}} = 0.22$. Therefore, we can estimate the expected number of SNPs having at minimum one association as $n_{loci} * P(x > 1) = 25.0$. We observed $y = 45$ loci with enrichment of TFBS at *trans* CpG sites, hence the P-value to obtain this results is $P(y > 45)\ Binom(p = P(x > 0), n_{loci}) = 7.4 \times 10^{-6}$.

Finally, we performed a sensitivity analysis to determine the influence of our selected CpG window size of 100bp to detect significant TF signatures. The results of this analysis are shown in Supplementary Figure C.3 for varying window sizes of 2, 100, 500, 1,000, 5,000 and 10,000bp. With larger window sizes we observe an increased amount of discoveries which potentially mirror regional correlations between CpG methylation betas. However, in order to be conservative and therefore underestimate the true number of *trans* TF signatures we chose a window size of 100bp.

4.2.9. Random walk analysis on locus graphs

For each meQTL hotspot ($\geq 5 trans CpGs$) we sought to identify the most likely candidate SNP gene from the full set of SNP genes (i.e. all genes located in a 1Mbp window around the sentinel) which mediates the observed *trans* effects. To this end, we linked all SNP

genes to the sentinel associated CpGs via a cascade of protein-protein interactions (PPI) and protein-DNA interactions.

Definition of a context specific protein-protein interaction network including protein-DNA interactions

For connecting the *cis* and *trans* loci we obtained PPI which showed either experimental or database curated evidence as available in the STRING database (see Section 2.3.5). This STRING subset comprised a total of 12,769 individual proteins and 186,674 protein interactions. We further aimed to generate a context specific protein interaction network and hence restricted the initial network to 8,880 proteins that showed expression in whole-blood data from the GTEx v6p data set [82]. A gene was defined as expressed if it exhibits a median reads per kilobase per million sequenced (RPKM) value of > 0.1. In addition, we chose only the largest connected component of the network to get a fully connected network on which to perform the analysis which entailed 8,668 proteins and 99,143 protein interactions.

Specifically, the obtained PPI network can be formulated as a network (or graph) $P = (V_P, E_P)$, in which the set V_P contains the nodes/vertices representing individual proteins and the set E_P with each element $E_{Pi} \in V_P x V_P$ are undirected edges in the network which connect the individual nodes (i.e. representing the actual protein-protein interactions). Protein-DNA interactions can be formulated similarly as a graph $D = (V_D, E_D)$, where V_D again represents the nodes consisting of all 145 distinct TFs available in the ChIP-seq data as well as all CpG sites which are bound by any of the DNA binding proteins (i.e. within 50bp of the binding sites).

Locus graph definition

Our goal is to prioritize SNP genes and generate candidate pathways for each hotspot locus. We hence defined 'locus graphs' which are created from the PPI network P augmented by locus specific CpG sites and their transcription factor bindings. For each locus, we collected the set of CpG sites S associated with the sentinel SNP in *trans* in the full set of all cosmopolitan meQTL hits. We further extended the set S by all *trans* associated CpGs of meQTL SNPs which 1) are associated with a *trans* CpG of the sentinel and 2) reside in *cis* of the sentinel SNP (i.e. within 1Mbp). Finally, we included all CpGs in S together with their respective protein-DNA interactions $E_D(S)$ in the large PPI network P to establish the locus graph $G = (V_P + S, E_P + E_D(S))$. For each locus, we identified the set of candidate genes C as all genes encoded at the SNP locus that are part of the PPI network. Locus regions were defined based on the results of the pruning analysis that identified sentinel SNPs.

Ranking of candidate genes through random walks

The last remaining step is to rank/prioritize the set of candidate genes C using the random walk on the topology of the locus graph for each analyzed locus. Candidate genes are defined as all genes which are part of the PPI network P and are encoded in *cis* (i.e. within 1Mbp) of the sentinel SNP's location. The prioritization of candidate genes based on random walks is similar to applications in previous studies [223, 224]. To implement the analysis, we utilize the adjacency matrix representation of graphs, i.e. an undirected graph $G = (V, E)$ with $E \in (VxV)$ can be represented as a symmetric matrix A with nodes in the columns and rows and where each entry $a_{ij} = a_{ji} = 1$, if edge $(i, j) \in E$ and $a_{ij} = a_{ji} = 0$, if $(i, j) \notin E$.

In a first step, we set up the symmetric transition matrix T of the same dimensions as A and with entries (t_{ij}), where $t_{ij} = a_{ij}/sqrt(d(i) \times d(j))$ and where $d(i)$ gives the degree (i.e. the number of adjacent edges) of a node i. T specifies for each node i the probability to move to node j in a single step of the random walk [225]. Further, the transition probability for paths of length t to 'walk' from node i to j can be computed by taking the t-th power, i.e. computing T^t. We are interested in random walks between the CpG sites S and the candidate genes C and consequently computed t-step transition probabilities T^t_{sc} for each CpG site $s \in S$ and each candidate gene $c \in C$.

However, the lengths of the paths t are not known beforehand. Therefore, we set out to compute the sum of transition probabilities over all possible path lengths $t_i \in \{0, ;\infty\}$. The random walk includes a stationary state reflecting the degree distribution of the nodes which we removed as this state is of no interest to our application. The stationary state corresponds to the first eigenvector Ψ_0 of T with the eigenvalue $\lambda_0 = 1$ [226]. We subtract Ψ_0's contribution from T and compute the aggregated (i.e. containing the sum over all path lengths) transition probability matrix $M = \sum_{t=0}^{\infty}(T - \Psi_0^T\Psi_0)^t$. Although this formula has a closed form solution [225] the computation of M consumes a large amount of memory since it is not sparse. However, M can also be approximated by using spectral decomposition of the transition matrix T as was shown in Haghverdi, Büttner, Wolf, et al. [225]:

$$M = \sum_{i=1}^{n-1}(\frac{\lambda_i}{1-\lambda_i} - \Psi_i^T\Psi_i)^t \tag{4.1}$$

To obtain an approximation of M without running into memory issues we used the first n=500 eigenvectors and ultimately retained only the part of M containing the transitions from the CpG sites S to all candidate genes $c \in C$, thereby obtaining rankings for all candidate genes. Finally, we obtained a single ranking R_c for each candidate gene in C by averaging the aggregated transition probabilities over all CpG sites, i.e. $R_c = 1/|S| \sum_{s \in S} M_{sc}$, for each $c \in C$.

In addition, we investigated the significance of the obtained scores R_c by comparing them to scores obtained from a randomized background setting. To this end, we executed the random walk analysis on $B > 100$ randomized graphs and calculated B scores R_c^b for all candidate genes. From these B scores, we then can calculated an

empirical P-value for each candidate gene c for a locus p given the maximum score $max_{c \in C}\ p_c^b$ of all candidate genes at that locus as $P(p_c) = 1/B \sum_{b \in B} I(p_c > max_{c \in C}\ p_c^b)$. Using the empirical P-value we then defined the set of significant candidate genes for each analyzed locus as $C^* = \{c | P(p_c) < 0.05\}$. To generate the randomized graphs we randomly sampled an equal number of $|S|$ CpGs S_b where we matched CpGs for methylation levels (mean and standard deviation, see also Section 4.2.8). We then extended the PPI graph P by adding all sampled CpGs s_b from S_b to construct the background locus graph $G_b = (V_P + S_b, E_P + E_D(S_b))$. By constructing random graphs we can empirically assess the probability of getting scores as extreme as the ones we observed. For this, we executed a random walk from a randomly sampled set of CpG sites to each of the candidate genes through the same PPI and ChIP-seq network as we used for the original analysis.

Candidate pathway extraction

Next, we sought to visualize our results from the network analysis in the context of the created locus graphs by extracting the part of the network best supported by the scores from the random walk. Therefore, for each locus graph G and for each node i, we defined weights w_i which aggregate the random walk score to arrive at node i starting from any of the CpG sites in S and of the scores for moving from node i to the candidate genes in C^*. We then normalized and inverted the w_i, i.e. set $w_i^* = max_i(w_i) - w_i$, to assign the highest-scoring nodes with the lowest weights and vice versa. This preparation was necessary to be able to apply a minimal node weights shortest path detection algorithm starting at the CpG sites S to the candidate genes in $C*$ based on all w^*, thus being able to extract optimal paths representing high random walk scores. We recorded all nodes on these minimal weights (maximum score) paths in the set Q, and subsequently defined a candidate pathway G_C for each locus as the sub-graph of G induced using only the nodes in the union of $C*$, Q and S.

Using functional data to corroborate random walk networks

For the random walk analysis we only used annotated PPI and TFBS data. However, having functional genomics data for Europeans (KORA) and South Asians (LOLIPOP) available we sought to further corroborate and extend our results by including these data in the network analysis. This allowed us to set our results which were originally obtained from QTL hotspots calculated in whole blood data into a functional context.

To this end, we again processed each of the 115 meQTL hotspots separately. We added CpG genes downstream of the *trans* CpGs for each locus to the generated locus graph (see Section 4.2.9) to be able to investigate which genes are ultimately affected by the *trans*-acting locus. For this, we included all genes overlapping or in direct vicinity (upstream and downstream) of a CpG, thereby retrieving at most 3 'CpG genes' per *trans* CpG. We then collected all functional data from both cohorts for each particular hotspot, i.e. genotypes for the genetic locus, gene expression for all collected genes and

methylation data for all *trans* CpGs, and calculated correlations between 1) the SNP and the SNP genes, 2) all genes in the network and 3) *trans* CpGs and CpG genes. All data were pre-processed as described in Section 3.2.

In this analysis it is possible for genetic variation in *cis* of the included genes and CpGs to influence the observed expression and methylation values, respectively, and hence to confound the association tests. In order to get rid of these confounding effects we used linear regression to adjust expression and methylation data for *cis*-meQTL (identified in this study) as well well as previously reported *cis*-eQTL [81] by regressing out the SNP effect. Genes and CpGs for which no *cis* acting SNPs were identified were used 'as-is' in the analysis. Next, we tested all associations between genotypes, expression and methylation in both cohorts separately. Correlation results from the distinct cohort data were subsequently meta-analyzed using inverse-effect meta-analysis (see Section 3.1.4). P-values obtained from the meta-analysis were then adjusted for multiple testing using the Benjamini-Hochberg procedure [203] to control the false discovery rate (*p.adjust()* method in R with parameter *method='BH'*, see also Section 3.1.6).

Finally, we constructed a network from the obtained associations which contains the SNP, genes, and CpGs as nodes and where edges between the nodes indicate significant correlations ($FDR < 0.05$). All edges between CpGs and CpG-genes present in this network were then added to the respective locus' candidate pathway identified in the random walk analysis (Section 4.2.9) and already existing edges were annotated with whether or not they showed correlation in the functional data. This procedure enabled us to use evidence from functional data to 1) reinforce edges already in the random walk networks and to 2) add additional edges connecting so far absent CpG genes to the CpGs of the network, thereby adding one additional layer of information.

4.2.10. Experimental validation of novel regulators

As a proof of concept and in order to experimentally validate the candidate gene *ZNF333* identified at the *rs6511961* locus, we performed a ChIP-seq experiment to investigate the binding of the *ZNF333* protein at *trans* associated CpGs of the locus. The experiments (ChIP-seq and IP-MS) and analysis of the ChIP-seq data were executed by our collaboration partners from the AME, whereas we performed the statistical over-representation analysis of the *ZNF333* interactome derived from the IP-MS experiment.

ChIP-seq analysis of ZNF333 binding sites

Raw sequencing data from the ChIP-seq experiments (see Appendix B.1) were mapped to the human reference genome (hg19) using the burrows-wheeler aligner (BWA, H. Li and Durbin [227]) and possible polymerase chain-reaction (PCR) duplicates removed from the aligned sequences. At the binding sites of the immunoprecipitated protein on the DNA an accumulation of mapped reads can be observed and the identification of so called 'peaks', i.e. regions of a statistically enriched number of sequence reads indicating protein binding, is a crucial analysis step for ChIP-seq data. In this case, we

used the peak-calling implemented in the *Dfilter* tool [228] to identify significant peaks across the genome setting parameter values $k_s = 60$, $b_s = 100$ and $l_{pval} = 6$. Finally, we investigated the overlap between the *trans* associated CpGs of the ZNF333 locus and the called *ZNF333* ChIP-seq peaks obtained from taking the union of the Myc and FLAG based experiments (see Appendix B.1 for details). Overlaps were defined for ChIP-seq peaks which overlapped the region 250bp upstream or downstream of the respective *trans* CpG. We further determined statistical significance of the overlap based on random background modeling to generate a null distribution and then applying Fisher's exact test on the derived contingency table. Lastly, we performed a sensitivity analysis with respect to the chosen window size around the peak, evaluating windows starting from 100bp up to 1,000bp with 100bp steps in between. We found that enrichment is robust with respect to the interval size around the *ZNF333* peaks (3.0 fold enrichment for 100bp, 3.4 fold enrichment for 1,000bp).

Over-representation analysis in *ZNF333* interactome

If *ZNF333* protein is indeed binding directly at the *trans* associated CpGs and associating with other proteins to form local chromatin complexes, we would expect to observe physical interactions between the *ZNF333* protein and the TFs identified in the network analysis or one of their direct interaction partners from the PPI network. To investigate this, we set out to evaluate whether transcription factors binding at the *trans* CpGs collected in the network analysis (set P_{ChIP}), show enrichment for the proteins pulled-down together with *ZNF333* in the IP-MS experiment (set P_{ZNF333}, interactor shortlist, see Section 3.3.2 and Appendix B.2 for details). In addition to these proteins, we included proteins indirectly associated to one of the ChIP-seq TFs via one intermediate step by utilizing the STRING PPIs (as defined under Section 4.2.9) to extend the two sets of proteins, P_{ChIP} and P_{ZNF333}. We defined two new sets P_{ChIP_ext1} and P_{ZNF333_ext1} by adding proteins which show a direct PPI to at least one of the proteins in the respective set.

Finally, we constructed a 2×2 contingency table of a random background set of proteins with the two protein lists to test for over-representation of the network derived set (P_{ChIP_ext1}) in the IP-MS derived set of proteins ($PZNF333\_ext1$) by use of a Fisher test. We formed a null (background) set of proteins from all TFs initially included in the network analysis, extended by their nearest neighbors from the STRING PPIs (set BG_{ChIP_ext1}). To construct the final table we count the overlaps between the P_{ZNF333_ext1} and P_{ChIP_ext1} protein sets with the background set, i.e. we determine the total number of proteins in both sets overlapping and not overlapping the BG_{ChIP_ext1} set. On the final table, we then applied a Fisher's exact test using the *fisher.test()* function in R (parameter alternative='greater').

In addition, we sought to analyze if proteins in the P_{ChIP_ext1} set show overall stronger signals (fold changes) in the IP-MS data in comparison to all other pulled-down proteins. To achieve this we obtained a ranking for all pulled down proteins (not only the shortlist, set $P_{ZNF333\_long\_ext1}$) using the fold changes between the *ZNF333* antibody and the IgG

control (see Section B.2) We obtained a P-value indicating whether the P_{ChIP_ext1} proteins found in the $P_{ZNF333\_long\_ext1}$ show overall stronger signal (higher fold changes) as the rest of the proteins in the $P_{ZNF333\_long\_ext1}$ set (excluding P_{ChIP_ext1}) by applying a two-sample Wilcoxon test (Mann-Whitney test) on the fold changes obtained for both sets (*wilcoxon.test()* in R with parameter alternative="greater"). We removed any "zero" fold changes and performed a log_{10}-transform prior to calculating the Wilcoxon-test.

Lastly, we set out to identify enriched gene ontology terms (GO terms) for the proteins identified in 1) the IP-MS experiment and 2) the transcription factors matched from the network analysis. For this, we conducted a gene ontology enrichment analysis for both of the two sets of proteins, i.e. the set P_{ZNF333} and the subset of proteins P_{ChIP_ext1} which was also identified in the pull-down analysis (same as for the Wilcoxon test). As a background, we used the long-list of proteins from the IP-MS experiment for set P_{ZNF333} and the one extended by the respective nearest neighbors in the PPI network (set $P_{ZNF333\_long\_ext1}$) for set P_{ChIP_ext1}. Enrichment was calculated for the two lists against all GO terms in the three GO categories (cellular component, molecular function, and biological process) using a hypergeometric test as implemented in the *hyperGTest()* function of the R-package *GOstats* (v 2.52.0). Significantly enriched GO terms were determined by setting an FDR threshold for the FDR adjusted P-value of $P_{adj} < 0.05$.

4.3. Genome-wide analysis of genetic effects on DNA methylation

For this project, we used the genotype and methylation data from European and South Asian population cohorts introduced in Section 2.2, i.e. data from the KORA, LOLIPOP, NFBC, and SYS cohorts amounting to 6,994 individuals in total. We aimed to derive an ethnicity independent ('cosmopolitan') set of genome-wide meQTL by measuring and processing all cohorts separately and then meta-analyze individual association results, collecting only SNP-CpG pairs showing significant associations in both ethnicities. Figure 4.2 shows the analysis plan followed in this project and the details on how computations were performed are given in methods Section 4.2.1. In this section, we highlight the results obtained by following this analysis plan, summarizing the cosmopolitan meQTL findings and the results of the subsequently applied two-step pruning strategy which was employed to obtain independent meQTL loci. Moreover, we will show that identified meQTL replicate well in independent data sets, both within the same and across different tissues. Finally, we describe our findings for the functional enrichment and network analyses we performed to unravel the functional mechanisms underlying meQTL associations. Specifically, we will also detail two of the main gene regulatory networks which we identified and used to explain *trans* regulatory hotspot signatures.

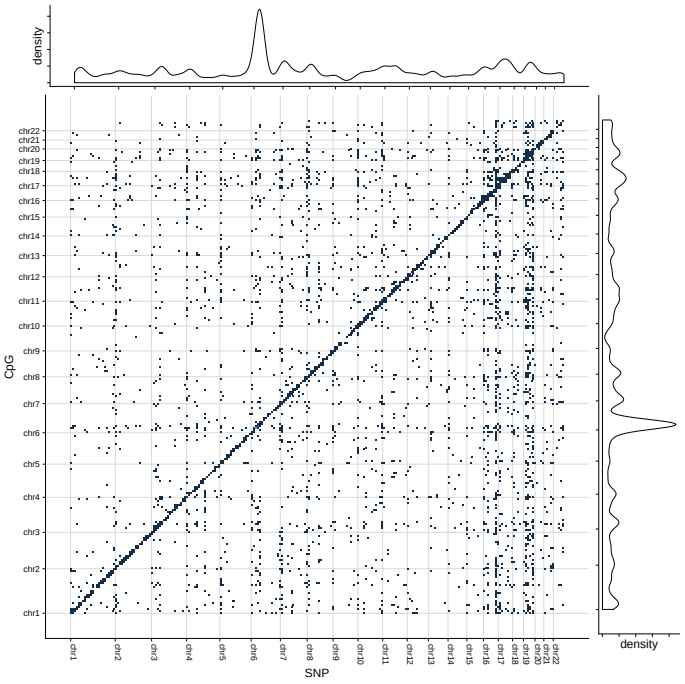


Figure 4.3.: Overview of all cosmopolitan methylation quantitative trait loci and their distribution across chromosomes. The x-axis shows SNP locations, the y-axis the CpG locations. Margin plots show densities of the number of SNPs (top margin) and the number of CpGs (right margin) summarized over all chromosomes. Dots indicate binned meQTL pairs (250 bins in both x and y direction).

4.3.1. Identification of a cosmopolitan set of meQTL

We generated a global cosmopolitan set of methylation quantitative trait loci encompassing *cis*, *longrange*, and *trans*-meQTL. Figure 4.3 shows the distribution across the genome of all cosmopolitan meQTL pairs. The diagonal indicates the large number of *cis*-meQTL (n=10,346,172 pairs). The *longrange* (n=351,472) and *trans* pairs (467,915) are overall distributed equally across the genome with an exception on chromosomes 17 and 19, which show a stronger signal of distal meQTL. Interestingly, chromosome 6 shows an overall higher density for SNPs and CpGs which could be caused by the Human Leukocyte Antigene (HLA) locus on this chromosome. Indeed, of all *cis* pairs, 1,431,186 pairs (13.3%) have a SNP located within the HLA region (6p21.3-22.1, [229]) although its size amounts to only about 0.3% of the human genome.

Due to local correlations between SNPs (linkage disequilibrium, see Section 2.1.1) and between neighboring CpG sites the meQTL analysis led to redundant SNP-CpG pairs that effectively represent the same methylation quantitative trait locus. We employed a two-step pruning approach to identify independent SNP and CpG loci over all cosmopolitan pairs which we termed 'sentinel pairs', or with respect to the individual entities sentinel SNPs and sentinel CpGs (see Methods). We applied this approach separately to SNPs and CpGs for each of the defined meQTL categories. This resulted in a conditional set of 84,456 genetic loci that are associated with at least one CpG site after the conditional analysis and yielded a final set of 77,953 pruned pairs with

1. 34,001 independent genetic loci associated with 46,664 independent methylation loci in *cis*

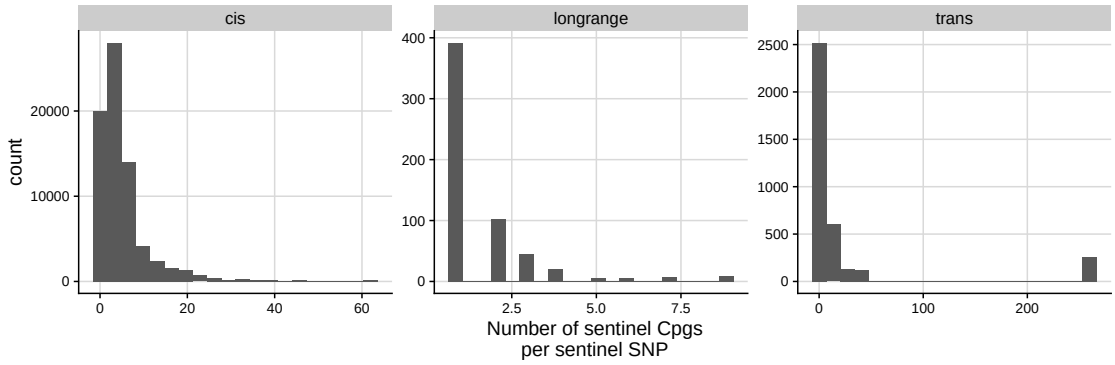


Figure 4.4.: Histograms showing the number of associated sentinel CpGs for all sentinel SNPs in the three meQTL categories *cis*, *longrange* and *trans*.

2. 467 independent genetic loci associated with 499 independent methylation loci in *longrange*

3. 1,847 independent genetic loci associated with 3,020 independent methylation loci in *trans*

after the LD pruning. For each of the identified loci we defined the SNP/CpG with the strongest association as representative sentinel SNP and CpG sites for downstream analysis. Figure 4.4 gives an overview of the sentinel pairs in each category. For each category, numerous SNPs are associated with more than one sentinel CpG (N: *cis*=14,043, *longrange* = 75, *trans*=466). The sentinel SNP with the most *trans* associations is linked to a total of 261 methylation sites in *trans*, with the next highest having a total of 46 associations.

4.3.2. Replication in isolated leukocytes, adipocytes and adipose tissue

Since the identification of meQTL was performed in whole-blood data containing diverse blood cell types we confirmed that the observed associations were not due to unobserved differences in cell subset composition. For this, we tested all cosmopolitan SNP-CpG pairs in independent data generated from isolated white blood cell subsets separated by fluorescence-activated cell sorting (monocytes, neutrophils, CD4+ lymphocytes, CD8+ lymphocytes). The experimental part of this analysis was performed by our collaboration partners (see Section 3.3.1) and the respective overlap analyses conducted by Rory Wilson (see Section 4.2.2). With $N = 57$ samples each, we expect to recover 30% of the meQTL with an effect size of 2.0% change in methylation per allele copy (at $P < 0.05$, MAF=20%). We find that 26%-37% of the 11.2M cosmopolitan SNP-CpG pairs replicate in the white cell subsets at $P < 0.05$ (26-37% for *cis* pairs, 21-28% for longrange *cis* and 27-37% for *trans*) in line with our expectations. Furthermore, we assessed whether replicated associations also show a consistent direction of effect which we observed for 80%-87% of all cosmopolitan pairs. We determined statistical significance at $P < 2.2 \times 10^{-16}$ for

all cell subsets by applying a binomial test, i.e. testing how likely it is to obtain the same number of hits or more given a random background. Figure 4.5 shows a comparison of effect sizes between the cosmopolitan hits and the respective individual white cell subsets.

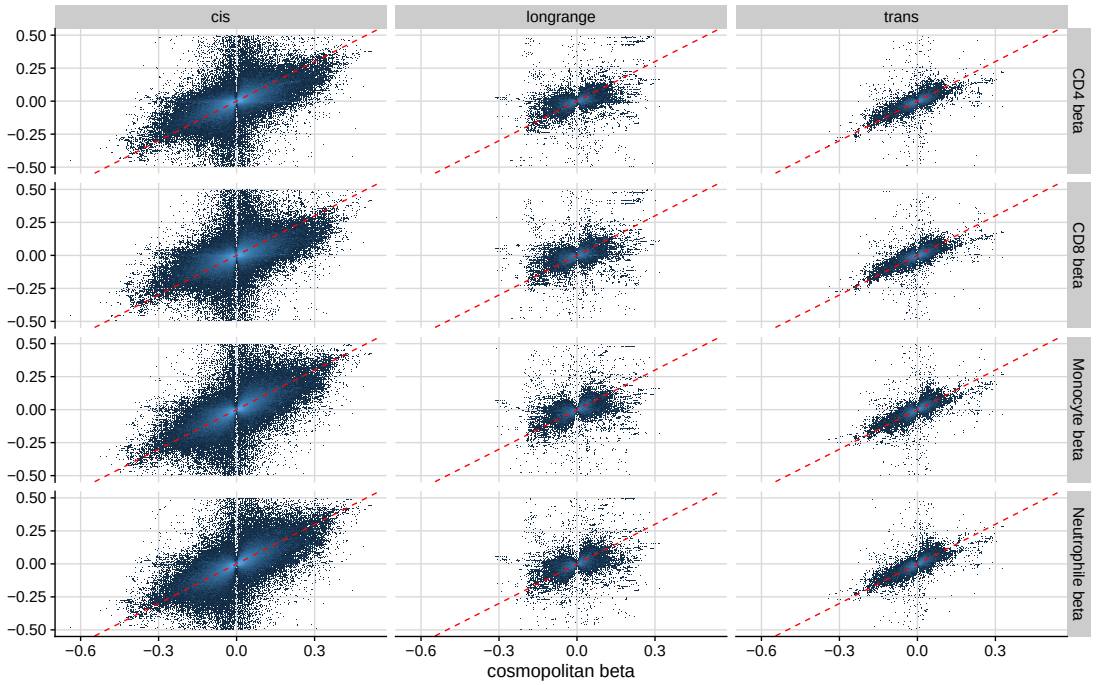


Figure 4.5.: Comparison of effect sizes for meQTL derived in isolated white cell subsets and the ones derived from whole-blood in our study. x-axis always shows cosmopolitan effect size, y-axis effect size in isolated cells. Results are stratified in *cis* , *longrange* and *trans* (columns). Plots are divided in 500 bins, fill color indicates region density (lighter blue indicates higher density and vice versa) and red lines show the diagonal for each plot.

We further sought to assess whether the identified meQTL SNPs exhibit their effects also in different contexts such as a tissue different to blood. Therefore, we obtained DNA methylation measurements from isolated subcutaneous and visceral adipocytes for $N = 47$ samples each and adipose tissue from the MuTHER cohort with $N = 603$ samples (see Section 3.3.1). Similar to the within tissue replication, this analysis showed a consistent direction of effect for all three datasets ($P < 2.2 \times 10^{-16}$ using a binomial test), amounting to 72%-86% of all cosmopolitan meQTL tested. The comparison of effect sizes for this analysis is shown in Figure 4.6.

Similar to the isolated blood cell data we replicated 19.2% of meQTL in isolated visceral and 19.4% in subcutaneous adipocytes ($P < 0.05$ and same direction of effect). These proportions are also consistent with the expectations based on the sample size, where we would expect to recover 25% of SNP-CpG pairs with an effect size of 2.0% at $P < 0.05$ for SNPs with a minor allele frequency of 20%. For the adipose tissue, we replicated 44.2% of all pairs for the same replication criteria.

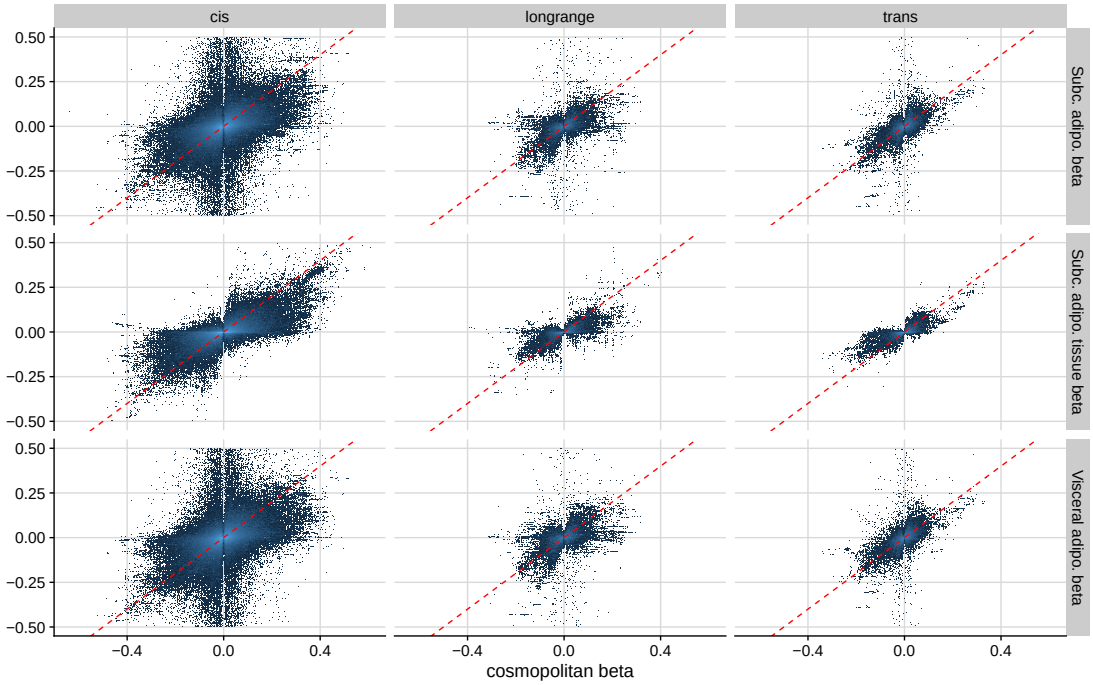


Figure 4.6.: Summary of cross tissue replication of cosmopolitan meQTL. x-axis shows cosmopolitan effect size (from whole-blood tissue) and y-axis effect size obtained from replication in isolated visceral and subcutaneous adipocytes as well as in subcutaneous adipose tissue. Results are stratified in *cis* , *longrange* and *trans* (columns). Plots are divided in 500 bins, fill color indicates region density and red lines show the diagonal for each plot.

We thus provide strong evidence that the observed genetic effects on DNA methylation are indeed 1) independent of the variation in cell subset composition and 2) shared across diverse cell types.

Lastly, in order to show that our associations are independent of the platform used for quantification of methylation, we performed an additional replication using previously published MeDIP-seq methylomes obtained from peripheral-blood [220]. Relatively few of the MeDIP-seq based associations were testable in our results ($N = 328$, see Section 4.2.2 for details). We replicated a total of 155 of these pairs (47%, $P < 0.05$) which indicates a significant proportion at $P < 0.01$ according to an empirically derived null distribution based on matched background SNP-CpG pairs (see Section 4.2.2, the background failing to identify any significant associations). These results suggest that our meQTL also generalize well across platforms.

4.3.3. Functional enrichment analyses

Even though meQTL SNP-CpG pairs show statistical associations, this does not necessarily imply a functional relationship. Hence, a main focus of our work was the functional annotation and understanding of the discovered meQTL pairs for which we

considered each of the three categories (*cis*, *longrange*, *trans*) separately. In addition to the functional data collected in this study, we curated additional public datasets for each of the analyses: To determine functional relevance of *cis*-meQTL we used the 15 state chromHMM genome segmentation from the Roadmap Epigenomics project [48, 49]. For *longrange*- and *trans*-meQTL we further leveraged Hi-C and promoter-capture Hi-C data from the BLUEPRINT consortium [75, 230] and to further corroborate *trans*-meQTL findings, we utilized a collection of transcription factor binding sites from the ReMap catalogue [71] and ENCODE [60]. All public data were matched with respect to their tissue of origin, i.e. we only used data established from blood related cell-lines or from primary blood cells in order to be able to integrate these with our data.

cis-meQTL pairs show enrichment in functional chromatin states

The genomic closeness of *cis*-meQTL SNPs and CpGs implies a potential functional relationship due to them being located in a common regulatory element. A possible explanation is that the observed polymorphisms change the sequence in the regulatory element, leading to altered transcription factor binding and gain or loss of DNA methylation (see Figure 4.1). To investigate this hypothesis we utilized chromHMM states, functional chromatin states derived from combinations of histone modifications indicating e.g. closed or accessible chromatin regions and performed an enrichment analysis of *cis*-meQTL in enhancer (potential distal regulatory element) and promoter (potential *cis* regulatory element) states. We generated a null distribution of matched background meQTL pairs and evaluated whether observed meQTL SNPs and CpGs reside more often in the same state as would be expected by chance (see Section 4.2.3 for details).

The results for this analysis are shown in Figure 4.7 and display a clear shift of the distribution of overlap fractions for 'observed' pairs compared to the background. Overall, meQTL are significantly enriched for being in the same regulatory state in comparison to the matched background ($P < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test), suggesting that observed *cis*-meQTL indeed tend to reside within the same regulatory element. This further hints at a specific mechanism of action, namely that the genetic variants act in *cis* via directly changing the DNA sequence of the regulatory element, affecting gene expression for instance through a change in transcription factor binding and direct impact on DNA methylation.

Longrange and *trans* SNP-CpG pairs are enriched in chromosomal contact regions

Especially for *longrange*- and *trans*-meQTLs it is difficult to establish a direct functional relationship due to the relatively large genomic distance between the involved entities. In this study, we identified independent *longrange* and *trans* genetic effects on DNA methylation for 467 loci and 1,847 loci, respectively. A possible mechanistic explanation for these distal effects is the 3D structure/folding of chromosomes which can connect

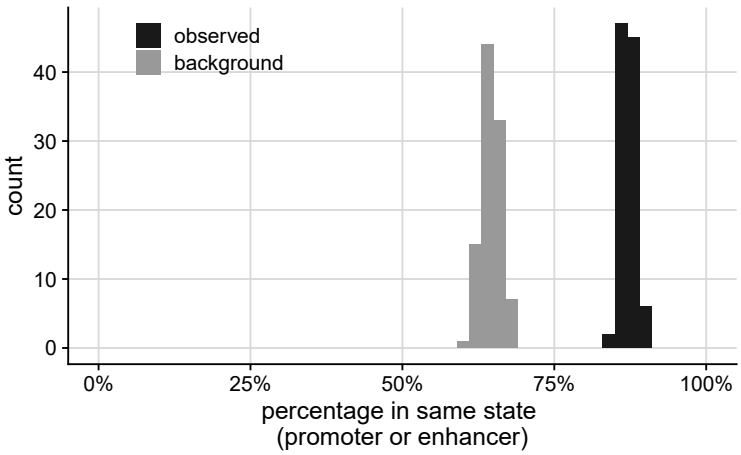


Figure 4.7.: Histograms showing the amount of meQTL and sampled background pairs residing in the same regulatory class (enhancer or promoter) for 100 iterations. y-axis shows total counts per bin, x-axis the percentage of matching states for each sampling. Dark grey distribution indicates the observed meQTL pairs, light grey the corresponding background pairs.

distal DNA regions (compare Figure 4.1). In recent years, 3D chromatin structure has been the focus of many works in the field of genomics due to the development of NGS assays, such as HiC, which make it possible to determine intra (within the same chromosome) and inter (between different chromosomes)-chromosomal chromatin contacts [69, 73–75, 231, 232]. It has been shown previously that, for instance, epigenetically marked enhancers are linked to active gene promoters and that genetic variants at chromosomal contacts are associated with the expression of genes they are in contact with [75]. Here, we used published Hi-C and promoter-capture Hi-C (PCHi-C) contacts as well as published topologically associating domains (TADs) in primary blood cells [75] to assess the potential functional relationship of *longrange*- and *trans*-meQTL pairs. TADs are derived from intra chromosomal contact information and reflect regions of high contact frequencies on the same chromosome e.g. established via chromatin loops [233]. We generated a matched background individually for the *longrange*- and *trans*-meQTL pairs and assessed whether the observed pairs are more often located in the same TAD/PCHi-C contact (*longrange*) or Hi-C contact (*trans*) as compared to the background based on odds ratios (details in Section 4.2.4). The results of these analyses are summarized in Figure 4.8.

For all three analyses, the results show a clear enrichment of observed meQTL pairs compared to the sampled background with an empirical P-value amounting to $P < 6.6 \times 10^{-3}$ in each case (100 samples, see Methods). This enrichment of pairs for both *longrange*- and *trans*-meQTL suggests an involvement of 3D chromatin interactions to establish the functional relationship between the individual meQTL entities and underlines the regulatory importance of our findings.

Enrichment of observed meQTL entities for association with gene expression

We set out to establish the functional relevance of our findings by assessing, whether or not SNPs and CpGs which show at least one significant meQTL association are more often also associated with gene expression changes as expected by chance, implicating

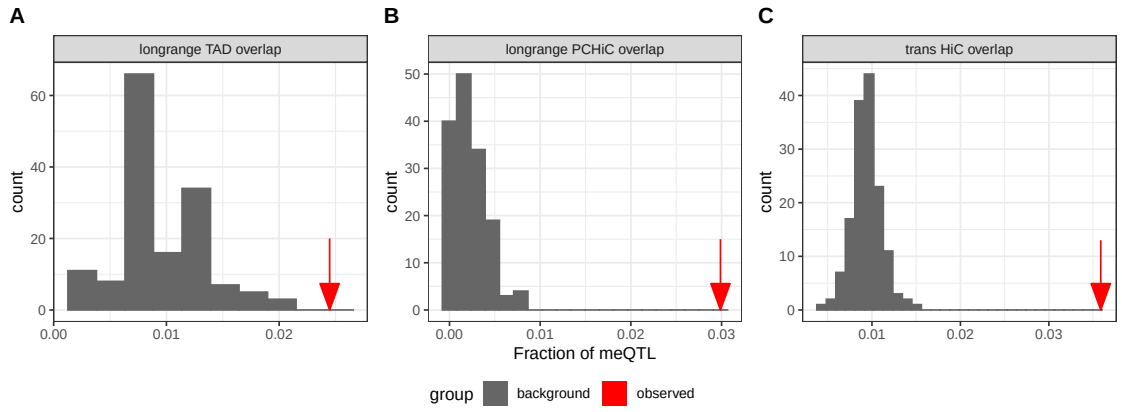


Figure 4.8.: Enrichment of *longrange*- and *trans*-meQTL pairs in HiC data. X-axes show the fraction of pairs in the same TAD (panel A) or in chromatin contacts (panel B and C) for 150 sets of sampled background pairs (grey distribution) and the observed pairs (red arrows). Panels A and B show results for *longrange* pairs, panel C for *trans* pairs.

functional relevance (impact on gene expression) of the identified entities. To this end, we calculated expression quantitative trait loci (eQTL) and expression quantitative trait methylation (eQTM, association between gene expression and DNA methylation) for all independent meQTL SNPs and CpGs. We avoided confounding by genetic effects for eQTM associations by regressing out previously identified *cis* -eQTL and -meQTL and generated a null distribution of background meQTL pairs using random, but matched, SNPs and CpGs (see Section 4.2.5). We stratified SNPs and CpGs according to their effect sizes in the meQTL associations to address the question whether enrichments would vanish for low effect size meQTL.

Figure 4.9 shows the results for this analysis for each of the three meQTL categories (*cis*, *longrange* and *trans*) and stratified by effect size (low, medium and high), separated by SNP (eQTL) and CpG (eQTM) enrichments. All categories show significant enrichment for SNP expression association. For CpGs, on the other hand, all categories except *longrange* (low) indicate depletion for gene expression association. These results indicate that meQTL SNPs impact gene expression and thereby indeed have a functional relevance. Also, association with gene expression is likely established through a genetic (SNP) effect rather than through a direct epigenetic (CpG) effect.

Enrichment of known regulatory genes in *trans*-meQTL loci

We investigated the type of genes encoded in the respective *trans* regulatory LD blocks, i.e. in the regions around the identified sentinel SNPs. Specifically, we sought to test whether the *trans*-acting genetic variants are over represented for genes involved in epigenetic processes, DNA binding, or transcriptional regulation, highlighting their functional importance. For this analysis, we used the three gene lists presented in

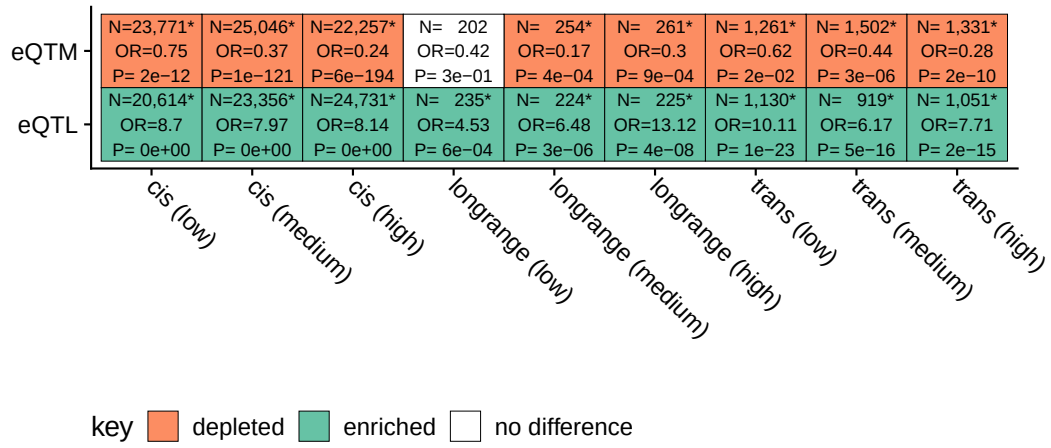


Figure 4.9.: Enrichment of meQTL for association with gene expression. Rows indicate eQTM (for meQTL CpGs) and eQTL (for meQTL SNPs) enrichment and columns indicate stratification by category (*cis*, *longrange*, and *trans*) and by effect size of meQTL association (low, medium, high). Cell contents indicate the total number of meQTL (N), odds ratio (OR) and Fisher P-value (P) per stratum.

Lemire, Zaidi, Ban, et al. [211], which entail a list of epigenetic regulator genes, a list of curated TFs and a list of zinc-finger genes as a subset of these transcription factors (see Section 4.2.7 for details). We compared the number of genes in the *trans*-meQTL LD blocks overlapping the respective gene sets to the number of overlapping genes obtained from random but matched background regions and assessed enrichment in the observed meQTL based on odds ratios (see Methods). The results of this analysis are shown in Figure 4.10. For all three gene sets a strong enrichment of the regulator genes in the sentinel regions defined by our meQTL as compared to the sampled background regions is evident (empirical $P < 5.99 \times 10^{-3}$ for all three gene sets). This further corroborates the functional relevance of the *trans*-meQTL identified in our study.

Trans associated CpG sites are enriched for TF bindings sites

For 1,847 genetic loci, we detected an influence of the genetic variation on DNA methylation in *trans*, with the total number of associated *trans* CpG loci ranging from 1 to 298 in the cosmopolitan set of meQTL. A possible explanation for these observed associations is that transcription factors (TFs) mediate the observed inter-chromosomal effects. This is a mechanism of action that has also been proposed by Bonder, Luijk, Zhernakova, et al. [23] and Lemire, Zaidi, Ban, et al. [211]. To assess the validity of this hypothesis, we set out to identify the respective TF bindings at the *trans* associated CpG-sites. For this, we used publicly available binding site information derived from uniformly processed ChIP-seq data gathered by ENCODE [60] and in the ReMap database [71] and performed an enrichment analysis for the obtained binding sites (Supplementary Figure C.2, see

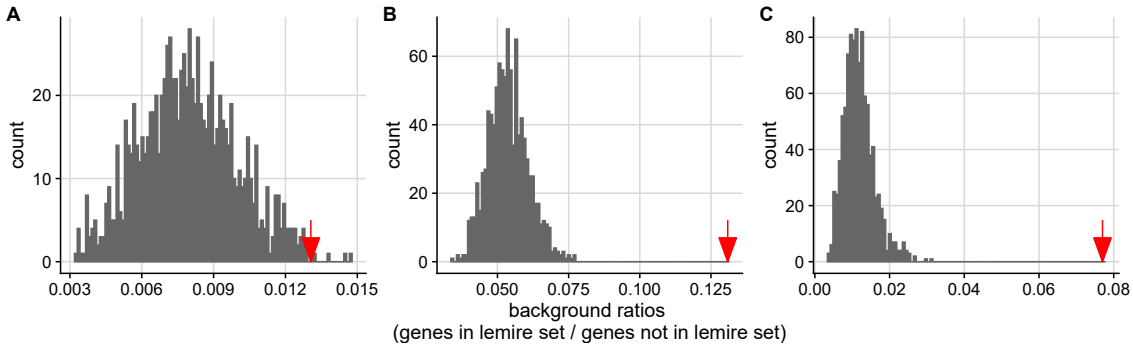


Figure 4.10.: Histograms showing the functional enrichment of *trans* regulators. Panels show enrichments for the three gene lists presented in [211] which represent epigenetic regulators (A), transcription factors (B) and zinc finger genes (C), respectively. The grey histogram indicate the distribution of the overlap fraction between the genes making up the gene lists and the genes at sampled background SNPs for the 1,000 iterations. Red arrows highlight the fraction of overlap observed between genes present in the regulatory sets and the trans regulator associated genes.

also Section 4.2.8). We utilized the TFBS of 246 transcription factor data sets obtained from blood-related cell lines. The analysis was limited to 115 *trans*-meQTL hotspots, i.e. *trans* genetic loci with a *trans* effect involving at least 5 associated CpG sites. Of the 115 *trans*-meQTL hotspots, 45 (39%) show enrichment of binding sites for at least one TF at their respective *trans*-CpGs. We then assessed whether these numbers indicate significant enrichment under the null hypothesis by evaluating the TFBS overlap for a sampled background set of *trans*-CpGs (see Section 4.2.8). Our result of n=45 TFBS enriched loci represents a 1.8 fold enrichment compared to expectation under the null hypothesis (n=25) generated from the random background (binomial test $P = 7.4 \times 10^{-6}$, see Methods for details).

For a total of 4 of the 45 enriched sentinel loci we further identified nuclear transcription factors at *trans* CpG sites which are encoded in *cis* of the respective sentinel SNP (see Figure C.2). These loci and transcription factors include the previously reported NFKB1 and CTCF [23, 211] as well as the novel REST and NFE2 loci (Fisher's exact test $P = 1.7 \times 10^{-5}$ to 3.4×10^{-89}). Therefore, for these four loci, our data indicate that the transcription factor encoded at the locus directly mediates the observed *trans*-meQTL signature.

4.3.4. *Trans*-meQTL reveal novel regulatory patterns

A total of 41 loci remain for which we could not find a direct explanation of the genome-wide genetic effects via a *cis* encoded transcription factor. For these 41 loci an important next step is to identify the means by which the genetic variant affects DNA methylation in *trans*. Based on the initially identified statistical associations, however, it is not immediately evident which gene is influenced locally (i.e. in *cis*) by the regulatory SNP and hence is the causal gene to mediate the genome-wide methylation changes. Although in some cases it might be the gene closest to the SNP, in other cases it could

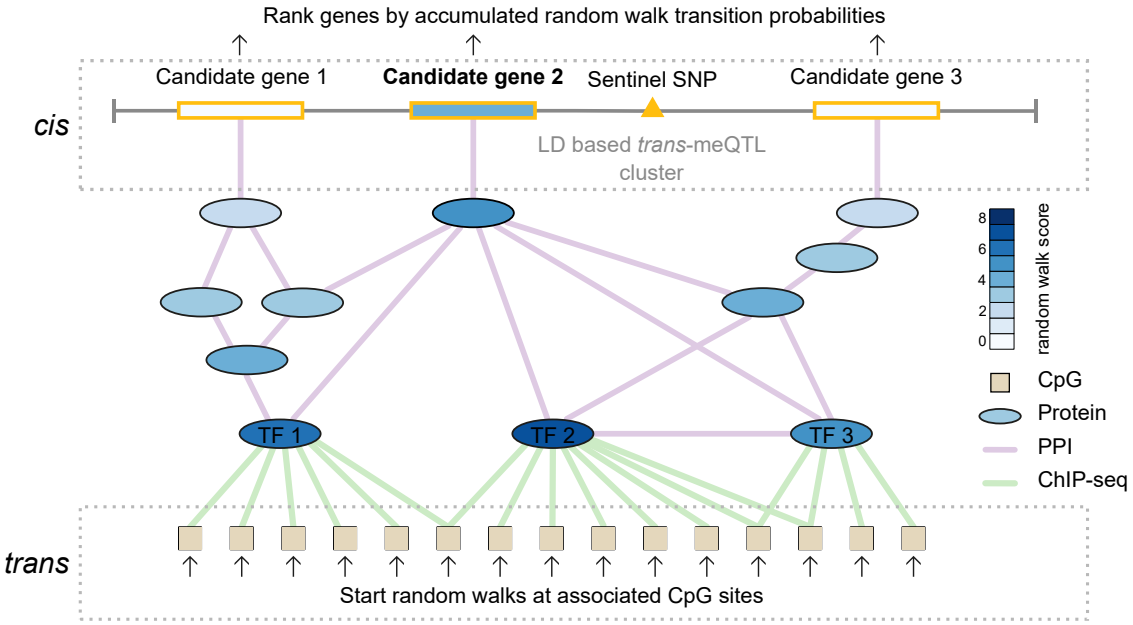


Figure 4.11.: Schematic illustrating the random walk approach. For each *trans*-meQTL locus, we connect genes located in *cis* to the sentinel SNP with the *trans* CpGs using transcription factor binding sites curated from ChIP-seq data (green edges) and a context specific protein-protein-interaction network obtained from the STRING database (purple edges). A random walk to prioritize SNP genes is then applied starting at the *trans* CpGs and accumulated scores for the SNP genes are assessed. Figure adapted from Hawe, Wilson, Loh, et al. [1]

be a more distant gene (e.g. due to local chromatin structures, see also Section 4.2.4) and a globally acting regulatory pathway is likely at the root of these hotspots. We hence set out to identify the most likely candidate gene and corresponding regulatory pathway for the remaining 41 *trans*-meQTL hotspots. To this end, we employed a two-step network analysis approach based on random walks, where we integrated publicly available protein-protein interaction (PPI) networks and transcription factor binding sites (TFBS) with the functional association data (SNP-methylation, SNP-expression, and methylation-expression) established in this study (see Section 4.3.4, Figure 4.11). We applied this approach individually to all 41 hotspots in order to identify the most likely causal candidate gene from the sets SNP genes (genes within 1Mbp of the SNP) and to connect the respective SNP genes to the associated *trans* CpGs via the PPI network and TFBS. For this, we created a high confidence PPI network specific to our whole-blood analysis context (see Methods).

We then prioritized genes in the *trans* genetic locus by a score that reflects the probability of reaching each gene through the curated PPI/TFBS network from the associated CpG sites (see Section 4.2.9). We established statistical significance for the selected candidate genes being connected to the respective set of associated *trans* CpG sites by comparing the observed random walk scores to a null model of scores calculated for a set of randomly sampled CpGs of the same size for each locus. This strategy

identified candidate genes with their corresponding regulatory molecular network for 19 independent *trans*-meQTL loci. To check, whether the identified candidate genes indeed represent relevant functional entities, we assessed the number of *trans*-eQTM (expression quantitative trait methylation) for each locus by associating the *trans* CpG methylation with the expression of each of the putative candidate genes. We determined significant enrichment of *trans*-eQTM for the analyzed candidate genes in comparison to all remaining genes encoded at the respective meQTL loci ($P = 4.5 \times 10^{-6}$, Wilcoxon test), further strengthening the results from our network analysis approach.

Finally, in order to illustrate the results of our strategy we choose the genetic locus identified by the sentinel SNP *rs730775* which is associated with 49 *trans* CpG sites. The corresponding locus network obtained from the random walk analysis is depicted in Figure 4.12A. Here, we selected the gene NFKB inhibitor epsilon (*NFKBIE*) (empirical $P < 0.01$) as the most likely *trans*-acting candidate gene. The SNP is a *cis*-eQTL for *NFKBIE* in whole-blood (eQTLGen $P = 1.2 \times 10^{-23}$) and is situated in the first intron of *NFKBIE*. *NFKBIE* is a direct inhibitor of *NFKB1* activity and shows significant co-expression with *NFKB1* in our data ($P = 2.2 \times 10^{-4}$), which further shows binding sites at 31 of the 49 trans-associated CpG sites (odds ratio=7.8, $P = 9.1 \times 10^{-7}$). In addition, the methylation levels for 7 of the 49 trans-CpG sites show significant association with the expression of 5 of the respective neighboring genes, all of which are also significantly co-expressed with *NFKBIE*. Moreover, the *trans* CpG sites of this locus are located close to other genes of the NFKB pathway, such as *TRAF6* and *IKBE*, and are further enriched for the gene ontology (GO) term 'regulation of interleukin-6 (*IL-6*) biosynthetic process' (GO:0045408; $P = 3.75 \times 10^{-05}$, hypergeometric test). The *trans*-acting locus at *rs730775* has previously been associated with rheumatoid arthritis (RA) [234], which has been characterized by Emery, Keystone, Tony, et al. [235] with *IL-6* mediated autoimmunity and which can be treated with drugs targeting *IL-6* [236]. We corroborated our results by performing a formal colocalization analysis using fastENLOC [237, 238], a Bayesian method to test for colocalization of molecular QTL with GWAS signals The analysis was implemented using each of the 49 *trans* CpG sites as molecular QTL and assessing SNP-wise posterior probability of a shared underlying causal variant between CpG sites and RA. The average posterior colocalization probability was 70%, providing strong support for a shared causal variant for the majority of CpGs and further strengthening our initial findings. Therefore, our findings highlight a potential regulatory mechanism for the association of genetic variation at the *NFKBIE* hotspot locus with rheumatoid arthritis, mediated through DNA methylation regulation at CpG sites in *cis* to genes important for the regulation of *IL-6* biosynthesis.

In addition to describing the novel *NFKBIE* locus, we also replicated and extended previous results for the known *trans* locus around the *SENP7* gene for the sentinel SNP *rs9859077* [211]. The random walk approach prioritized *SENP7* (Sentrin specific protease 7, empirical $P < 0.01$) as the most likely mediator gene for this *trans* locus. The identified regulatory network around the rs9859077/SENP7 locus and the associated

trans CpGs is shown in Figure 4.12. Located in an intronic region of SENP7, the 'C' allele of *rs9859077* is associated with increased mRNA expression of SENP7 in our data ($P = 3.1 \times 10^{-11}$) as well as in GTEx whole-blood data ($P = 3.0 \times 10^{-11}$) and in lymphoblastoid cell lines (LCLs, $P = 0.002$) [82, 239], which corroborates previous findings in CD4 and CD8 lymphocytes [211]. Moreover, the methylation of 85% of the *trans* CpG sites is correlated to the expression of *SENP7*, as estimated by Storey's π_0 method [240] on the *trans*-eQTM P-values. Furthermore, our integrated network analysis approach highlighted a potential regulatory molecular network which connects the *SENP7* locus to the respective *trans*-meQTL CpGs: *SENP7* interacts with *BCL6* [241], in turn interacting with the transcription factors (TFs) *YY1*, *PML*, *EBF1*, *SP1*, *SPI1* and *ELF1* [241–244]. All these TFs show DNA binding sites which overlap with the *trans*-meQTL CpG sites of the *SENP7* locus. Finally, similar to the *NFKBIE* locus the expression levels of 16 of the genes neighboring CpG sites are associated with the methylation of 43 of the 57 *trans*-CpGs of which 13 genes can be independently replicated in LCLs [239], whole blood [22, 245] or CD4 or CD8 lymphocytes [211]. These results hence replicate and largely extend previous findings which described the SENP7 locus, by supporting a molecular mechanism which underlies the regulation of DNA methylation at a cluster of zinc-finger genes on chromosome 19 by SENP7, possibly related to DNA repair mechanisms [246, 247]. Overall, our analyses showed that using our computational random walk approach in conjunction with functional data constructs functionally relevant networks and can be used to generate novel hypotheses about *trans*-meQTL mechanisms.

4.3.5. Experimental validation confirms novel regulators

We highlighted several candidate genes at multiple loci which have as of yet not been associated with genome regulation. For instance, we proposed *ZNF333*, the putative candidate gene at the genetic locus identified by the *rs6511961* SNP, as a candidate regulator potentially mediating the observed *trans* effects. In our data, *rs6511961* is an eQTL of *ZNF333* and the expression of *ZNF333* co-varies with the expression of genes known to encode for nuclear transcription factors (e.g. *TAL1*, *CDK9*). To validate the hypothesis that *ZNF333* is indeed DNA binding and mediates the observed relationship between *rs6511961* and its *trans* CpG signature, we performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) using FLAG/Myc-tagged ZNF333 constructs. Details on this analysis and the ChIP-seq data processing are given in Section 4.2.10. The results of the ChIP-seq analysis confirmed site-specific DNA binding ($P = 7 \times 10^{-3}$, Fisher exact test), including a high overlap of DNA binding regions for *ZNF333* between the FLAG and Myc tags (Supplementary Figure C.4). We further identified a putative binding motif for *ZNF333* based on the ChIP-seq data, *TG[AG]*TCA*, and the *ZNF333* binding sites are strongly enriched for motifs of other, known transcription factors ($P < 2.2 \times 10^{-16}$, including *FOSL2*, *Jun-AP1*, *FRA1*, *BATF* and *ATF3*). In addition, we found that 35% of the *rs6511961* associated *trans* CpGs are within or close to ($< 500bp$) *ZNF333* DNA binding sites, which represents an approximate 3-fold enrichment as

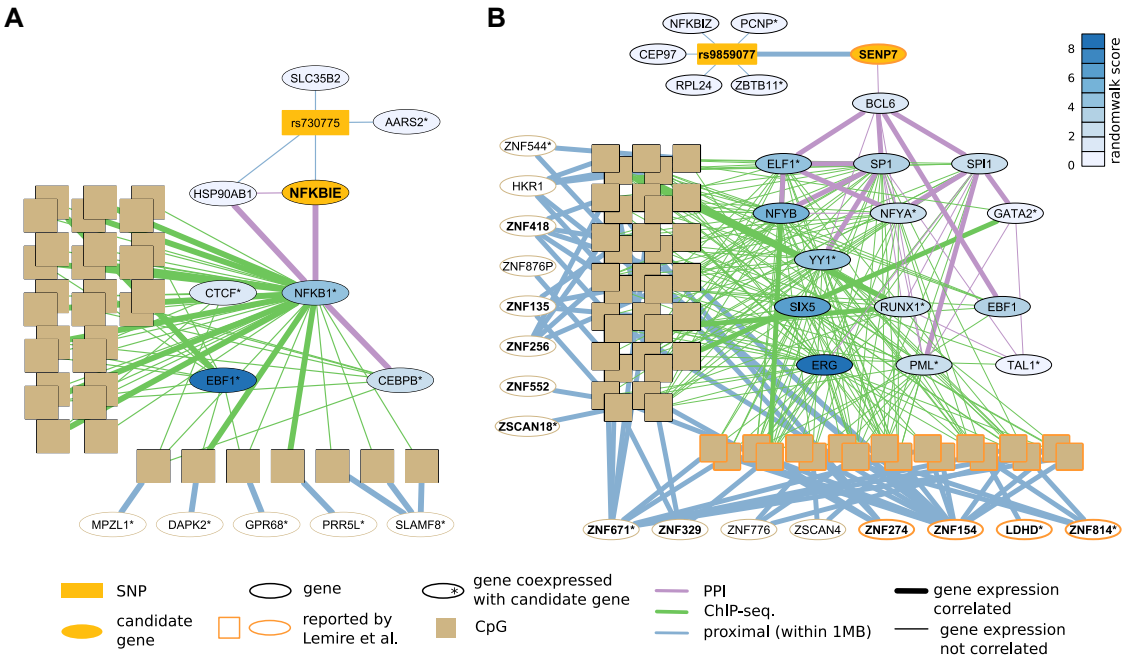


Figure 4.12.: Two networks inferred during the random walks analysis for the *NFKBIE* locus (panel A) and the *SENP7* locus (panel B). Yellow rectangles indicate SNPs, yellow ellipses SNP genes prioritized by the random walk. Blue edges indicate entities genomically close to one another, purple edges indicate PPIs, green edges TF binding at CpG sites. Bold edges illustrate observed correlations in functional data. CpG-genes indicate in bold font are *trans* associated to the locus SNP, genes marked with an asterisk (*) are co-expressed with the selected SNP gene. Figure adapted from Hawe, Wilson, Loh, et al. [1].

compared to what is to be expected under the null ($P < 0.05$, Fisher's exact test). As an additional validation step, we performed an immunoprecipitation mass-spectrometry (IP-MS) experiment to pinpoint potential binding partners of *ZNF333* (see Sections 4.2.10 and B.2). We were able to identify interacting proteins of *ZNF333* for which we could further observe gene ontology (GO) enrichment for terms related to 'nucleic acid binding and processing' ($FDR < 0.05$, hypergeometric test) as well as to the *MLL1* complex which is an important epigenetic modifier. Finally, we sought to analyze whether the transcription factors identified in the computational overlap analysis are enriched for inclusion in the direct *ZNF333* neighborhood as compared to a random background model (see Section 4.2.10 for details). Significance of the enrichment for the computationally determined proteins in the experimentally derived *ZNF333* neighborhood was assessed using both a Wilcoxon test on the IP-MS derived list of fold-changes ($P = 5.4 \times 10^{-5}$) and a Fisher test based on the constructed overlap contingency table ($P = 2.87 \times 10^{-11}$). Thus, the experimental analyses provided additional support of the hypothesis that the *trans* CpG signature of *rs6511961* is, at least in part, determined by the DNA binding protein *ZNF333*.

4.3.6. Effect of increased resolution and coverage on meQTL results

Finally, we investigated whether increased genome coverage and resolution with respect to DNA methylation could allow additional discoveries and whether the possible incomplete and non-random coverage of the 450k array could generate false-positive findings. For this, we performed additional meQTL calculations based on DNA methylation assayed using the EPIC array available in the KORA FF4 cohort (N=1,848) using the same analyses steps as for the original European discovery analysis (see Methods). Overall the 450k and EPIC array overlap in 406,501 CpG sites and the EPIC array assays an additional 381,605 CpGs. The EPIC analysis replicated 96% of the original discovery findings from the 450k array ($P < 0.05$, same direction of effect), indicating high data quality and good agreement between studies and platforms. We set out to systematically quantify to which extend conclusions of our functional analyses would change depending on the difference in resolution and coverage of both arrays. For instance, the availability of additional correlated markers on the EPIC array could 1) improve the functional characterization of sentinel markers and 2) identify previously not evaluated functional features. In addition, markers on the EPIC array independent of our sentinels could increase the overall discovery of genomic features. To determine which EPIC-specific markers correlate with the sentinel markers, we computed pairwise correlations between the additional EPIC-specific CpGs and our 49,580 sentinel CpGs. We determined the closest sentinel CpG for all N=381,605 EPIC-specific markers and obtained the respective R^2 from the methylation data. With increasing distance between CpGs we find that correlations vanish rapidly (Figure 4.13) in line with previous reports [248]. While 7,138 EPIC CpG markers showed $R^2 > 0.2$ with at least one of our sentinel CpGs, 98% (374,467) of EPIC-specific CpGs do not correlate and are therefore independent of our sentinels.

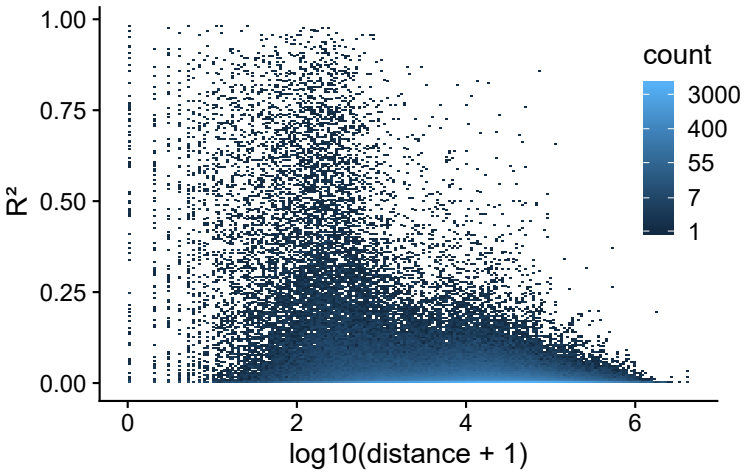


Figure 4.13.: Relationship between distance of sentinel and EPIC array specific CpGs (x-axis) and the correlation of their respective β values (R^2, y-axis). Zoomed in to $0 \leq X \leq 10^6$ on the x-axis. Increasing distance between entities leads to a drop in correlation. Figure adapted form Hawe, Wilson, Loh, et al. [1].

Effect of increased coverage on identification of TFs

We investigated the effect of increased genome coverage of CpG markers on the TFBS enrichment for *trans*-acting sentinels. For this we repeated the TFBS enrichment under inclusion of EPIC *trans*-meQTL results. Specifically, we defined four distinct sets of meQTL:

1. **450k**: EPIC meQTL also available on 450k array
2. **450k + correlated**: 450k set incl. EPIC-specific markers correlated with a 450k *trans* meQTL ($R^2 > 0.2$ at distance < 1Mbp)
3. **450k + independent**: 450k set incl. EPIC-specific markers independent of 450k *trans* meQTL ($R^2 < 0.2$ at distance < 1Mbp)
4. **EPIC**: All meQTLs discovered using the EPIC array

Definition of those sets allowed us to investigate the effect of improved local resolution (2) and addition of independent markers (3) and their combined effect (4) on TFBS enrichment as compared to the baseline (1).

Our results showed that EPIC-specific content improves the number of discovered transcription factors overlapping *trans* loci which originates both from correlated and independent CpGs (Table 4.2) and where independent markers provide the largest increase in identified TFs. All except one of the TFs found in the initial 450k based analysis were also identified in the full set of EPIC *trans*-meQTL associations. In total, we identified approx. 14% more enrichments from about double the number of CpGs available as compared to the 450k array.

Analysis	Test set	SNP	TF
Adding correlated meQTLs	450k + correlated	37 / 0 / 5	108 / 0 / 3
Adding independent meQTLs	450k + independent	36 / 1 / 20	106 / 2 / 12
Adding all EPIC-specific meQTLs	EPIC	37 / 0 / 23	107 / 1 / 15

Table 4.2.: Results from individual comparisons in the EPIC based TFBS enrichment analysis. Assessed are the total number of SNPs and TFs identified through enrichment testing for three different marker sets. Numbers in columns 3 and 4 indicate: overlap / specific to 450k / specific to Test set.

Effect of increased resolution on functional enrichment of *cis*-meQTL

Next, we evaluated the effect of the increased resolution of the EPIC array on the *cis* regulatory element analysis (based on chromHMM state enrichment) of our *cis*-meQTL. Here, we utilized all EPIC *cis*-meQTL ($P < 10^{-14}$) which map to any of the *cis*-meQTL from the European discovery ($R^2 > 0.2$ and distance < 1Mbp for the CpG) and analyzed their overlap in chromHMM promoter/enhancer (same as for 450k analysis, see Section 4.2.3). We generated a EPIC specific set of background pairs (matched for

cis-meQTL dataset	Number of CpGs in an enhancer / promoter state	Observed CpG has an associated SNP in same state	Matched background pair in same state (median)
450k marker set	11,172	9,930 (89%)	65%
EPIC marker set	16,099	14,973 (93%)	82%

Table 4.3.: Table compares the results for the chromHMM enrichment in *cis*-meQTL CpGs that are in an enhancer or promoter state for the EPIC and 450k analysis. Percentages indicate the proportion of pairs sharing the same state with an associated SNP.

CpG methylation mean and SD, compare 450k analysis) and performed 10 iterations of background sampling. The EPIC enrichment results were then compared to the results obtained from the 450k and the matched background pairs. The results of this comparison are displayed in Table 4.3 and show, that the fraction of EPIC pairs residing in the same state is higher as the fraction of 450k pairs for observed meQTL (93% vs 89%, respectively) and the sampled and matched background pairs (82% vs 65%, respectively).

4.4. Project summary

With the work described in this chapter, we reported the first large-scale meta-analysis of methylation QTL (meQTL) with a focus on providing novel insights into the regulatory mechanisms underlying genome-wide *trans*-meQTL effects. To this end, we analyzed the relationships of 9.1 million single nucleotide polymorphisms (SNPs) and DNA methylation at CpG dinucleotides (360,000 sites) in a genome-wide scan using blood samples of 6,994 individuals of European (N=1,731 discovery; N=2,068 replication) and South Asian (N=1,841 discovery; N=1,354 replication) descent. We performed a comprehensive evaluation of the associations between genetic variants and CpG methylation and identified 11,156,559 unique SNP-CpG associations in peripheral blood, including a total of 10,346,172 pairs acting in *cis* (same chromosome, SNP-CpG distance $< 1Mbp$), 351,472 acting in *longrange* (same chromosome, SNP-CpG distance $> 1Mbp$) and 467,915 SNP-CpG associations across genome boundaries (*trans*-meQTL). Our associations comprise a total of 2,709,428 SNPs and 70,709 CpGs which replicate in both ethnic groups ($P < 10^{-14}$) and which we confirmed in independent data from isolated leukocytes, isolated adipocytes and adipose tissue. Moreover, we replicated 96% of the European discovery associations in independent EPIC array data from KORA (N=1,848) indicating good data quality and strong agreement between studies and platforms. Additional conditional analysis on our cosmopolitan set of meQTL identified 34,001 *cis*-acting genetic loci, 467 in *longrange* and 1,847 in *trans*, as independent drivers of underlying regulatory mechanisms. The main association results generated in this project are publicly available at `https://qtldb.helmholtz-muenchen.de`.

For this project, we contributed extensive functional and network analyses for the identified meQTL pairs which shed light on the cellular processes underlying *trans* regulation. By performing elaborate enrichment analyses using transcription factor binding sites (TFBS), chromatin conformation capture, and chromatin state data, we highlighted the functional relevance of meQTL both in local (*cis*) and global (*longrange*, *trans*) contexts. We found that genetic variants associated with DNA methylation are enriched for being located in active chromatin regions ($P = 3 \times 10^{-40}$) and for association with gene expression ($P = 8.1 \times 10^{-18}$ and $P = 2.5 \times 10^{-66}$ for *cis* and *trans* meQTL, respectively). For the *longrange* and *trans* associations we further observed enrichment in Hi-C derived topologically associating domains (TADs, for *longrange*) and intra- and inter-chromosomal chromatin contacts (empirical $P < 6.6 \times 10^{-3}$ for *longrange* and *trans*), indicating the involvement of putative enhancer-promoter interactions to realize genome-wide SNP-CpG associations.

Moreover, we performed a systematic assessment of the regulatory processes underlying the pruned set of 1,847 *trans*-meQTL loci, including *trans* hotspots, which are associated with 3,020 methylation sites and are of particular importance to understand genome regulation. We combined *trans*-meQTL with additional TFBS information of 256 ChIP-seq experiments to corroborate the functional importance of *trans* hotspots including potential master regulators through TFBS enrichment. Additional investigations of EPIC array data, which provide higher resolution and more genome coverage compared to the 450k methylation data, corroborated and extended our enrichment analyses. Here, we recovered transcription factors initially enriched in the 450k *trans*-meQTL associations and identified novel enrichments. Moreover, by including additional protein-protein interaction data and integrating diverse functional data we identified transcription factor pathways and likely candidate genes linking the effect of genetic variants to *trans* methylation and expression for 104 identified *trans*-meQTL hotspots. Candidate transcription factor networks were established for several regulatory proteins, which include *NFKBIE*, *RELA*, *SENP7*, *CTCF*, and *NFKB1*. The identified candidate genes at *trans* hotspots showed enrichment for the encoding of transcription factors and their interacting proteins. We followed up three of our loci with additional analyses including the sentinels *rs6511961*, *rs9859077*, and *rs730775*, and their identified candidate genes (*ZNF333*, *SENP7* and *NFKBIE*, respectively) and established further insights regarding their effects on *trans* methylation and gene expression. Importantly, additional experimental validation via ChIP-seq and IP-MS for the novel *ZNF333* locus which encodes *ZNF333* in *cis* confirmed DNA binding of *ZNF333* at regions overlapping the locus' *trans*-CpG signature, thereby corroborating our results and extending our insights into *trans*-meQTL mechanisms.

In summary, we generated new insights into the regulatory pathways underlying observed statistical associations between genetic variants and DNA methylation. These included comprehensive details on the regulation of nuclear function and molecular phenotypes and therefore advanced our understanding of *trans*-acting and disease-

associated genetic variants. Moreover, we provide a rich database of novel relationships between SNPs and CpG sites, particularly in *trans*, thus providing a scaffold for novel hypothesis-driven experimental studies to unravel complex molecular mechanisms.

5. Prior based network inference

Chapter glossary

multi-omics data	A data set in which for each biological sample at least two different kinds of molecular information (such as genotype, gene expression, or DNA methylation information) is available.
***trans* association**	Association involving traits (e.g. SNP and CpG) on different chromosomes
QTL hotspot	A SNP statistically associated with at least 5 quantitative *trans* traits (such as expression of genes)
meQTL	**me**thylation **Q**uantitative **T**rait **L**ocus - A genetic variant associated with DNA methylation
eQTL	**e**xpression **Q**uantitative **T**rait **L**ocus - A genetic variant associated with gene expression
graph	Consists of a set of nodes/vertices and edges connecting the nodes
GGM	Gaussian graphical model - Graph where nodes reflect normally distributed random variables and edges the conditional dependence of the variables
(edge-wise) prior	Encoding of previously derived knowledge (about interactions) ranging from 0 (not likely, not found previously) to 1 (likely, found previously with high confidence)
MCC	**M**atthews **C**orrelation **C**oefficient - Correlation measure for imbalanced classes such as a relatively small number of edges compared to a large number of 'non-edges' in networks

In the previous chapter, we established genome-wide meQTL and looked at the regulatory networks underlying *trans*-meQTL hotspots by investigating established protein-protein interaction (PPI) and protein-DNA interaction networks in a functional context. Interestingly, we made the observation that interactions derived from available functional data aligned well with the identified networks. In addition, these data could be used to complement our network analyses by providing an additional layer of information by adding eQTL for network genes and eQTM for *trans* associated CpGs. Based on these observations we set out to design a unified approach that uses all available information simultaneously. To this end, we integrate functional multi-omics data with prior information derived from large-scale biological data sources to dissect

the regulatory mechanisms underlying expression and methylation *trans*-QTL hotspots using de-novo regulatory network inference. Contrary to the previous chapter, we utilized the PPI, protein-DNA interaction, and other biological knowledge, which is provided through numerous large databases (see Table 1.2), as prior information to guide the inference rather than to view these as fixed interactions.

The work presented in this chapter is part of a collaborative effort with Prof. Battle from the Johns Hopkins University as well as Prof. Chambers (ICL) and Dr. Gieger and Dr. Waldenberger (AME) and describes and adapts parts of the manuscript submitted for publication in *Genome Medicine* [2]. Our goal was to unravel the complex molecular underpinnings of genomic master regulators by providing a unified approach to integrate multi-omics data through regulatory network inference. The chapter details the derivation of comprehensive prior information from public interaction databases (e.g. BioGRID [188] and ReMap [71]), and large-scale functional data resources (e.g. GTEx [82, 93], ARCHS^4 [119] and the Roadmap Epigenomics Project [48]) and shows how we integrated these priors with human population-scale multi-omics data to explain important trait associated *trans* hotspots.

5.1. Inferring multi-omics networks from functional data

Trans quantitative trait loci (*trans*-QTL) represent genomic master regulators [83] and are particularly interesting for genetics studies as they tend to be enriched for disease-associated variants [22, 23, 26]. However, the mechanisms underlying their genome-wide effects are difficult to explain [21]. Based on the results from the previous chapter we reasoned that a fully integrative approach to network inference for *trans*-QTL hotspots has the potential to yield novel insights into their underlying regulatory processes. A limitation of the random walk based network analysis approach of Chapter 4 is that it is not able to detect unknown edges and purely relies on edges already present in PPI and protein-DNA interaction databases. Additional information, e.g. in the form of eQTM edges, has to be added in a separate correlation analysis. An advantage of de-novo inference of regulatory networks from functional data is that it does not restrict the search space for edges and hence has the potential to give a more complete description of regulatory mechanisms. Moreover, it has been shown that simultaneous integration of omics data can be used to obtain more detailed insights into the investigated systems as compared to using e.g. pairwise correlation approaches [3, 84, 99].

We used multi-omics data from the KORA and LOLIPOP cohorts (see Section 2.2) to infer regulatory networks across three distinct genomic layers: the genotype, gene expression, and DNA methylation layers. The inference of regulatory networks from molecular high-throughput data has been studied intensively [28, 29, 130, 249, 250]. For instance, several works set out to infer networks from single omics data [84, 251] or to individually combine distinct omics layers to infer interactions, such as genotypes and gene expression levels [21, 245, 252, 253] or chromosomal aberration [254] data. Nowa-

days, with more and more multi-omics data being generated, it is possible to reconstruct interaction networks across more than two genomic layers in a fully integrated approach and hence obtain more detailed insights into regulatory patterns [34].

While several efforts have been carried out to reverse engineer networks from such data, methods to successfully construct networks from multi-omics data are still lacking [3, 10, 99, 255]. Another novel aspect of network inference emerged due to the steady growth of genomics data and their availability through large, publicly accessible databases: the established biological knowledge can serve as a-priori information to guide the inference process [99, 128, 256]. As an example, several resources have been made available which provide PPI networks or large-scale eQTL results (interactions between genotypes and genes) and hence provide prior evidence for specific regulatory interactions which can be used to alleviate network reconstruction in novel contexts (see Table 1.2). The inclusion of prior knowledge into network inference has been investigated previously [28, 30–32, 102, 124, 128, 129, 249] and their benefits confirmed. However, studies have either been lacking in the curation of prior knowledge or were not applied to a human context. Inference of molecular interaction networks from human multi-omics data proves a daunting task, but the curation of comprehensive prior information can facilitate inference by prioritization of interactions between and within distinct genomic layers [3, 99].

In this work, we add to previous efforts by devising a novel strategy to infer regulatory networks seeded around *trans*-QTL hotspots from cohort-scale multi-omics data and prior information. The main goal is to systematically derive mechanistic explanations for these hotspots which exhibit coordinated effects across chromosome boundaries through regulatory networks. We tackle the $N << P$ problem, i.e. where the number of variables P (corresponding to nodes in the network) largely outstrips the number of samples N [3] by 1) stringent creation of locus sets (see Section 5.2.1) to reduce the problem dimension P, 2) careful curation of continuous prior information for possible network edges (see Section 5.2.2) and 3) application of suited models employing e.g. regularization procedures (see Methods). Moreover, we benchmarked several state-of-the-art network inference methods for their capability of reconstructing networks, including an extensive simulation study to investigate the effects of sample size and prior noise as well as a cross-cohort replication analysis. To showcase the benefit of our strategy, we provide detailed evaluations of networks inferred for two trait-associated genetic hotspots related to schizophrenia and lean body mass.

5.2. Methods for regulatory network inference on *trans* hotspots

5.2.1. Generation of locus sets

Similar to the network analysis described in Section 4.3.4 the analyses in this chapter focus on *trans*-QTL hotspots, i.e. genetic loci associated with ≥ 5 quantitative traits in *trans* . We curated hotspots based on both genome-wide methylation and expression quantitative trait loci (meQTL and eQTL, respectively) and inferred regulatory networks individually for each curated hotspot thereby reducing the number of variables per inference task significantly. For each meQTL or eQTL hotspot we carefully selected sets of genomic entities including genetic variants (SNPs), DNA methylation sites (CpGs) and genes, and collected their corresponding data which were then supplied to the inference method.

Hotspot extraction

In a first step, we utilized the list of pruned *trans*-meQTL identified in the previous chapter. We removed sentinels with no annotated *cis* genes or where no expression probes where available (Supplementary Table C.1), as these are needed for locus set definition (see below). Remaining independent genetic loci were added to the set of *trans*-meQTL hotspots M_H if the total number of *trans* associated CpG sites is ≥ 5 yielding a set of $|M_H| = 107$ hotspots.

Next, we obtained the *trans*-eQTL as published by the eQTLGen consortium [22] directly from their website at `https://eqtlgen.org/trans-eqtls.html`[1]. As these results have not been pruned for independent genetic loci we performed manual pruning of the list of 59,786 *trans*-eQTL. To this end, we merged all SNPs within a $1Mbp$ genomic window and with $R^2 > 0.2$ in our data to independent genetic loci. Representative sentinel SNPs for each locus were defined by selecting the SNP with 1) the highest minor allele frequency (MAF) and 2) the largest number of *trans* associated genes. Finally, we added all sentinel SNPs with at least 5 *trans* associations to the set of *trans*-eQTL hotspots E_H, yielding a total of $|E_H| = 444$ hotspots.

In addition to the whole-blood based analyses based around the two QTL sets M_H and E_H we also applied our approach in a different tissue context. We took *trans*-eQTL from the current GTEx v8 release (N=163 *trans*-eQTL over all tissues) to obtain tissue specific eQTL hotspots. To this end, we merged SNPs into independent genetic loci similar to what we did for the eQTLGen loci, i.e. combining SNPs with $R^2 > 0.2$ and distance $< 1Mbp$, but keeping all individually associated genes as *trans* genes for the merged locus. A sentinel SNP was selected again by taking the SNP with the highest MAF. This procedure yielded a single *trans*-eQTL hotspot G_H in Skeletal Muscle tissue from GTEx (see Section 5.3.4).

[1] file `2018-09-04-trans-eQTLsFDR-CohortInfoRemoved-BonferroniAdded.txt.gz)`

The set $\mathcal{H}$ was finally defined as the union of all collected hotspots, i.e. $\mathcal{H} = M_H \cup E_H \cup G_H$.

Locus set definition

We set out to construct locus sets for each of the collected hotspots in $\mathcal{H}$ with the goal of including all entities, which can be used in the network analysis to explain the observed QTL effects. An important aspect of the locus set definition is, that we need to overcome the inter-chromosomal gap between *cis* genetic loci and *trans* associated traits.

We executed the following steps in order to generate all entities for a particular locus set S_L for a *trans* -acting locus $L \in H$:

1. Add QTL entities (SNP $\mathcal{S}$ and *trans* genes/CpGs $\mathcal{T} = \{T_1, \ldots, T_q\}$, where q is the number of associated *trans* entities for L) to S_L
2. Add all genes encoded in 1Mbp window of $\mathcal{S}$ as **SNP-Genes** to S_L (set $\mathcal{G}_C$)
3. For hotspots in M_H, add genes in the vicinity of each $\mathcal{T}_i \in \mathcal{T}$ (previous, next and overlapping genes with respect to the location of $\mathcal{T}_i$) as **CpG-Genes** to S_L (set $\mathcal{G}_T$)
4. Add all **TFs**, where a TFBS overlaps the 50bp region around a CpG or the promoter region of a gene over all $\mathcal{T}_i \in \mathcal{T}$ to S_L (set $\mathcal{G}_{TF}$)
5. Add PPI based **shortest path genes** G_{SP} (genes which connect $\mathcal{G}_C$ with $\mathcal{G}_{TF}$) to S_L

To achieve this, we used several annotation data including the ChIP-seq derived transcription factor binding sites (TFBS) provided by *ReMap* [71][2] and *ENCODE* [60, 257][3] and the protein-protein interaction (PPI) information gathered in the *BioGRID*[188][4] (version 3.5.166). These data were subsequently filtered to match our whole-blood context for sets M_H and E_H, i.e. we only selected TFBS which were identified in blood related cell lineages and filtered the *BioGRID* PPI network for genes expressed in whole-blood, i.e. achieving a median reads per kilobase and million (RPKM) of > 0.1 in GTEx v6p whole blood data. For the single Skeletal Muscle hotspot in G_H we filtered *BioGRID* interactions for genes expressed in Skeletal Muscle. We further used muscle tissue gene expression data downloaded from the ARCHS[4] resources (see Section 2.3.3). We employed our ARCHS[4] data loader[5] and obtained Muscle tissue expression data by setting the filter keyword 'Skeletal_Muscle', yielding N=194 samples. We normalized these data using *ComBat* implemented in the R package *sva*, where we provided the dataset series ID as the *batch* parameter. TFBS were not available for a large number of

[2] `http://tagc.univ-mrs.fr/remap/download/All/filPeaks_public.bed.gz`

[3] `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz`

[4] `https://downloads.thebiogrid.org/Download/BioGRID/Release-Archive/BIOGRID-3.5.166/BIOGRID-ORGANISM-3.5.166.tab2.zip`

[5] `https://github.com/jhawe/archs4_loader`

TFs in muscle related cell lines in ReMap and ENCODE. We therefore used *FactorNet* [258], a neural network based method for predicting TFBS from genomic sequence data, to derive TFBS for numerous TFs from DNAse-seq chromatin accessibility data obtained from muscle cell lines. To this end, we trained *FactorNet* on K562 cell line based TFBS available in ReMap which function as the ground truth. As inputs we supplied DNAse-seq data from ENCODE[6] as well as the raw DNA sequence information. All trained models (N=205) were then used on DNAse-seq data obtained from the *LHCN-M2* muscle cell line[7] and with DNA sequence information as input to *FactorNet*, using default parameters to obtain TFBS predictions. We extracted high confidence binding sites by setting a score cutoff of 0.999 and then merged overlapping sites for the same TF. Only regions with $width < W^T_{0.95}$, where $W^T_{0.95}$ is the 95th percent quantile of the widths of all regions obtained for a specific TF T, were kept to permit only relatively narrow peaks typical for TF binding. This resulted in a set of size N=179 TFs for which we had additional binding sites available in muscle derived data, which were then used for locus set construction.

For the definition of G_{SP} (step 5 above) we extracted genes residing on the shortest minimal node weight path between all *trans* traits $\mathcal{T}$ and the SNP-Genes $\mathcal{G}_C$. In detail, we first added the CpGs (in case locus $L \in M_H$) to the constructed PPI network. We then added connections for each $TF \in \mathcal{G}_{TF}$, if the TF overlaps the 50bp window around the CpG ($L \in M_H$) or the promoter region of another gene ($L \in E_H$). Node weights for each node where then calculated based on network propagation as described in the previous chapter for the random walks (see Section 4.2.9 for details). Next, to be able to apply a shortest minimal node weight path algorithm in order to extract nodes with maximal propagation scores PS, we adjusted the weights of nodes to be proportional to $PS^*_i = max_i(PS_i) - PS_i$ for each node i. We then obtained the minimal node-weight paths connecting elements in $\mathcal{T}$ and the SNP-Genes $\mathcal{G}_C$ by applying the *sp.between()* function as implemented in the *RBGL* R library (version 1.56.0, see also [259]) and collected the genes located on the shortest paths and not yet in S_L.

Functional data were subsequently collected for all nodes present in the locus sets and supplied to the respective network inference algorithm (see below) as inputs.

5.2.2. Sources and formulation of biological priors

In this project, we used several public resources to curate a comprehensive set of biological prior information across different omic levels for locus set based network inference. Numerous large-scale data sources have been established, often through the efforts of large international consortia, which we aim to leverage for prior information in network inference. A list of resources collecting functional omics data and molecular interactions is given in Table 1.2. Priors are defined on a per-edge basis and we defined priors for four distinct types of edges:

[6] dataset ENCFF971AHO
[7] ENCODE dataset *ENCFF639MPM*

1. **SNP-Gene**: Edges between S and collected SNP-Genes (set $\mathcal{G}_C$)
2. **Gene-Gene**: Edges between all genes in S_L (except TF-target edges)
3. **CpG-Gene**: Edges between CpG sites $\mathcal{T}$ and genes encoded in their vicinity (set $\mathcal{G}_T$)
4. **TF-target**: Edges between TFs and their (ChIP-seq based) targets

SNP-Gene priors (item 1) are generated from previously published eQTL results, and Gene-gene priors (item 2) by combining PPI information from BioGRID with tissue matched gene expression data from GTEx (whole-blood context) and ARCHS[4] (Skeletal Muscle context). For the CpG-Gene priors (item 3), we used the 15-state chromHMM model from Roadmap [48, 49]. For the **TF-target** priors (item 4), we set a fixed, large prior value of 0.99 for all possible interactions between TFs and their respective target gene or CpG. The reasoning behind this is, that ChIP-seq peaks represent strong evidence of protein-DNA interactions and hence should receive high 'a priori' interaction probabilities. We defined targets for a TF as either 1) genes, where the region 2,000bp upstream and 1,000 downstream of the TSS overlap a TFBS or 2) CpG sites, which position $\pm$50bp overlaps a TFBS.

In the following, we describe in detail how we constructed priors for the individual edge types 1)-3) defined above.

Priors between SNPs and SNP-Genes

One important component of our multi-omics network inference is the detection of links between the genetic variants and *cis* genes in $\mathcal{G}_C$, to identify the gene most likely mediating the observed *trans* effects.

We used publicly available whole-blood eQTL results from GTEx v6p[8], i.e. SNP-gene interaction results, to define priors for the SNP and SNP-Gene combinations (between sets S and $\mathcal{G}_C$). After obtaining the complete result table containing the results of all association tests we calculated for each S-$\mathcal{G}_C$ pair the local false discovery rate $lFDR$, which reflects the Bayesian posterior probability of a true null hypothesis given a test statistic [260, 261]. For this, we used the *fdrtool* R library, version 1.2.15. Based on the $lFDR$ we then defined a prior for a specific S-$\mathcal{G}_C$ pair A and B as $p_{\mathcal{AB}} = 1 - lFDR^{GTEx}_{\mathcal{AB}}$.

For the application of our approach to GTEx Skeletal Muscle tissue we cannot use the same genotype and gene expression data on which we perform the inference as prior information. Instead, we obtained independently generated Skeletal Muscle based eQTL published by Scott, Erdos, Huyghe, et al. [262][9] and used the $lFDR$ method to generate SNP-Gene priors for this context.

[8] file Whole_Blood_Analysis.v6p.all_snpgene_pairs.txt.gz from `https://www.gtexportal.org/home/datasets`

[9] obtained from `https://theparkerlab.med.umich.edu/data/papers/doi/10.1038/ncomms11764/`

Definition of gene-gene priors

For the intermediate component of the putative regulator networks, i.e. the gene-gene interactions connecting the genetic locus with the transcription factors and *trans* entities, gene to gene edge-priors can help clarify the regulatory sequence mediating the associations. To define the gene-gene edge-priors we utilized GTEx v6p whole-blood gene expression data in conjunction with the PPI obtained from BioGRID. Summary gene expression data were downloaded directly from the GTEx portal[10], filtered for samples with RNA integrity number (RIN) ≥ 6 (high quality samples), log2 transformed, quantile normalized and subsequently moved to standard normal distribution. We removed the first 10 principal components from these data in order to correct for confounding factors [263]. To obtain a high quality set of priors we only set priors for gene-gene edges which also show up in the PPI network. For each such pair, we then correlated their respective gene expression and gathered the association P-values. We calculated $lFDR$ on the set of generated P-values and set the prior for an edge between genes $\mathcal{G}_1$ and $\mathcal{G}_2$ to $p_{\mathcal{G}_1\mathcal{G}_2} = 1 - lFDR^{GTEx}_{\mathcal{G}_1\mathcal{G}_2}$, similarly as we did for the SNP-Gene priors.

Priors between CpGs and their gene neighbors

Edges between CpGs (set $\mathcal{T}$) and the genes encoded in their vicinity (set $\mathcal{G}_T$) are specific to hotspots obtained from the meQTL results (locus sets in M_H), To formulate priors for these CpG-Gene edges we make use of the genome-wide ChromHMM states [49], specifically the 15 states model, as reported in light of the Roadmap Epigenomics project [48] (see also Section 2.3). State definitions, segmented into 200bp windows across the genome were directly obtained from the Roadmap web portal[11]. The chromHMM states have been derived from multiple histone ChIP-seq experiments on the same cell types and represent functional annotations of the genome, classifying segments for instance in enhancer or promoter related states. We formulated priors for CpG-Gene edges for which the genomic distance of the involved gene G and the CpG C is no larger than 200bp. For each such pair, we quantified the prior probability of the CpG influencing the expression of the gene by calculating the fraction of Roadmap blood-related cell-lines for which the CpG resides in an 'active transcription' state including 'TssA', 'TssAFlnk', 'TssBiv' and 'BivFlnk' states (see Table 1.1 for detailed state information). This fraction, p_{Tx}, is finally adjusted for white blood cell proportions by weighting in the available Houseman white cell subset estimates. For this, we multiplied the chromHMM state proportions with the population mean for each of the Houseman cell estimates yielding P_{Tx}. A specific CpG-Gene prior for a CpG $\mathcal{C}$ and a gene $\mathcal{G}$ is then set to $p_{\mathcal{C}\mathcal{G}} = P_{Tx}$, if the genomic distance $d(\mathcal{C}, \mathcal{G}) <= 200bp$.

Priors can typically not be created for all possible edges between the entities in the locus sets indicating a certain lack of evidence for the specific edges. For instance, gene-gene priors might be missing because a PPI between these genes has never been

[10] `https://www.gtexportal.org/home/datasets`

[11] `https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html`

reported. We hence set a low pseudo-prior for each such edge as $p_{pseudo} = 1 \times 10^{-7}$.

5.2.3. Simulation study design and replication analysis

Many methods have been proposed to perform regulatory network inference from (multi-)omics data. However, depending on the specific context (e.g. $P >> N$, data modalities, use of priors) specific methods might be better suited for particular inference tasks than others. We performed simulation and replication analyses to identify the method best suited for our context, i.e. network inference from QTL hotspots (relatively low P) on multi-omics data under the inclusion of prior information. In the simulation study, methods were tested independently on simulated data and evaluated for 1) their ability to infer simulated ground truth networks, 2) the impact of low versus high sample sizes on their performance and 3) their sensibility to noise in the provided priors.

For the replication analysis, we do not have a ground truth network available but seek to investigate how stable networks inferred from the same method and hotspot are across different datasets. To this end, we infer networks independently on both datasets and cross-compare resulting networks, i.e. comparing networks inferred in LOLIPOP against the ones inferred in KORA and vice versa.

For both the simulation and replication study, we compared the reconstructed networks in terms of Matthews Correlation Coefficient which represents a balanced correlation measure suited for our imbalanced class labels (i.e. few edges vs many non-edges) [124, 264]. Specifically, Matthews Correlation Coefficient only results in a good (high) score if all four basic criteria including false positives, false negatives, true positives, and true negatives yield good results [265]. It is defined as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{5.1}$$

Simulation study

For the simulation study, ground truth graphs and data were simulated 100 times for each of the meQTL hotspots in set M_H. For each locus set $\mathcal{L}_S$, we first generated a prior matrix $\mathcal{P}_S$ as described above. Then, a possible hotspot graph $\mathcal{G}_T$ is sampled considering all entities available in $\mathcal{S}_L$ by sampling edges uniformly from the prior matrix $\mathcal{P}_S$. Therefore, a specific edge e_{ij} is present in $\mathcal{G}_T$ for a prior entry $p_{ij} = p_{ji}$ of $\mathcal{P}_S$ (symmetric matrix), only if $p_{ij} > p_{pseudo}$ and if in addition $runif(0,1) <= p_{ij}$, where $runif(0,1)$ samples random values from a uniform distribution between [0,1].

This sampled graph forms the basis for one simulation iteration (out of 100). For each generated $\mathcal{G}_T$ we constructed 10 noisy ($\mathcal{G}_N$) ground truth graphs $\mathcal{G}_N^{10}, \mathcal{G}_N^{20} \ldots \mathcal{G}_N^{100}$ by rewiring the graph edges such that $10, 20, \ldots, 100$ percent of the prior information is still true and the degree distribution of each $\mathcal{G}_N$ stays the same as in $\mathcal{G}_T$. For example, rewiring 10% of the edges in $\mathcal{G}_T$ generates $\mathcal{G}_N^{10}$ and by making sure the introduced edges receive no priors according to $\mathcal{P}_S$ this effectively generates 10% of noise in $\mathcal{P}_S$. Lastly, we

included one additional comparison to evaluate prior knowledge about the density of the observed graph. To this end, we estimated a single prior probability representing for all edges based on a binomial model for edge probabilities. We utilized the total number $|E_{\mathcal{G}_T}|$ of edges over all graphs in $\mathcal{G}_T$ within a single simulation run. The number of possible edges is given by $|E_T| = (N * (N-1))/2$, where N is the total number vertices. The binomial prior is then defined as

$$p_{rbinom} = max(\frac{1}{N_S} * \frac{\sum_{\mathcal{G}_T} |E_{\mathcal{G}_T}|}{|E_T|}, p_{pseudo}),$$

where N_S represents the number of sampled (randomized) graphs in this simulation run.

Data were then generated according to the structure of each graph $\mathcal{G}_S \in \{\mathcal{G}_T, \mathcal{G}_N^i; i \in \{10, 20, \ldots, 100\}$. For this, we made use of the *bdgraph.sim()* method of the *BDgraph* R package and supplied the following parameters:

1. **p**: $|\mathcal{S}_L|$ (number of nodes)
2. **graph**: $\mathcal{G}_S$ (graph structure as adjacency matrix)
3. **N**: 612 (sample size in LOLIPOP)
4. **mean**: 0 (mean vector for sampling from the multivariate normal)

The method then generates a data matrix of the specified size, i.e. $N \times p$, sampled from a multivariate normal distribution and adhering to the covariance structure as defined by the structure of the supplied graph.

One remaining issue is to simulate the discrete genotypes in these data to assess their impact on network inference in the simulation study. To achieve this, we created discrete genotype dosage information (values 0, 1 or 2) based on the normally distributed SNP variable in the simulated data set, which reflects the observed allele frequencies of the hotspot genetic variant in the LOLIPOP data. We took the Gaussian data and transformed them to discrete values based on the individual dosage frequencies which we took as quantile cut points for the discretization.

Finally, we simulated data and reconstructed networks for each of the hotspots individually which were then compared to the respective ground truth networks $\mathcal{G}_T, \mathcal{G}_N^{10}, \ldots, \mathcal{G}_N^{100}$.

5.2.4. Estimation of transcription factor activities

Proteins are the final product of many genes encoded on the DNA. While it is possible to measure protein expression directly (e.g. using Mass Spectrometry (MS) experiments or protein arrays) measuring gene expression is typically preferred as it is more comprehensive and cost-effective than MS. However, transcript expression of genes represents a mere proxy for the activity of protein-coding genes and methods have been developed to estimate e.g. the actual activity profile of transcription factors [266, 267]. For

this study, we investigated both gene expression and estimated transcription factor activities (TFA) which aim to represent the true activity of TFs in the data. TFA were calculated for all TFs extracted from the ReMap and ENCODE databases using the *plsgenomics* R package's $TFA.estimate()$ method (version 1.5-2) [266]. This approach uses the normalized expression data of the TF and its target genes (as defined via the binding profile obtained from our curated TFBS) to estimate TFA using partial least squares regression. For each cohort, we supplied as input the full gene expression from KORA and LOLIPOP, respectively, and the incidence matrix I_{TFBS} containing the binding profiles for all available TFs across all measured genes. An entry $a_{ij}^{TFBS} = a_{ji}^{TFBS} \in I_{TFBS}$ is set to 1, if TF i has a binding site within the promoter (2,000 bp upstream and 1,000 bp downstream) of the respective target gene j and is otherwise set to 0.

5.2.5. Graphical model based network inference

In this section, we recap the basics of network inference and describe the methods employed to derive hotspot specific networks in our study. We utilized graphical models to infer the structures of graphs (or networks) to understand regulatory mechanisms in cellular contexts. Graphs consists of a set of P nodes (or vertices) $V = \{v_1, v_2, \dots v_P\}$ and edges $E \in V \times V$ between these nodes. In a graphical model nodes are random variables (RV) and edges represent the conditional dependence between these variables. Alternatively, the absence of edges indicates conditional independence (compare Section 3.1.1). For instance, two variables A and B are conditionally independent given the rest of the variables in the graph (A $\perp\!\!\!\perp$ B $\mid$ rest) if there is no edge between A and B in the graph, i.e. $e_{AB} \notin E$. For inferring network structures, i.e. all individual edges in a graph, diverse methodologies have been proposed. We first introduce pairwise correlations and then move on to conditional independence graphs (graphical models) which are often employed in this context and which we applied in this project. There, we will discuss the basic graphical lasso (gLASSO), tree-based approaches such as GENIE3 and finally a fully Bayesian treatment of graphical models (BDgraph).

Pairwise associations

A straight forward technique to reconstruct networks from omics data is by application of pairwise association measures involving for instance the calculation of Pearson's Correlation Coefficient (PCC) or Spearman's Rank Correlation Coefficient between all measured pairs of RVs in the data individually. Correlations between individual RVs significant according to certain criteria, e.g. $PCC > 0.8$ and $PCC_{pvalue} < 0.05$, then form the edges in the network. As an example, take two genes A and B and their expression measured over multiple samples. Calculating PCC between the measurements of these two genes gives information about co-expression, i.e. the gene A is expressed whenever gene B is expressed or vice versa. Similarly, gene A could be repressed while gene B is expressed (anti-correlation). Although alternatives to correlation based approaches have been proposed, for instance based on mutual information (MI) [268, 269] to detect

non-linear relationships, improved correlation based methods such as the biweight midcorrelation [270, 271] outperformed MI for network module identification [272].

Partial correlations

One issue of using straight forward, pairwise correlation approaches on genomic datasets is that these measures cannot distinguish between direct and indirect effects and edges manifest in the network in both cases as direct and indirect associations [273, 274]. Typically, this issue yields networks of very high density, i.e. a large number of edges [84] which makes the network less accessible and interpretable. This issue has been approached by using partial correlations. The idea is to associate two Gaussian variables but accounting for the effect of all other variables while doing so thereby alleviating the problem of indirect associations. Briefly, this can be achieved by regressing out the effect of all remaining variables prior to testing two variables of interest. For instance, the indirect dependencies between two RVs B and C which originally arose from direct dependency on a mutual third variable, CA, will then no longer manifest in the inferred network as the influence of A has been accounted for and hence only direct dependencies (partial correlations or conditional dependencies) are retained. Consider again the expression of the two genes (B and C) but also include another gene A, such as a transcription factor that regulates both genes. Regulation by the TF A represents a direct association between the expression of A and the expression of the regulated genes B and C. This dependence on a mutual source of both target genes can introduce a pairwise correlation and by using conditional dependencies this spurious correlation would be removed (if the influence of A is the sole determinant of their association in this example). Importantly, in this case, the direct relationship between A and B as well as A and C would be preserved. This concept is illustrated in Figure 5.1.

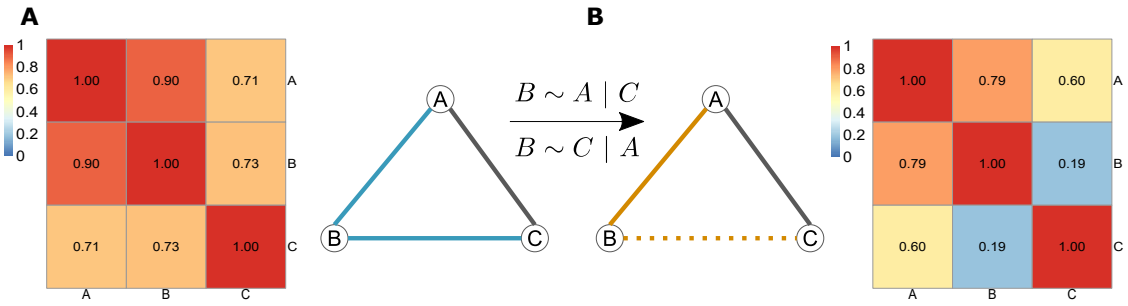


Figure 5.1.: Illustration of the concept of partial correlations. A) When performing pairwise associations, spurious associations between two variables (B and C) can emerge, although in truth this is due to a third variable (A). B) Spurious correlations vanish when considering influence of other variables in a partial correlation based approach. Here, correlation between B an C vanishes when the effect of A is considered. Figure adapted from Hawe, F. J. Theis, and Heinig [3].

Graphical models

The idea of partial correlations forms the basis for *graphical models* which are also known as conditional dependence or partial correlation networks and where edges are only included in case of a conditional dependence between the involved RVs [86, 87, 275]. For this project, we explored different methods based on Gaussian Graphical Models (GGMs) which make the assumption of normally distributed RVs and which have successfully been applied to gene expression data [273], metabolomics data [84] and to infer links between genetic variation and the metabolome [276]. A GGM can be written as multivariate normal distribution in the following way:

$$\mathcal{M}_G \sim \{\mathcal{N}_P(\mu, \Sigma) | \Sigma^{-1} \in \mathcal{P}_G\} \tag{5.2}$$

In Equation 5.2 $\mathcal{M}_G$ is the graphical model with respect to a graph G, $\mathcal{N}_p(\mu, \Sigma)$ is a P-dimensional multivariate normal distribution with mean μ and covariance matrix Σ, Σ^{-1} is the inverse of the covariance matrix also termed the *precision matrix*, and $\mathcal{P}_G$ is the set of all positive semi-definite matrices of size $P \times P$. The conditional independence structure of a graphical model is given via the *precision matrix* Σ^{-1}. An entry p_{ij} of Σ^{-1} is 0, if and only if the RVs i and j are conditionally independent given all other nodes, i.e. $i \perp\!\!\!\perp j | N \setminus i, j$, and > 0 if they are conditionally dependent. For Gaussian Graphical Models the entries in the precision matrix directly reflect the partial correlation values. According to the definition of a graphical model the structure of the network for P nodes (RVs) is directly given via the off-diagonal entries of the precision matrix, i.e. each p_{ij} of $\mathcal{P}_G$ with $p_{ij} > 0$ *and* $i \neq j$ corresponds to an edge in the network between nodes i and j [87]. Network inference methods based on graphical models typically either seek to estimate the full Σ^{-1} [273] or only its non-zero entries [86]. One issue inference methods face with genomic data, aside from spurious correlations, is the $N << P$ problem, i.e. the number of samples N is significantly smaller compared to the number of observed variables P. For instance, a typical experiment could involve several hundred samples (N) and $> 20,000$ genes (P)[12]. To be specific, if $N << P$ more variables than data points are present in the data which is statistically challenging: A model fit will involve many degrees of freedom and the mathematical formulation will be underdetermined, therefore the system is prone to be overfit to the data [195].

GeneNet and the graphical LASSO

One approach of handling the dimensionality burden is via the application of regularization procedures as it is for example implemented in GeneNet [273, 274]. GeneNet uses a Bayesian shrinkage method to obtain a stable estimate of the full Σ^{-1} matrix. Here, the authors obtain the inverse of the correlation matrix using singular value decomposition and generate stable estimates of Σ^{-1} using a bootstrap aggregation (bagging) approach [277], i.e. by repeatedly sampling from the available data and then aggregating the

[12] An exception are single-cell gene expression experiments measuring thousands of cells and the same number of genes

obtained estimates [273]. Moreover, Schäfer and Strimmer [273] derive P-values to detect significant partial correlations by estimating a mixture distribution with two components reflecting 1) the partial correlations of truly dependent variables and 2) partial correlations for independent variables (which should vanish). As there is a larger number of potential edges ($|E| = \frac{|N|*(|N|-1)}{2}$), this also poses a multiple testing problem for which the authors apply false discovery rate (FDR) correction [203] (see also Section 3.1.6).

In our case, we used the GeneNet R package implemented by Schäfer and Strimmer [273] to obtain estimates of the precision matrix. Data are filtered for any missing values and then supplied to the *ggm.estimate.pcor()* method of the package followed by application of the *network.test.edges()* and *extract.network()* methods. We further set a FDR cutoff of $FDR < 0.2$ to obtain regulatory networks from the partial correlation estimates. Although GeneNet cannot make use of pre-existing prior information it has been one of the standards in regulatory network inference and we, therefore, use it as a baseline comparison for prior based methods.

Another approach employing regularization for graphical model inference was proposed by Meinshausen and Bühlmann [86]. The authors proposed a strategy for estimating the non-zero elements of the precision matrix by applying LASSO (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator) regression for each variable individually, using all remaining variables as predictors. This method assumes a sparse precision matrix Σ^{-1} and makes use of L_1 regularization (see Section 3.1.3) to restrict the total number of non-zero estimated parameters $\beta = \{\beta_1, \beta_2, \ldots \beta_p\}$ (i.e. the variable coefficients of the regression analyses, see Section 3.1.2). Variable selection is performed implicitly during the regression by pushing the least important β coefficients to 0 which alleviates the $N << P$ problem. The method can be used to generate a Σ^{-1} where an element σ_{ij} for a pair of RVs i and j is non-zero in case either β_{ij} (i is dependent, j the independent variable), β_{ji} or both parameters are non-zero. Therefore, weaker and potentially spurious dependencies are being discarded from the network. This approach approximates the likelihood of the multivariate normal distribution underlying the graphical model, which for a fixed mean vector ¯ is given as:

$$ll(\Theta) = log(det(\Theta)) - trace(\mathcal{S}\Theta) \tag{5.3}$$

Here, Θ is the precision matrix to be estimated and $\mathcal{S}$ is the sample covariance matrix [195]. Friedman et al. proposed as an alternative the graphical LASSO (*gLASSO*), which directly evaluates the L_1 regularized log-likelihood given as

$$ll(\Theta) = log(det(\Theta)) - trace(\mathcal{S}\Theta) - \lambda||\Theta||_1 \tag{5.4}$$

employing a block-wise gradient descent algorithm [87, 278]. Here, $||\Theta||_1$ is the L_1 norm, the sum of all absolute values of the elements of Θ.

The authors implemented the *gLASSO* in their R/CRAN package *glasso* which we used for inferring networks in our project. Specifically, we execute the *glasso* method of

the package and specify as parameter *penalize.diagonal = FALSE*. For the application of the *gLASSO* an important consideration is how to optimally set the L_1 regularization parameter λ and typically a range of λ is screened and resulting models evaluated using e.g. Bayesian Information Criterion and cross validation [102] (see also Methods). High values of λ imply strong penalization of edges and result in sparser graphs, whereas low λ values imply weak penalization and yield denser graphs. The penalization term can also be provided for each possible edge individually by providing a matrix Λ of size $P \times P$ (with P the number of variables) where each element $\Lambda_{ij} \in \Lambda$ specifies the regularization constant for dependencies between nodes i and j. This directly enables the use of prior knowledge to guide the inference, specifically as the L_1 regularization can also be viewed as a Laplace prior on the respective beta values and can provide edge-specific information (see also Section 3.1.3) [195]. Additionally, the 'weight' the prior information encoded in Λ receives can be screened, similarly to the original λ screening, for instance by element-wise multiplication of Λ with a weight λ, i.e. $\Lambda = \Lambda \times \lambda$ (compare also [31].

In our case, we performed two parallel inference tasks, one using and one not using prior information. For the non-prior case we applied the *gLASSO* for a range of penalization parameters $\lambda \in \{0.01, 0.015, ..., 1\}$, and selected the best model over five cross validation runs (80% training data each) with the minimal mean BIC value obtained from the test data. For the prior case we performed the same procedure but included the prior matrix $\mathcal{P}$ by setting $\Lambda = (1 - \mathcal{P}) \times \lambda$ for each $\lambda \in \{0.01, 0.015, ..., 1\}$, which is similar to what has been proposed in Z. Wang, Xu, Lucas, and Y. Liu [30] and Y. Li and Jackson [31].

name	**version**	**repository**	**attribute**	**reference**
BDgraph	2.61	CRAN	Bayesian/ MCMC	[98, 124]
gLASSO	1.11	CRAN	Graphical lasso	[87]
GENIE3	1.2.1	bioconductor	Random forests	[88]
GeneNet	1.2.13	CRAN	Shrinkage/ FDR	[273, 274]
iRafNet *	1.1-2	CRAN	Random forests	[32]

Table 5.1.: The inference methods and their respective implementation used in this project. Table adapted from Hawe, Saha, Waldenberger, et al. [2].
* *adjusted to make use of parallel processing*

Tree-based inference methods

The Gaussianity assumption on all RVs for inferring homogeneous networks might not hold in settings involving heterogeneous multi-omics data, for instance when integrating discrete genotypes with continuous gene expression measurements. Tree-based methods such as *GENIE3* [88] or *iRafNet* [32] form an interesting alternative to the graphical LASSO and are particularly suited for multi-omics settings as they are free of any distributional assumptions and can detect non-linear relationships. They hence have

the potential to successfully integrate mixed data types (e.g. discrete and Gaussian, Gaussian and non-Gaussian, etc.) across multiple omics layers.

The idea of tree-based methods is similar to the idea underlying *gLASSO*: For each variable A in the data, a random forest model [279, 280] including remaining variables $B \in \mathcal{P}$ as independent variables is built and interactions between A and all $B \in \mathcal{P}$ inferred based on their importance for explaining A (i.e. variables are ranked according to their influence, similar to e.g. [86]). Link rankings are then merged from the P distinct models to obtain a single list of rankings of edges between individual entities. The rankings do not equate to any statistical measure [96] and therefore the optimal number of edges (ranked by their variable importance) needs to be determined [88, 96] (similar to the screening for the best regularization parameter λ in the *gLASSO*). This can for instance be done by employing Stability Selection as proposed by Meinshausen and Bühlmann [281] to control false positive findings [96]. Alternatively, the top N edges can be selected such that the resulting network fits an expected network topology (e.g. scale-free topology for biological networks).

We make use of both *GENIE3* [88] and *iRafnet* [32] in our project to infer networks. Although *GENIE3* cannot use prior information, we nevertheless included it in our project as it was one of the top performers of recent network inference challenges (DREAM4/5, [89, 90], see also Section 1.4.3). When using *GENIE3* for the network inference we first removed missing entries from our data matrix (see also application of *GeneNet* above) and variance normalized the input data as suggested in [88]. We then apply the *GENIE3()* method of the GENIE3 R/bioconductor package to generate the basic model and utilized the *getLinkList()* method to obtain the ranked list of edges. In both cases, we used default parameters. To define a weight cutoff for the ranked list of edges we initially divide it into 200 quantiles, i.e. defining 200 possible cutoffs for the list, in case the total number of obtained distinct weights is larger than 200. For each such cutoff $c \in C$ we then created the corresponding regulatory network N_c for links L with a weight $w_L > c$ and used a scale-free topology as a reference for network evaluation. Specifically, we followed the approach proposed by B. Zhang and Horvath [270] to evaluate each N_c. We divide the degree distribution d containing the number of edges for each node for N_c into 20 distinct bins and then formulate a linear model $log10(d_p)\ log10(d)$, where d is the mean degree for each bin and d_p is the frequency of degrees in that bin. From this model, we extract the R^2 and the fitted coefficient β. We only consider models for which $\beta < 0$ (negative slope), as for a scale-free network the frequency of nodes with high degree should be lower than the frequency of nodes with low degree. The generated R^2 values indicate how well a network relates to a scale-free topology ('goodness-of-fit'). The authors suggest keeping only networks with R^2 values above 0.8 to obtain scale-free networks, however, in our case if none of the N_c for all cutoffs $c \in C$ has an $R^2 > 0.8$ we keep the network with the highest R^2 over all networks. In case there are multiple networks with the same R^2 we constrain further by the mean connectivity of the network.

In contrast to *GENIE3* the tree-based *iRafNet* is specifically designed to include prior

information during the inference process [32]. Here, Petralia et al. build upon the concept employed by *GENIE3* but use weights (priors) to prioritize specific entities when constructing individual decision trees for a selected variable. This increases the probability of important variables to be chosen early on in constructing the trees [32]. For our study, we utilize the implementation provided by Petralia and colleagues. As it is not possible to execute *iRafNet* without prior information we cannot directly compare a prior-based to a non-prior run and hence only infer networks including prior information. We use the data filtered for missing values with the *iRafNet()* method and apply parameters *ntress = 1000*, *mtry = round(sqrt(ncol(data)-1))*, and *npermut = 5*, followed by the *Run_permutation()* method supplied with the same parameters. Internally, *iRafNet* performs a permutation procedure to generate empirical P-values. We select the network based on $FDR < 0.05$ using the *iRafNet_network()* method and setting the parameter *TH = 0.05*. To enable parallel computation, we modified the original (no longer maintained) version of the *iRafNet* package and made it available under `https://github.com/jhawe/irafnet_custom`.

Bayesian treatment of network inference

Mohammadi and Wit proposed a fully Bayesian approach to GGM estimation [124] and provided an extension for application on mixed data [98]. Here, non-Gaussian variables are not modelled explicitly, but are rather transformed to a Gaussian distribution using the semi-parametric copula modelling approach proposed in Dobra and Lenkoski [282].

In the approach by Mohammadi and Wit the authors perform efficient Markov-Chain-Monte-Carlo (MCMC) sampling to sample graph structures and propose an efficient sampling scheme to obtain samples from the distribution of precision matrices $\mathcal{P}_G$. This is a necessary step as the Monte Carlo sampling requires frequent sampling to cover a large fraction of possible graph structures $\mathcal{G}$ which is exponentially large in the number of nodes, i.e. $|\mathcal{G}| = 2^{\frac{P*P(-1)}{2}}$ (with P the number of nodes/variables). They formulate a posterior probability for a graph G given a specific instance of a precision matrix K and data D as

$$P(G, K|D) = P(D|G, K)P(K|G)P(G) \tag{5.5}$$

Therefore, their MCMC method can make use of prior information, particularly in the form of edge-wise priors. To explore the graph space they propose a birth-death approach where in each iteration of the MCMC algorithm an edge is either added (birth) or removed (death) from the current graph structure, depending on birth/death rates of the individual edges which themselves follow a Poisson process. For instance, the probability for the birth of an edge e at a specific state determined by (G, K) where G is the current graph structure and K is the precision matrix, is given as

$$P(birth\ of\ e) = \frac{\beta_e(K)}{\beta(K) + \delta(K)}, \tag{5.6}$$

where $\beta_e(K)$ is the rate for the Poisson process of birthing edge e and $\beta(K) = \sum_{e \in \overline{E}} \beta_e(K)$ and $\delta(K) = \sum_{e \in E} \delta_e(K)$ are the overall rates for all edges $e \in E$ and all non-edges $e \in \overline{E}$, respectively, according to the current precision matrix K [124]. The birth and death rates are designed such that the algorithm converges to the desired posterior distribution for a large enough number of iterations [124].

The bulk of the computation in this approach arises in computing the individual $\beta_e(K)$'s and $\delta_e(K)$'s and obtaining new samples from the precision matrix K. For example, computation of $\beta_e(K)$ and $\delta_e(K)$ involves calculation of the normalizing constants for the prior distribution of the precision matrix and a new sample of K. For sampling K Mohammadi and Wit propose a new approach for obtaining direct samples from a G-Wishart distribution [124, 195]. Nevertheless, the algorithm is computationally more expensive than any of the other methods with the exception of *iRafNet*. Finally, by iteratively adding and removing edges to the network structure according to their specific birth and death rates, the Markov-Chain is designed to ultimately converge to the desired posterior distribution [124] and the graph with the highest posterior probability can be extracted.

When applying *BDgraph* to our data we inferred networks both under consideration of prior information and without prior information supplying only a uniform non-informative prior for the latter case. We utilized the implementation of the *BDgraph* R package, specifically the *bdgraph()* method. We set parameters *method = "gcgm"*, *iter = 10000*, *burnin = 5000*, i.e. we perform a relatively large number of iterations to achieve stable results (good convergence to the desired posterior distribution). The *g.prior* parameter was set according to the gathered prior information matrix for each hotspot. In addition, we specified the *g.start* parameter which takes an incidence matrix for the graph structure G_S at which to start the sampling. To determine G_S we generated an incidence matrix $I_G S$ for which we set each entry k_{ij} to 1, if the corresponding value p_{ij} in the prior matrix is > 0.5 and to 0 otherwise. After the model fit we extracted the graph structure using the package's *select()* method with parameter *cut = 0.9* which controls the posterior probability of edges to be included in the graph.

5.2.6. Network prioritization and final network creation

We inferred networks for a total of 107 meQTL and 444 eQTLGen hotspots which yielded networks with a median number of 67 and 20 edges for $gLASSO_P$ and 72 and 27 for $BDgraph_P$, respectively. In order to select the most interesting networks we filtered and ranked networks based on the following criteria.

GWAS filtering. For detailed evaluation we only considered genetic loci which have previously been linked to a complex trait in a GWAS and used the current version of the GWAS catalog [283] (v1.0.2) to annotate genetic loci accordingly. In addition, we extracted proxy SNPs in high LD ($R^2 > 0.8$) with our sentinel SNPs using the SNiPA tool [284] and tested them for GWAS hits. If either the original SNP or one of its proxy SNPs had a GWAS hit the respective network was included in downstream analysis.

Network ranking. To rank remaining networks for further investigation, we utilized a simple graph score reflecting desirable biological properties which could be assumed for *trans*-QTL derived networks. It is designed such that we 1) rate the adjacency of SNPs and SNP-genes positively, 2) rate the inclusion of *trans* entities positively in case they are not directly linked to the genetic locus and 3) rate high graph densities negatively (i.e. more parsimonious graphs yield higher scores than dense ones). We defined our graph score as:

$$S_G = -log10(D_G) * [\frac{1}{|\mathcal{G}_C|}(\sum_{i=1}^{|G_S|} 1 - \sum_{i=1}^{|\overline{G_S}|} 1) + \frac{1}{|\mathcal{T}|}(\sum_{i=1}^{|G_T|} 1 - \sum_{i=1}^{|\overline{G_T}|} 1)]$$

Here, D_G is the density of the graph, $\mathcal{G}_C$ are all SNP-Genes, $\mathcal{T}$ is the set of all *trans* entities, G_S are all SNP-genes adjacent to the SNP in G or linked directly to a different SNP-Gene, $\overline{G_S}$ are all SNP-Genes in G not directly linked to the SNP or another SNP-Gene, G_T are the *trans* entities in G linked via an arbitrary path to any SNP-Gene which does not include the SNP or another *trans* gene and $\overline{G_T}$ are all *trans* genes linked directly to the genetic locus. Only the cluster containing the SNP, i.e. the SNP itself and any nodes reachable from the SNP via any path in G, is considered for calculating S_G. If the SNP is not present or no SNP gene has been selected in the final graph the score is set to 0. We further ranked networks based on straight forward network statistics such as the total number of edges and nodes in order to prioritize smaller networks for detailed analysis.

Graph merging. Finally, based on the individual networks obtained from the two cohorts for a single hotspot we constructed a final graph containing only high confidence edges. We created the combined graph by including only those edges and nodes which are present in both cohort networks. Any nodes not connected to the network were subsequently removed.

5.2.7. Colocalization analysis to corroborate networks

We corroborated our inferred networks by performing a formal colocalization analysis of GWAS and *trans*-eQTL signals observed at the schizophrenia locus. For this, we obtained GWAS summary statistics for schizophrenia found using the GWAS atlas [285][13]. *Trans*-eQTL signals in whole-blood were downloaded from eQTLGen for all SNP-Gene pairs[14]. We then employed *fastENLOC*[15] [237, 238] to calculate SNP-level colocalization probabilities following the guidelines published in the *fastENLOC* Github README and using default options. Probabilistic eQTL annotations were generated

[13] `https://atlas.ctglab.nl/` which we downloaded from `http://walters.psycm.cf.ac.uk/clozuk_pgc2.meta.sumstats.txt.gz`

[14] `https://www.eqtlgen.org/trans-eqtls.html`, file 'Full trans-eQTL summary statistics'

[15] `https://github.com/xqwen/fastenloc`

using *DAP-G* [286, 287][16] and the required PIP files constructed using *TORUS* [288][17]. Finally, the required LD block definitions were obtained from LDetect [289][18].

[16] https://github.com/xqwen/dap/
[17] https://github.com/xqwen/torus
[18] https://bitbucket.org/nygcresearch/ldetect-data/src/master/

5.3. Multi-omics integration for *trans* hotspot regulatory network inference

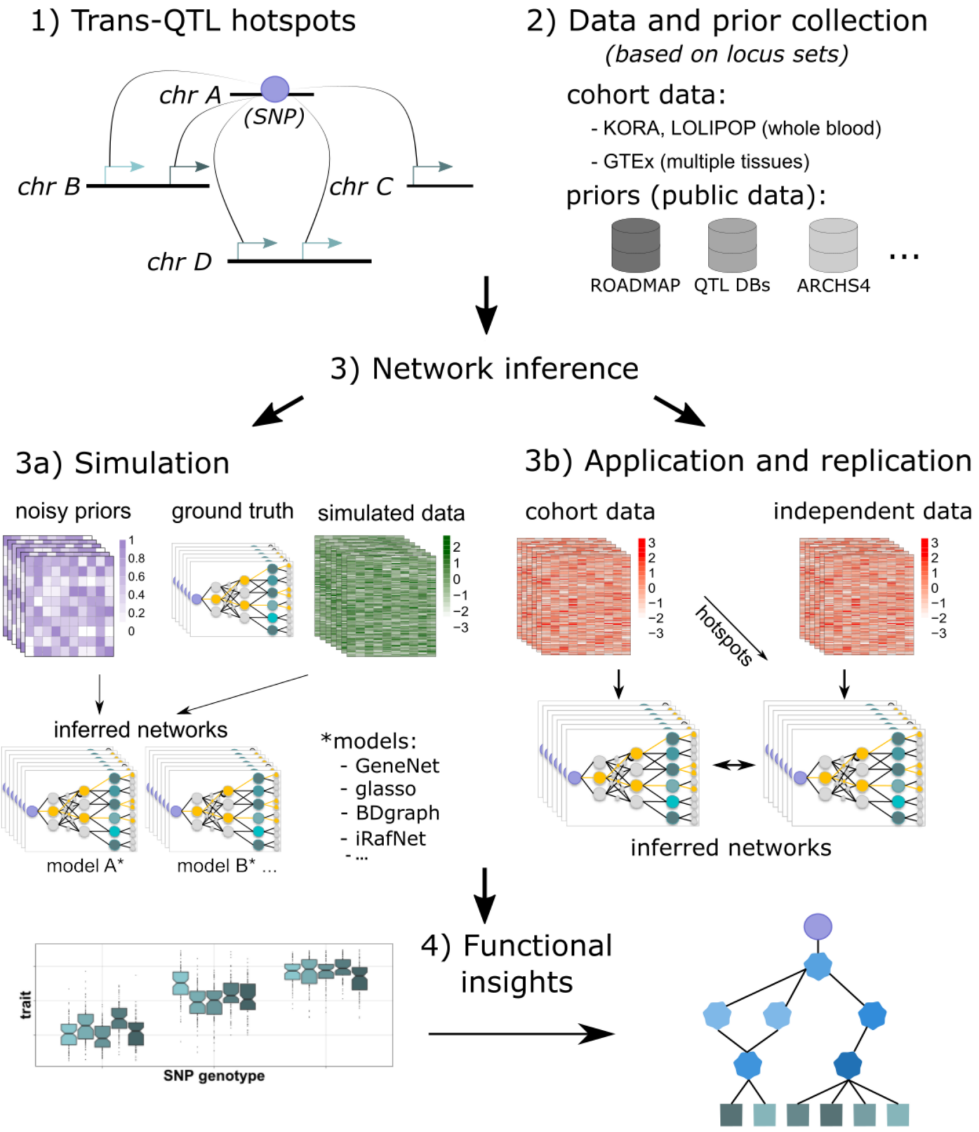


Figure 5.2.: The analysis plan followed in this project. We curate *trans*-QTL hotspots **(1)** and collect locus sets, for which we obtain functional data and interaction prior information **(2)**. Subsequent benchmarking **(3)** allows us to select the method best suited for application and interpretation of real-world networks **(4)**. Figure adapted from Hawe, Saha, Waldenberger, et al. [2].

In this study, we set out to explain the global effects of genomic master regulators for which we developed a new approach for prior based network inference on curated *trans*-QTL hotspots (see Figure 5.2). It involves a strategy for generating sets of entities important for each hotspot locus ('locus sets', Section 5.3.1) and deriving a compre-

hensive set of biological prior information from large-scale genomic databases such as Roadmap, GTEx, and BioGRID (see Section 2.3 and Table 1.2, results in Section 5.3.2). These priors are then integrated with the multi-omics data collected for the locus set using state-of-the-art network inference approaches such as the graphical LASSO [87], GENIE3 [88] and BDgraph [124]. To benchmark available methods and to select the one best suited for our application we performed an extensive simulation study followed by a replication analysis in the KORA and LOLIPOP cohorts (same as in Chapter 4, results in Section 5.3.3, see also Section 2.2 for details on the data). By applying selected methods on real-world population data we showed, that prior based network inference can replicate and extend previous findings and generate novel insights into the underlying mechanisms of disease-related *trans* hotspots. The strategy followed in this project is depicted in Figure 5.2 and entails four general steps:

1. curate QTL hotspots (1)
2. define locus sets and obtain functional data and priors (2)
3. select the best method based on simulation study and replication (3)
4. infer and interpret final networks (4)

5.3.1. Leveraging *trans*-QTL hotspots to reduce complexity

Trans hotspots represent important genetic variants which are enriched for disease-associated loci [22, 26, 210] and investigation of how they execute their influence on the associated *trans* traits can improve our understanding of regulatory patterns and disease.

For our analysis, we obtained *trans* hotspots from previously published QTL studies. We used the methylation QTL (meQTL) discovered in the Hawe et al. study also described in parts in the previous chapter in this thesis [1]. Further, we curated the *trans* expression QTL (eQTL) reported by Võsa et al. in light of the eQTLGen consortium [22]. In both cases, data were measured from whole-blood samples and the QTL obtained using a meta-analysis of multiple cohorts to increase detection power. Using these QTL results we were able to define 107 and 444 *trans* hotspots with at least 5 *trans* associations for the meQTL and eQTL sets, respectively (see Methods). Figure 5.3 shows an overview of all collected hotspots for meQTL (green) and eQTL (orange) binned according to chromosomal positions.

The figure shows that hotspots are equally distributed across autosomes 1-22 and that eQTLGen provides more hotspots with overall more associations than the meQTL results. This could stem from the larger effective sample size used in eQTLGen (yielding higher power to detect *trans* effects). We observe a maximum number of 477 (unbinned) *trans* associations for eQTLGen (mean 115) whereas for the meQTL the maximum number is 261 (mean 62). As we intended to demonstrate that our approach can be applied in arbitrary contexts we further curated hotspots from recently published GTEx

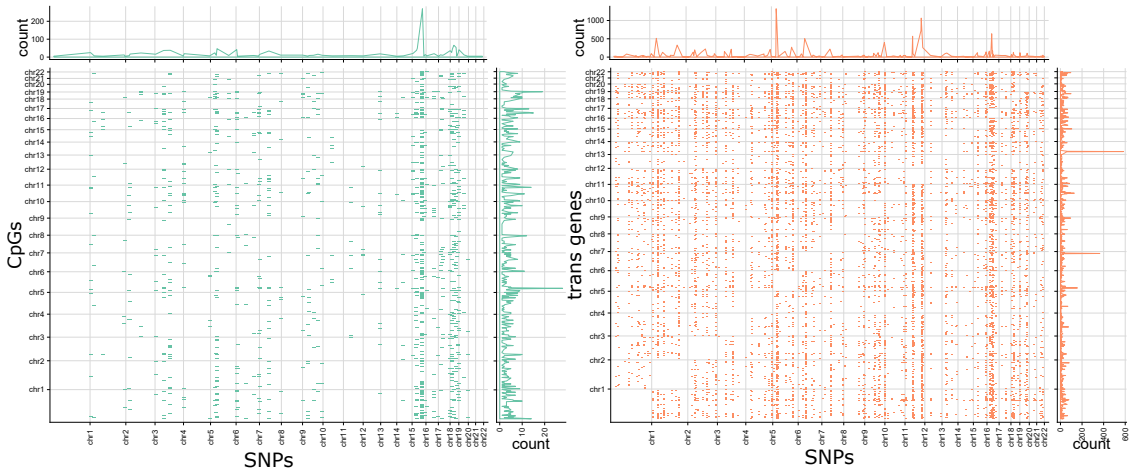


Figure 5.3.: Overview of all collected *trans* hotspots in this study, including 107 meQTL hotspots (green) and 444 eQTL hotspots (orange). x- and y-axes show genome locations of SNPs and genes/CpGs, respectively. Dots in the plot represent bins across the genome, margin plots indicate cumulative counts for individual bins.

v8 *trans*-eQTL [101]. The stringent filtering applied by GTEx to define *trans*-eQTL resulted in a single hotspot in Skeletal Muscle tissue for which we inferred regulatory networks (see Section 5.3.4).

To mitigate the $N << P$ problem we defined 'locus sets' which are intended to fully represent individual hotspots on a functional level. To this end, all relevant genes potentially mediating the observed *trans* relationships are included and used to bridge the gap between the involved chromosomes (via the inclusion of PPI and TFBS information, see Methods). Therefore, for each locus set, we collected the genetic locus (SNP) and corresponding *trans* traits (CpGs for meQTL and genes for eQTL) and additional genes, entailing genes encoded near the SNP (*cis* genes), genes encoded near CpGs (*trans* genes, meQTL only), proteins binding at *trans* entities (TFs) and genes of a PPI network, selected only if they are located on the shortest path between *trans* and *cis* entities. We show a summary of the collected entities over all 444 eQTL and 107 meQTL hotspots in Figure 5.4. The median number of entities amounts to 41 and 59 for meQTL and eQTL loci, respectively.

5.3.2. Collection of prior information

A central aspect investigated in this work is the application of prior knowledge to ease the reconstruction of regulatory networks. To achieve this, in addition to the functional data collected for each entity in the generated locus sets we gathered continuous priors for possible edges between these entities reflecting prior probabilities of observing the respective associations. We discriminate between four different kinds of edges for which we curate prior information and all priors are generated from data independent of the multi-omics cohort data used for inference but are matched for tissue context

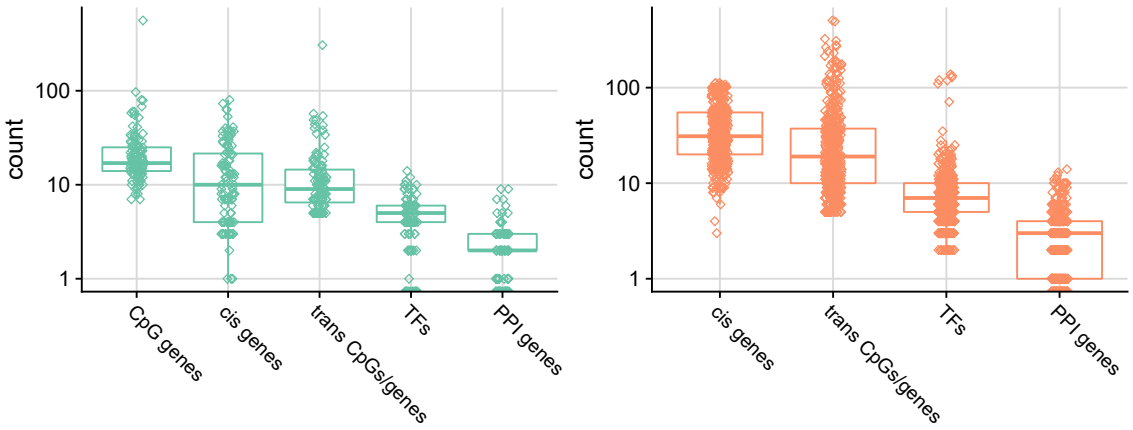


Figure 5.4.: Overview of the collected entities over all hotspots. x-axis indicates the different entity types, y-axis the total amount of these entities in the hotspots (log10-scale). Data are stratified in meQTL hotspots (green) and eQTL hotspots (orange). Figure adapted from Hawe, Saha, Waldenberger, et al. [2].

(see Methods). We display the overall number of edges for which we were able to derive informative priors for all hotspots in Figure 5.5A. Regardless of eQTL or meQTL hotspots, we can identify at least 2 edges with prior information (minimum 2 for meQTL, 3 for eQTL). Most hotspots overall receive relatively few priors (median 26 for meQTL, median 94 for eQTL) when compared to the total number of possible edges which could potentially be inferred. Overall, 8 and 209 hotspots are annotated with $\geq$ 100 edge-wise priors for meQTL and eQTL, respectively. In Figure 5.5B and 5.5C, we further see that, as expected, the number of edges annotated with prior information correlates positively with the number of possible edges, yet the fraction of all edges receiving prior information decreases with increasing numbers of possible edges.

5.3.3. Method comparison by simulation and replication study

We sought to perform a rigorous benchmark of inference methods (see Table 5.1) to determine the one best suited for our specific application context. To this end, we employed two distinct strategies:

1. an extensive simulation study evaluating the effect of sample size and prior information to reconstruct a simulated ground truth

2. a replication across the available cohort data sets to assess effect of priors on stability of reconstructed networks

In both cases we used Matthews Correlation Coefficient (MCC) [264] to evaluate the inferred networks either against the ground truth network structure (1) or against the network obtained on a different data set for the same locus and method (2), performing two comparisons selecting one of the inferred networks as the reference).

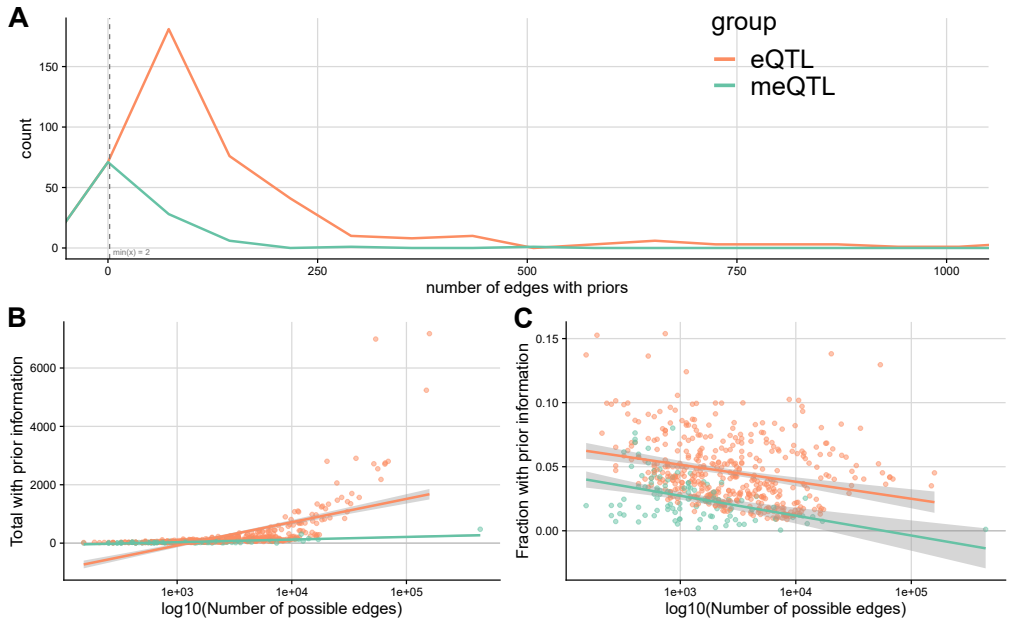


Figure 5.5.: Overview of the priors collected in the network inference study. A) shows the number of edges annotated with prior information (x-axis) over all curated hotpots (counts on y-axis). B) Shows the total number of edges with priors (y-axis) against the number of possible edges (x-axis, log10-scale) C) Shows the fraction of edges with priors against the total number of possible edges (x-axis, log10-scale). Lines in B) and C) represent the line of best fit from a linear regression model including standard error (shaded area). Figure adapted from Hawe, Saha, Waldenberger, et al. [2].

Simulation study shows benefit of data-driven priors

By executing an extensive simulation study we benchmarked five state-of-the-art network inference methods with respect to their capability of reconstructing simulated ground truth graphs in different scenarios (see Table 5.1 and Methods details). For this simulation study, we generated data and ground truth networks such that they reflect the collected real-world data for the 107 collected meQTL hotspots (see Methods). We considered 2 distinct scenarios: One, where we analyzed the impact of different sample sizes on network inference and one, where we investigated the benefit of priors and effects of varying degrees of noise in the priors. We generated 1,284 simulations for the prior-based analysis (12 different noise scenarios) and an additional 1,284 simulations for the sample size scenario (12 different sample sizes), for each of which we performed 100 iterations and generated all network inference models and obtained the MCC. The simulation strategy resulted in a total of 256,800 simulated data sets. For both gLASSO and BDgraph we set out to assess the relative difference in performance, each time training two distinct models once under consideration of prior information ($gLASSO_P$, $BDgraph_P$) and once neglecting it ($gLASSO$, $BDgraph$). As $iRafNet$ cannot be run without prior information, we did not assess relative differences based on priors. $GeneNet$ and $GENIE3$ cannot incorporate priors and hence were trained only

on the simulated data as a reference. The results for the sample size based analysis are presented in Figure 5.6 (see also Supplementary Table C.4).

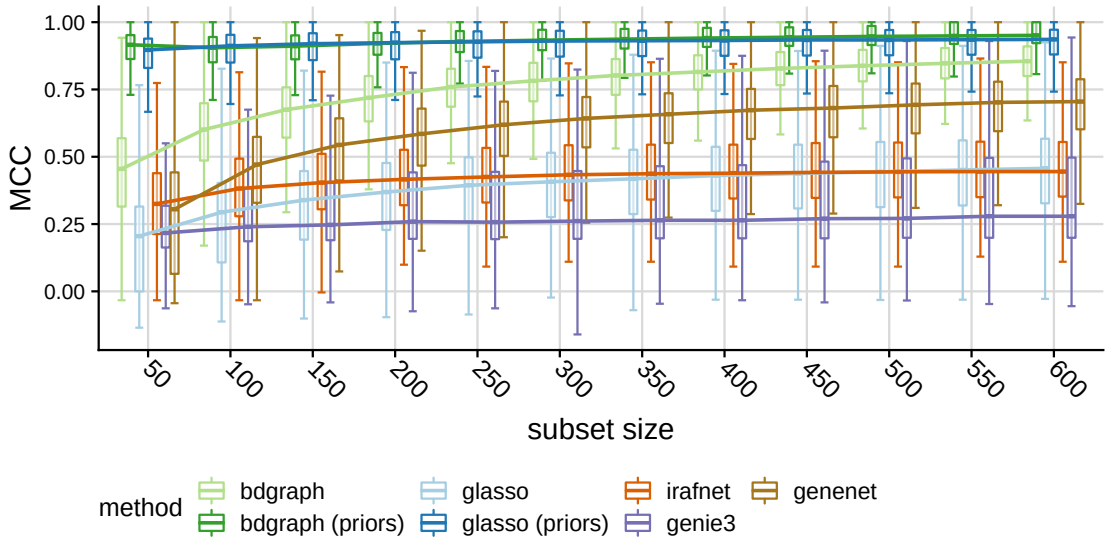


Figure 5.6.: Results of the simulation study for the sample size based analysis. y-axis shows the MCC, x-axis show the different sample sizes. 100 iterations are performed for each of the 107 loci and each sample size. Colors indicate different inference methods. Boxplots show median and upper and lower quartiles. Figure adapted from Hawe, Saha, Waldenberger, et al. [2].

The results on the sample size simulation show that including prior information can help to overcome low sample size problems, specifically for $BDgraph_P$ and $gLASSO_P$ which show almost steady performance across all sample sizes. For all other methods, an increase in sample size also improves inference as expected, although to different degrees. Notably, $GENIE3$ performance only increases marginally with increasing sample size compared to other methods.

In addition to the sample size based analysis, we evaluated the impact of noise in the prior information on network inference (Figure 5.7 and Supplementary Table C.3). The figure shows that the prior based methods $BDgraph_P$ and $gLASSO_P$ consistently outperform all other methods as long as prior noise does not exceed 30% or 20%, respectively, with $BDgraph_P$ outperforming $gLASSO_P$. Interestingly, $BDgraph$ is a close third and outperforms $BDgraph_P$ in high noise scenarios and *GeneNet* achieves relatively high performance followed by $gLASSO$. Although $iRafNet$ uses prior information it achieves only moderate performance. Finally, $GENIE3$ shows worse performance than $iRafNet$ which might be expected as these are similar approaches (tree-based) with $iRafNet$ using prior information and $GENIE3$ not using priors.

We further evaluated the performance of prior based methods when providing a prior merely on the sparsity of the true network (column 'rbinom' in Figure 5.7, see Methods). Our results show that knowing about the true density of the underlying graph structure (i.e. the total number of edges) helps to improve inference performance significantly.

Overall, our results suggest that carefully curated prior information indeed facilitates network reconstruction in a simulation scenario.

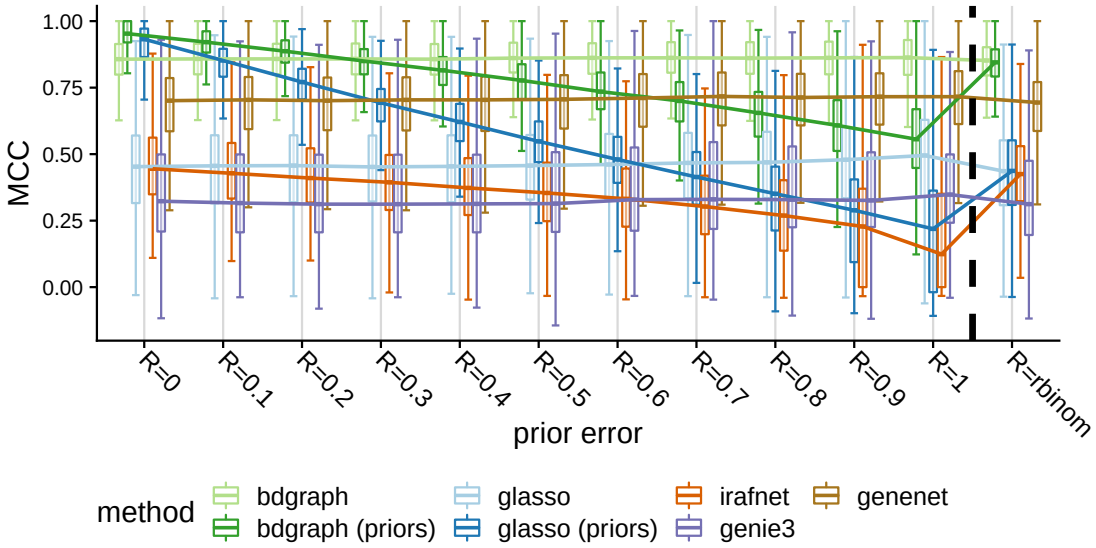


Figure 5.7.: Results of the simulation study for the prior assessment. y-axis shows the MCC calculated for inferred networks against ground truth networks, x-axis shows increasing levels of noise in the prior information from left to right. Column 'rbinom' indicates prior representing the degree distribution of the true graph. 100 iterations are performed for each of the 107 loci and each noise level. Colors indicate different inference methods. Boxplots show median and upper and lower quartiles. Figure adapted from Hawe, Saha, Waldenberger, et al. [2].

A final aspect of the simulation study is that we can investigate the performance of our methods in recovering mixed edges, i.e. edges between genotypes (discrete data) and gene expression of *cis* genes (continuous Gaussian data). This is of particular importance as the integration of mixed data modalities is one of the main obstacles in multi-omics data integration. Therefore, we specifically evaluated how well methods are able to recover edges between discrete and continuous nodes in the simulated data (see Methods) by assessing the overall fraction of recovered SNP-Gene edges over all simulations for each method. Figure 5.8 shows the results for the SNP-Gene recovery analysis.

One can see that prior based methods achieve best performance in this context, with the exception of *iRafNet*, matching the overall impression gained from the simulation results. Similar to the prior-noise simulation *BDgraph* performs better than other inference methods.

Inferred networks replicate in independent datasets

While the simulation study provides good evidence regarding method performance in a controlled setting, performance on real-world genomic data is typically harder to gauge due to a missing ground truth. As we have multi-omics data from two large population

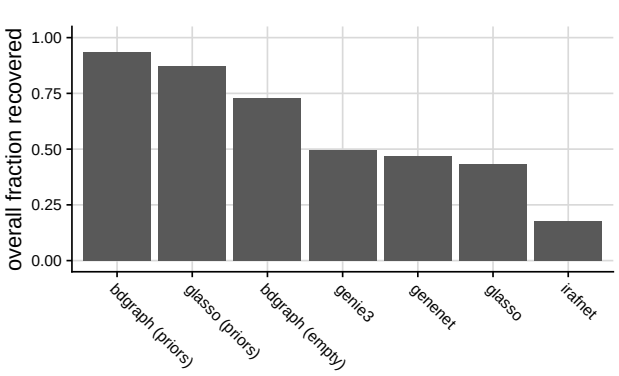


Figure 5.8.: Results for the SNP-Gene recovery analysis. Shown is the overall fraction of SNP-Gene links (discrete-continuous mixed edges) over all simulations (y-axis). X-axis shows the individual methods. Figure adapted from Hawe, Saha, Waldenberger, et al. [2].

cohorts available (KORA and LOLIPOP, see Section 2.2 and previous Chapter), which have both been measured in whole-blood data, we set out to perform a replication study on these data to assess method performance. These data consist of genotype, gene expression and DNA methylation data (microarray based, see also Section 2.1) and were measured over 683 and 612 individuals for KORA and LOLIPOP, respectively. We assessed method performance in terms of faithfulness of inferred networks across the two cohorts. Specifically, we performed cross cohort replication where we inferred networks from data of both cohorts separately and then obtained quantitative replication performance by calculating pairwise MCCs, always using one of the inferred networks as a reference. To this end, we collected data and priors for all entities over all hotspots in the respective locus sets. The results for this analysis are displayed in Figure 5.9 (see also Supplementary Table C.5).

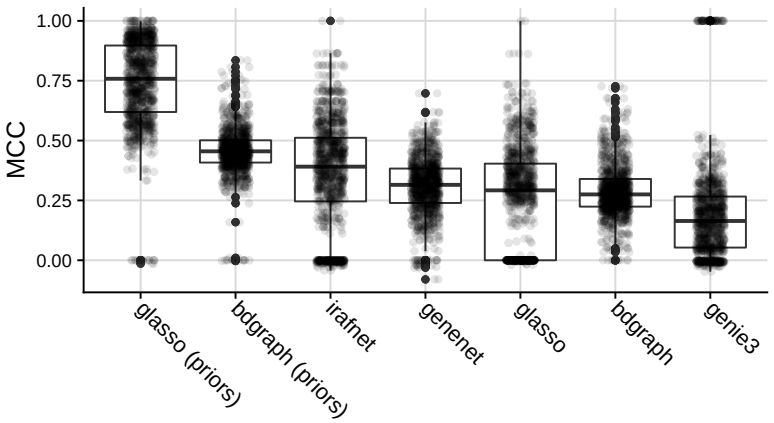


Figure 5.9.: Performance in the cross cohort replication analysis for all methods over all meQTL and eQTL hotspots. y-axis indicates cross-cohort MCC, two MCC are calculated per hotspot (using one of the networks as reference) and shown as individual dots. Y-axis indicates the different methods. Boxplots show medians (horizontal line) and first and third quartiles (lower/upper box borders). Whiskers show $1.5 * IQR$ (inter-quartile range). Figure adapted from Hawe, Saha, Waldenberger, et al. [2].

The results show, that all methods including prior information ($gLASSO_P$, $BDgraph_P$ and $iRafNet$) replicate better across datasets as compared to the non-prior methods, with $gLASSO_P$ performing best. Notably, $BDgraph_P$ shows much less variation in prediction performance than both $gLASSO_P$ and $iRafNet$. For the non-prior methods *GeneNet* achieves most robust performance followed by $gLASSO$ and $BDgraph$. Similar to the simulation study $GENIE3$ shows worst performance.

Estimated transcription factor activities improve network replication

Potential binding of transcription factors (TFs) at genomic loci *trans* associated to our hotspots is central to our analyses, as they effectively bridge the gap between *cis* and *trans* QTL entities. It has previously been proposed that gene expression measurements might not reflect the true activity of TFs (which could for instance be driven by a TF's phosphorylation state) [267]. We hence investigated the use of transcription factor activities (TFAs) instead of TF expression for network inference as has been proposed in [267]. TFAs can be estimated from the TF's expression together with the expression of its target genes, which can be defined via ChIP-seq experiments. Therefore, we estimated TFAs for each of the TFs in our data based on our expression data and collected TF binding sites [71] (see Methods for details). We then applied the same replication strategy as above to evaluate the robustness of the inferred networks and compared TFA based results to expression-based results. The results for this analysis are displayed in Figure 5.10.

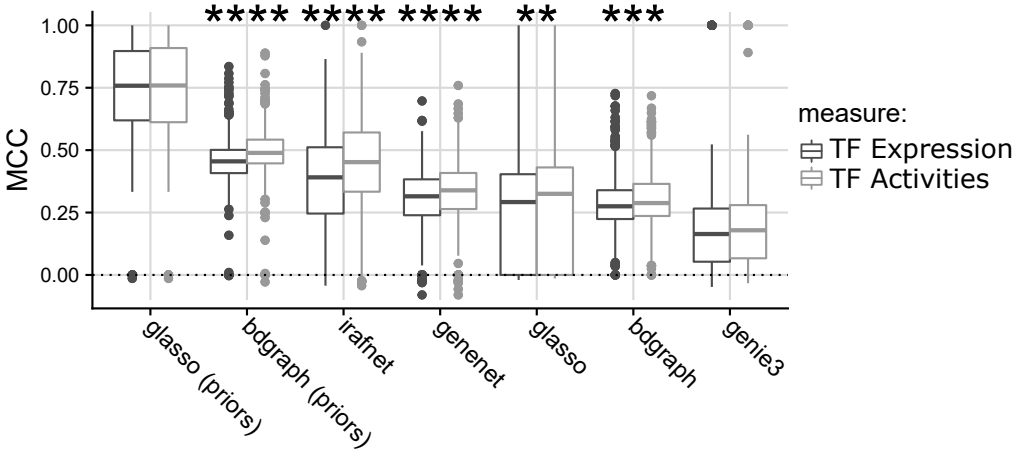


Figure 5.10.: Results of the TFA replication analysis compared to the original expression based cohort replication. y-axis shows the MCC across cohorts, x-axis indicates the different methods. Grey scale indicates the type of data used for TF measurements (TF Activities or Expression). Boxplots show medians (horizontal line) and first and third quartiles (lower/upper box borders). Whiskers show $1.5 * IQR$ (interquartile range). Points indicate outliers (beyond whiskers). Stars indicate significant difference between TFA and expression results for each method (Wilcoxon test, **: $P \leq 0.01$, ***: $P \leq 0.001$, ****: $P \leq 0.0001$). Figure adapted from Hawe, Saha, Waldenberger, et al. [2].

In this analysis, we found that for all models but $gLASSO_P$ and $GENIE3$ the use

of TFAs significantly improves the cross cohort MCC compared to the expression based analysis (Wilcoxon $P < 0.01$). Moreover, an increase in MCC is also visible for $gLASSO_P$ and $GENIE3$ indicating improved performance also for these methods.

Based on the overall method evaluation we decided to prioritize methods inferred from $gLASSO_P$ and $BDgraph_P$ for more detailed investigation of regulatory patterns and substituted TF expression values for TF Activities in all subsequent analyses.

5.3.4. Application to real-world population data

In the previous sections, we established the best method for inferring networks from *trans*-QTL hotspots under consideration of prior information, for which we evaluated summary performance measures for reconstructing ground truth graphs or replicating networks across datasets. A natural next step is to look in detail at inferred networks based on the methods deemed best suited for the task of explaining *trans* hotspots. In this section, we will first show the replication of our results from the previous chapter using the selected methods. We will then highlight two novel genetic hotspot loci identified in the whole-blood data from KORA and LOLIPOP and Skeletal Muscle data from GTEx [101], for which we inferred networks and generated novel regulatory hypotheses.

Replication of previous findings by simultaneous data integration

We previously constructed regulatory networks in a semi-integrative approach involving two separate steps, 1) performing a random walk on established locus graphs and 2) complementing network interactions with a subsequent local correlation analysis. The approach in this project is based on a fully integrative inference strategy, simultaneously integrating all available data and using annotations (e.g. PPI, ChIP-seq) as prior information. Generally, this approach enables the inference of additional edges which cannot be identified by the previous approach (e.g. Gene-Gene edges) or are tested isolated from the rest of the network (e.g. CpG-Gene edges). In this comparison, we analyzed three of the highlighted networks derived in the previous study and compared them to the data-drive networks established in this chapter. The results of this comparison are summarized in Table 5.2.

locus	num. nodes	num. edges	common edges	MCC
rs9859077	99 (89)	447 (287)	141	0.517
rs730775	58 (49)	98 (67)	48	0.689
rs7783715	25 (17)	24 (23)	5	0.65

Table 5.2.: Comparison of inferred networks from this study and the networks generated in the meQTL random walk analysis described in Chapter 4. Numbers in brackets indicate original network statistics. Table taken from Hawe, Saha, Waldenberger, et al. [2].

Concordance between the networks is relatively high with MCC values of 0.52, 0.69 and 0.65, respectively. As expected, the simultaneous inference yields overall more

edges and contains more nodes compared to the previously derived network (56%, 46%, and 4% more edges and 11%, 19% and 47% more nodes). We hypothesize, that the two-step approach might have missed these additional edges and nodes as it purely relies on already known PPI and ChIP-seq interactions. We set out to investigate one of the comparisons in detail and choose the locus around the *rs730775* SNP as an example due to its relatively small size (Figure 5.11). Here, we annotated the network derived from our integrative approach with the original network indicating novel (green) and replicated (orange) edges.

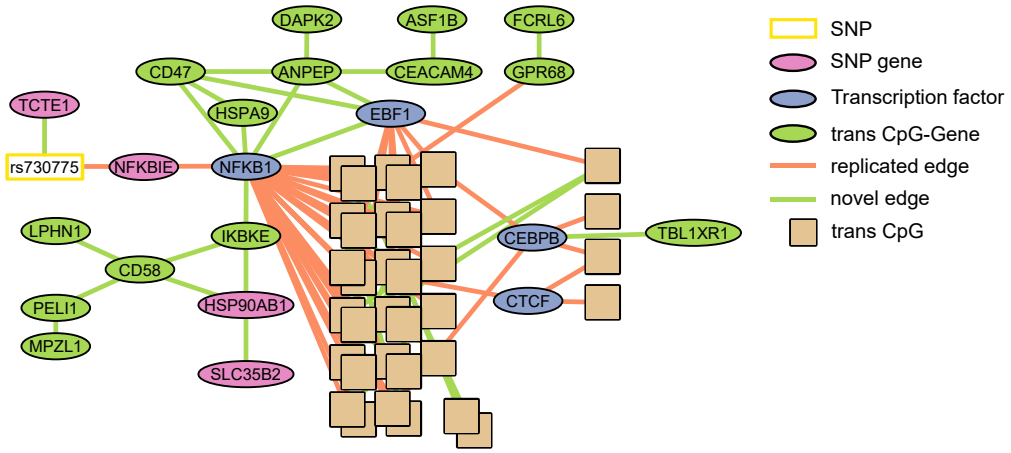


Figure 5.11.: Comparison of the two networks for the *rs730775* locus for the random walk approach and the integrative analysis. Network derived from the omics-data is used as the scaffold to indicate replicated (orange) and novel (green) edges. White box represents the SNP, pink nodes SNP-genes, blue nodes transcription factors, brown boxes CpG sites and green nodes CpG-genes. Figure adapted from Hawe, Saha, Waldenberger, et al. [2].

Our network recovers the main pathway described in the original network which connects the SNP *rs730775* via *NFKBIE* and *NFKB1* to the meQTL associated *trans* CpGs. Moreover, we discover several of the TFs reported in the previous network and find additional interactions between CpG genes and the transcription factors which could not be picked up with the other approach. These results show that the integrated inference (under consideration of biological prior information) can both be used to replicate and extend previous networks, potentially adding novel insights into the genetic and epigenetic regulation of target genes.

A *trans* regulatory network for a schizophrenia susceptibility locus

Next, we systematically applied our inference approach to all collected meQTL and eQTL hotspots to demonstrate how it can be used to derive novel mechanistic insights for *trans*-QTL hotspots and complex traits. In order to select interesting networks, we filtered results for loci which have been associated in a GWA study and devised a graph score, which reflects desirable biological properties of *trans* hotspot graphs and which is employed for additional ranking of our results (see Methods).

Following this strategy led us to select a genetic locus around the *rs9469210* SNP in the Human Leukocyte Antigen (HLA) region, which has previously been associated with schizophrenia (SCZ) [284, 290][19]. The resulting network is presented in Figure 5.12A. *rs9469210* is connected to three *cis* genes in the network (*PBX2*, *RNF5* and *HLA-DQA1*), for all three of which it is also a *cis*-eQTL in eQTLGen [22]. The locus-associated *RNF5* gene was differentially expressed in a study comparing SCZ cases vs controls, and its expression is also associated with a schizophrenia risk SNP independent of *rs9469210* (*rs3132947*, $R^2 = 0.14$ in 1000 genomes Europeans[20]) [291]. *PBX2* was linked to a SCZ related phenotype in a recent pharmacogenetics study (clozapine-induced agranulocytosis) [292, 293] and is linked in our network to *SPI1*, also showing ChIP-seq binding evidence of PBX2 to its promoter. *SPI1* (*PU.1*) is a well known transcription factor which has previously been implicated in Alzheimer's Disease by influencing neuroinflammatory response pathways [294] and interacts with its associated network neighbor *RUNX1* to modulate gene expression [295]. In addition, *RUNX1* has been associated with rheumatoid arthritis (RA). RA is inversely related to schizophrenia and implication of *RUNX1* in RA might therefore indicate a role in SCZ [296]. Moreover, HLA related genes are connected to both *SPI1* and *RUNX1*. The HLA locus has previously been implicated with SCZ and other psychiatric diseases [297–300]. The transcription factor *TCF12* has not been implicated in brain related disorders so far, however, its paralogs *TCF4* and *TCF3* are well known E-box TFs which are expressed in multiple brain regions [301]. In addition, *TCF4* loss-of-function mutations are causative for Pitt-Hopkins syndrome [302], which causes, amongst other phenotypes, mental disabilities and behavioral changes, and SNPs influencing *TCF4* activity have also been implicated in SCZ [303, 304].

The remaining TF in the network, *NFKB1*, is involved in the regulation and development of neural processes and has been implicated in a variety of disease phenotype, including SCZ [305]. The rest of the network is comprised of 9 out of the 40 initially discovered *trans* associated genes of the *trans* acting locus which are connected via the intermediate TFs. Several of these *trans* genes have been linked to neurological disorders related to SCZ, including *PSEN1*, *B9D2*, *CXCR5* and *DNAJB2* [306–309], and *SH3BGRL3* has directly been implicated in schizophrenia [310]. Additionally, Rodriguez and colleagues [311] linked the *trans* gene *RNF114* to the *NFKB1* pathway mentioned above. To establish the possibility of a common underlying causal variant between the selected eQTL genes and schizophrenia, we further executed a formal colocalization analysis of eQTL and GWAS [312] signals using fastENLOC (Figure 5.13) [237]. Briefly, *fastENLOC* provides a quantitative assessment for the enrichment of molecular QTLs in complex trait-associated regions and the colocalization of both signals (i.e. QTL and GWAS). Here, we observed a strong colocalization signal between the GWAS and three of the selected *trans* genes (*PSEN1*, *DNAJB2* and *CD6* with SNP-level colocalization

[19] *rs9274623*, which is an alias of the *rs9469210* SNP according to SNiPA: `https://snipa.helmholtz-muenchen.de/snipa3/`

[20] https://ldlink.nci.nih.gov/?tab=ldmatrix

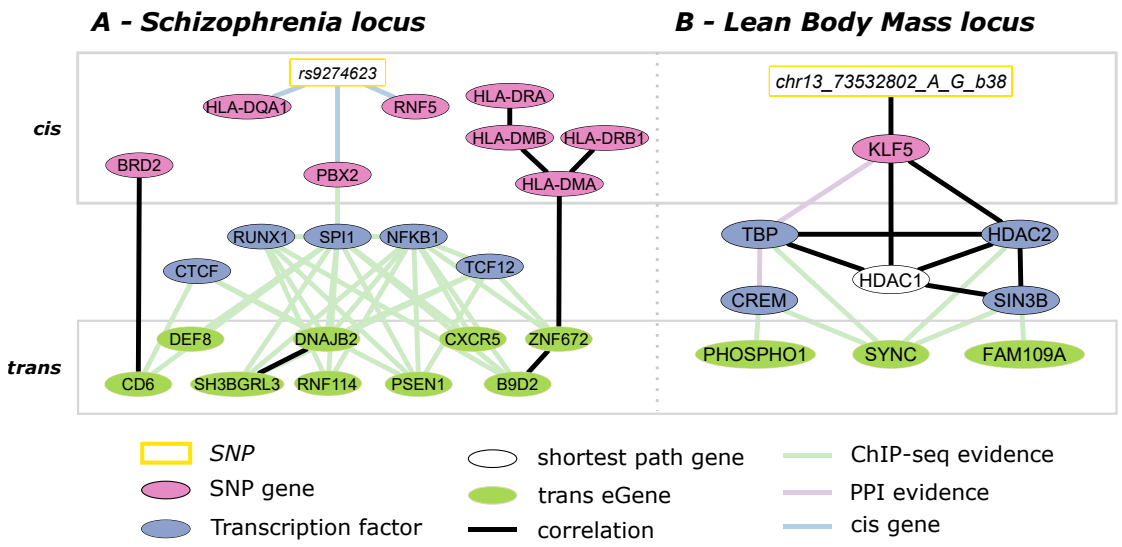
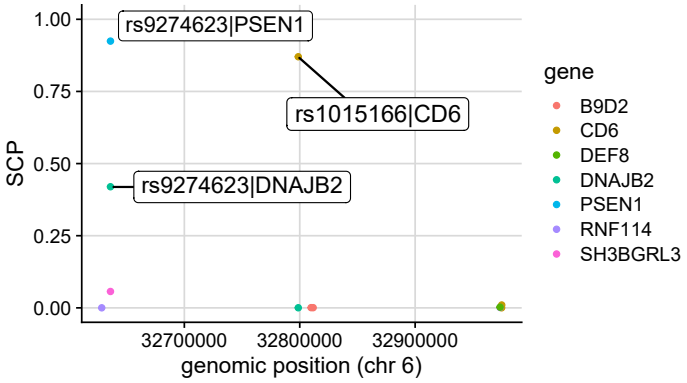


Figure 5.12.: The two highlight networks inferred from functional multi-omics data using prior information to guide the inference. A) Network around the *rs9469210* SCZ susceptibility locus. B) Network inferred in GTEx Skeletal Muscle data involving the lean body mass associated hotspot around *rs9318186* . Figure adapted from Hawe, Saha, Waldenberger, et al. [2].

probabilities 0.92, 0.87 and 0.42; see Methods for details).

Figure 5.13.: Results for the colocalization analysis for the SCZ locus for SCZ GWAS and eQTL signal obtained from eQTLGen. Y-axis indicates SNP-level posterior colocalization probability as reported by fastENLOC, x-axis chromosomal position of SNPs. Colors indicate different *trans*-eQTL genes. Figure adapted from Hawe, Saha, Waldenberger, et al. [2].



Using our inference strategy we were hence able to generate functional hypotheses for the regulatory pathways underlying a *trans*-QTL hotspot related to SCZ, highlighting several already disease related genes and potential novel targets for future investigations. The strong evidence from our colocalization analysis further indicated a potential common causal variant between discovered *trans* genes and schizophrenia.

Application to Skeletal Muscle

We established a network inference strategy and applied it to data obtained from whole-blood samples collected in light of the KORA and LOLIPOP cohorts. Now, we show that the strategy can be transferred to a different biological context. Specifically, we set out to investigate *trans*-eQTL hotspots identified in light of the GTEx project (see Section 2.3.2) [82, 93]. Based on the latest v8 release of GTEx [101] we could extract a single *trans*-eQTL hotspot identified in Skeletal Muscle tissue (see Methods) and defined its locus set as we did for the other hotspots. A particular obstacle for the application of our approach on Skeletal Muscle tissue is the lack of a comprehensive set of TFBS for this tissue. To overcome this we predicted TFBS from muscle-related DNAse-seq data and used the predicted binding sites to define the locus sets and priors (see Methods). In the whole-blood analysis, we also utilized GTEx data to generate priors and, as this is not appropriate when using (the same) GTEx data also during inference, we collected independent muscle specific gene expression and eQTL data to define priors (see Methods). Using the GTEx v8 Skeletal Muscle genotype and gene expression data together with the collected priors we inferred a regulatory network for the *trans* hotspot which is shown in Figure 5.12B (based on $gLASSO_P$).

In this network, the hotspot variant *rs9318186* which is also a *cis*-eQTL of *KLF5* in GTEx v8 Skeletal Muscle ($P = 6.1 \times 10^{-37}$) is associated with *KLF5*, and a SNP in high LD ($R^2 = 0.88$) has further been linked to the *lean body mass* phenotype (LBM). In a transcriptome-wide association study Singh et al. [313] associated *KLF5* with LBM by integrating LBM GWAS results with gene expression. The gene has also been linked to lipid metabolism via studies performed in *KLF5* knockout mice [314]. Other genes of the network have been implicated in similar phenotypes. For instance, in a study by Moresi et al. [315] the authors found that *HDAC1* and *HDAC2* are involved in controlling skeletal muscle homeostasis in mice. Moreover, both genes are involved in muscle development and interact with the SIN3 core complex (involving the network gene *SIN3B*, to regulate gene expression [316]. *TBP* (TATA-binding protein) has not yet been associated with LBM, but is an intensely studied transcription factor central in the regulation of numerous genes [317]. *CREM*, the final transcription factor picked up in the network has not been linked to LBM before. However, a genome-wide association study of elite sprinter status [318] indicated an association of a *CREM* related SNP (rs1531550, $P = 1.88 \times 10^{-6}$) in this trait. All three of the *trans* genes included in the network have been implicated in LBM related traits. The gene *SYNC*, for example, interacts with dystrobrevin to uphold muscle function in mice and has been linked to neuromuscular disease traits [319, 320]. Additionally, *PHETA1/FAM109A* expression was linked to Body-Mass-Index (BMI) in a study by Seim et al. [321]. The final *trans* gene *PHOSPHO1* has been identified in several studies investigating metabolic traits. Oldknow and colleagues [322] proposed a role of *PHOSPHO1* in energy homeostasis and Wahl et al. linked it to BMI via DNA methylation [323]. In addition, it has been associated with HDL levels, a trait inversely linked to LBM [324] and Dayeh et al. showed that DNA is

hypomethylated at the *PHOSPHO1* locus when comparing diabetic and non-diabetic skeletal muscle samples.

Therefore, our network indicates that the regulatory role of *KLF5* may be realized through interaction with the SIN3 core complex, potentially involving both *TBP* and *CREM*, to act on the linked *trans* genes *PHOSPHO1*, *SYNC* and *PHETA1/FAM109A*.

5.4. Project summary

In the work presented in this chapter, we developed a Bayesian framework to infer undirected gene regulatory networks from *trans*-acting genetic hotspots by simultaneous integration of human multi-omics data with existing biological prior knowledge. To this end, we generated a comprehensive set of edge-specific prior information from large-scale genomic databases including functional data and established interaction resources. We executed an extensive simulation study to evaluate recent network inference methods and where we investigated the impact of sample size and prior noise on inference performance. The performance was assessed on 256,800 simulated data sets, simulated according to parameters observed in real-world data (number of samples and number of nodes). For these data, we demonstrated that prior-knowledge informed methods outperform non-prior methods in reconstructing ground truth networks. As expected, increasing amounts of noise, i.e. percentage of incorrect edge priors, in the prior information lead to a significant decrease in method performance. Overall, we found that the Markov-Chain-Monte-Carlo based BDgraph outperforms all other methods both when including and neglecting prior information (in the latter case only for high noise scenarios). Both $BDgraph_P$ and $gLASSO_P$ outperform other methods in correctly identifying 'mixed' edges, i.e. edges between nodes of different data types such as SNPs (discrete genotypes) and genes (continuous gene expression) in our case. The tree-based methods, which are expected to perform better than regression-based models in mixed settings as they do not make any distributional assumptions, showed overall worst performance. Simulation results were confirmed by performing a cross-cohort replication analysis in the KORA and LOLIPOP data where all prior based methods replicate more faithfully across datasets than the non-prior methods.

The BDgraph and graphical LASSO inferred networks from real-world data were subjected to a detailed investigation based on their benchmark performance. Good replication and extension of the random walk based networks from the previous chapter showed that our fully integrative strategy can successfully reconstruct networks from multi-omics data. In addition, our new approach identified interactions not previously described, therefore yielding additional insights into the molecular regulatory cascades around transcription factors and DNA methylation. Systematic application of our strategy to data from KORA, LOLIPOP, and GTEx showed that it enables the recovery of known disease genes and the generation of novel molecular hypotheses underlying *trans* hotspots. Our analyses highlighted several known schizophrenia genes for a schizophrenia associated genetic locus, including HLA genes and *RNF5* [291, 297],

together with genes involved in related neurological disorders such as *PSEN1* and *PBX2* [292, 293, 306]. *RUNX1* has previously been hypothesized to be related to SCZ as it has been implicated in rheumatoid arthritis (RA), which is inversely related to SCZ [296]. We corroborated this hypothesis using our networks and implicate e.g. *BRD2*, *DEF8*, and *RNF114* as additional genes potentially playing a role in schizophrenia. These hypotheses are substantiated in the form of a colocalization analysis of SCZ GWAS [312] and independent *trans*-eQTL signals of the *trans* genes connected to the network which yielded high SNP-level colocalization probabilities for the involved molecular and disease traits. We applied our approach to a *trans*-eQTL hotspot derived from GTEx Skeletal Muscle tissue to show that it can be applied in diverse contexts. The hotspot has been associated with lean body mass (LBM) and our approach pinpointed several genes in the network related to this or a related trait including, for instance, *KLF5* (involved in LBM, lipid metabolism), *PHOSPHO1* (associated with BMI), *HDAC1* and *HDAC2* (control skeletal muscle homeostasis) and *SYNC* (maintains muscle function) [313–315, 320, 323].

Overall, in this project, we described a novel approach to alleviate gene regulatory network reconstruction from *trans*-QTL hotspots by curating a comprehensive set of priors from large-scale genomic databases and combining these priors with human population scale multi-omics data. The constructed regulatory networks can be used to generate mechanistic insights into the genome-wide *trans* effects of genetic hotspots and allow the derivation of new hypotheses about the functional consequences of disease-associated loci. Our extensive benchmark in simulated and real data of current inference algorithms provides guidance to choose the method best suited for specific inference tasks (see Discussion). Moreover, it showed that established data sources can be used to formulate biological priors which largely benefit network reconstruction.

6. Discussion

The downstream effects of disease-associated genetic variants remain largely unknown. Especially for *trans*-acting variants, which influence quantitative molecular traits across chromosome boundaries, it is not straight-forward to understand how their effects propagate through the cell. For instance, regulatory networks underlie important *trans*-acting hotspots, which affect numerous quantitative traits across the genome, and recovering these networks can help in explaining the effects of these genomic master regulators. The main contribution of this thesis is to generate new insights into genetic *trans* hotspots by devising novel techniques for network analysis and biological data integration, thereby advancing our understanding of complex traits.

6.1. Systematic assessment of the genetic effects influencing DNA methylation

Methylation of DNA is an important epigenetic mark and has been associated with complex diseases such as cancer and type 2 diabetes [52, 54]. In our study, we investigated methylation quantitative trait loci (meQTL) derived from 6,994 individuals and replicated our findings in isolated leukocytes, adipocytes and in adipose tissue. We generated a genome-wide catalog of 2,709,428 genetic variants influencing DNA methylation at 70,709 CpG sites in *cis*, *longrange* and *trans*. Other studies investigated associations between genetic variants and CpGs, however, they worked with smaller sample sizes or lacked replication [56, 59, 211, 325] or did not provide cross tissue replication [18, 56, 59, 325]. Comparison to previous publications [23, 211] indicated good agreement of results and highlighted numerous novel findings, showing the benefit of our large sample size and analysis strategy. Moreover, to our knowledge, we were the first to provide both a systematic investigation and a detailed characterization of often disease-associated *trans*-acting loci by employing novel strategies for network analysis and data integration. In Bonder, Luijk, Zhernakova, et al. [23], for instance, the authors provided a deep functional characterization of *trans*-acting loci but lacked a systematic analysis, whereas in Huan, Joehanes, C. Song, et al. [18] the authors systematically established *trans* hotspots but lacked a deep functional follow up of their identified loci.

We established the functional relevance of meQTL through extensive enrichment analyses using independent multi-omics data. We provided new evidence, that *cis* associations of SNP and CpGs reflect a change in the DNA sequence of *cis* regulatory elements, which alters DNA methylation and likely affects transcription factor binding [326]. Importantly, our study size allowed a systematic investigation of *longrange*- and

trans-meQTL effects. We identified novel enrichment of these genome-wide associations in topologically associating domains (TADs) and in inter-chromosomal chromatin contacts derived from Hi-C data, consistent with putative promoter-enhancer interactions: A genetic variant in an enhancer could lead to altered transcription factor binding and subsequently to a change in methylation in a promoter element, and the promoter and enhancer could form a connection via a chromatin loop [15, 326, 327]. While similar findings have been made in recent expression-QTL studies [22], we were the first to report these in an epigenetic context of this scale. In addition, in line with previous studies [23, 59] our analyses showed enrichment of meQTL for association with gene expression and enrichment of epigenetic regulators at genetic loci, substantiating their involvement in genome regulation.

Our meQTL analyses were based on the assessment of DNA methylation from the 450k array, and the non-random array content could introduce a bias in our analyses, leading to false-positive findings [169]. However, our strategy was designed specifically to avoid such a bias, particularly through the application of permutation testing for all enrichment analyses to quantify expectations under the null hypothesis based on array design. While the 450k array profiles methylation at approx. 450,000 CpG sites, recent advances in methylation profiling such as whole-genome bisulfite sequencing or the Illumina EPIC array allow potentially more complete coverage of the human genome. The incomplete coverage of the 450k array could hinder the identification of *trans*-acting methylation signatures and mask potential causal associations. Our additional analyses of EPIC array data showed that most TF enrichments from the 450k based analysis are also identified in the EPIC data, and that enrichment of identified TF signatures is stronger in EPIC compared to 450k derived results. Similarly, over-representation of meQTL SNPs and CpGs in chromHMM states is observed for both arrays, with the fraction of pairs in the same state being higher for the EPIC array compared to the 450k array. Therefore, our results from the EPIC analysis corroborated our previous findings of TF signature enrichment at *trans*-meQTL hotspots. They further showed little evidence of false-positive identification of transcription factor enrichment and indicated slightly more power to discover functional links for the larger and denser EPIC array as compared to the 450k array, as expected.

Our study was focused on SNP-CpG associations and the influence of methylation on gene expression was only studied for independent methylation loci. While the independent methylation loci represent important regulatory actors, a global analysis of CpG effects on genes could provide additional insights into the regulatory landscape surrounding DNA methylation. Moreover, novel developments in experimental assays, such as methyl self-transcribing active regulatory region sequencing (mSTARR-seq) [17], could be used to corroborate our independent sentinel SNPs and CpGs. For instance, mSTARR-seq allows determining methylation-dependent and non-methylation-dependent gene expression regulation on a genome-wide scale and thus could help delineate causal dependencies between methylation and gene expression [17]. Therefore, it could be applied to experimentally determine the functional impact of sentinel CpGs

by evaluating whether changes in their methylation directly affect gene expression.

6.2. Identification of *trans*-acting regulatory mechanisms underlying DNA methylation

The substantial sample size and replication strategy allowed the discovery and systematic analysis of important *trans*-meQTL hotspots, which highlighted several genetic loci involved in genome regulation. Recent analyses investigated *trans* hotspots through establishing binding sites at associated *trans* sites of an in *cis* encoded transcription factor [23], thereby only permitting a direct link between the genetic locus and the quantitative traits. Establishing indirect links, e.g. through interaction networks, is still a major bottleneck in *trans*-QTL analyses [22] and only a few works set out to investigate these [29]. We significantly advanced analyses of important *trans* hotspots and were the first to design and systematically apply a current network analysis approach to recover indirect connections through molecular interaction networks in human population data. By generating networks for 115 *trans*-meQTL hotspots we improved upon previous studies, which either lacked sample size to discover genome-wide *trans* effects [56, 328] or did not provide a systematic and detailed assessment of their (indirect) underlying mechanisms [23, 59, 325]. To this end, we implemented a sophisticated network analysis approach based on random walks on *trans*-acting hotspots integrated with curated protein-protein interaction and ChIP-seq networks. Networks obtained through the random walk approach were subsequently corroborated using functional cohort data in a local correlation analysis, thereby adding another layer of information and extending the functional interpretation of the networks. For numerous hotspots, we were able to pinpoint the candidate genes and cellular pathways mediating genome-wide genetic effects on DNA methylation and gene expression. We confirmed previous findings, involving for instance the genetic locus around *SENP7* [211], and provided novel hypotheses, for example for the *NFKBIE* and *ZNF333* *trans*-acting loci, thus advancing our understanding of the gene regulatory mechanisms underlying genetic and epigenetic control. The *NFKBIE* locus has been associated with rheumatoid arthritis (RA), a disease which can be medicated using IL-6 targeting drugs. Importantly, we were the first to provide a detailed molecular hypothesis of how the underlying genetic locus affects IL-6 regulation and further supported our network findings by performing a formal colocalization analysis of RA GWAS and *trans*-meQTL signals, indicating a common causal variant driving the observed phenotypes. Moreover, as a proof of principle, we validated the hypothesized mechanisms around the novel *ZNF333* locus experimentally, confirming the validity of our computational analyses.

We set out to extend our network analysis by devising a novel Bayesian framework for simultaneous multi-omics integration in network inference tasks under consideration of biological prior knowledge. This allowed an independent confirmation of the multi-step pathways established using the random walk approach. What is even more, the revised

strategy allowed us to propose novel molecular interactions for these pathways, yielding further insights into DNA methylation regulation. These interactions could not be detected by the previous approach, which focused on established interactions from PPI networks and ChIP-seq derived binding sites and therefore did not test all possible edges for associations. In the simultaneous inference approach, all edges regardless of presence or absence of prior evidence are tested, and therefore interactions will be discovered if they are strongly supported by either the functional data, the prior information, or both.

6.3. Biologically informed priors improve network inference

Based on our rigorous benchmark of state-of-the-art network inference algorithms, several practical considerations arise. We showed, that biologically informed prior information improves inference performance significantly, even in settings with low to medium amounts of noise, i.e. wrong prior edges, in the data. This is in line with previous results, e.g. in a modified graphical LASSO context [30] or a Bayesian regression context [128]. However, too much noise can severely impact inference performance, and thus care should be taken when curating prior information from public genomics resources to allow only high-quality data to enter the prior definition. For instance, by only considering experimentally validated protein-protein interactions, high-quality gene expression data, and other, experimentally derived interactions, noise levels can be kept low while simultaneously providing comprehensive edge-wise priors defined from e.g. gene co-expression analyses, allowing the use of available prior knowledge to its full extend. In addition, good replication performance across cohorts for prior-guided methods indicates that using prior information combined with functional data yields more faithful and, therefore, higher confidence networks as compared to not using priors. Overall, our results showed that prior based methods outperform prior-agnostic methods, an observation also previously made [128, 132, 256], and which highlights the benefit of using biologically informed priors for inference.

A critical aspect in genomics and specifically in gene regulatory network inference is the $N << P$ problem, i.e. the number of variables P (e.g. genes) is usually much larger than the number of available samples N. Specialized approaches have been proposed to alleviate this issue, for instance the graphical LASSO by applying regularization [3, 87]. Our results showed, that prior information significantly improves performance even in settings with comparatively low sample sizes. Moreover, applying our specialized framework yields locus sets with a relatively low number of entities as compared to the number of samples for which data are available. The benefit of alleviating the $N << P$ problem using locus sets comes with a small risk of excluding genes that would be needed to obtain a complete description of the observed *trans* effects of the hotspots. For example, we only included genes on the direct (shortest) path between the *cis* and *trans* entities, as we assume that most of the key regulators reside on that path in the extracted PPI network. Nevertheless, our stringent framework for defining locus sets,

collecting relevant TFs, PPI derived genes, and genes in the vicinity of the hotspot SNP and CpG sites, should include almost all of the key regulators in the network inference and lead to parsimonious, easily interpretable results.

We aimed to integrate multi-omics data and hence included methods able to cope with mixed data types, for instance, discrete genotypes and continuous gene expression values. While we would assume that tree-based methods perform well in mixed settings due to their lack of distributional assumptions on the data [32, 96], the copula model based BDgraph [98, 124] overall yielded the best performance for the inference of discrete-continuous interactions. Therefore, in mixed settings, BDgraph should be preferred, even more so if informative prior information is available. However, although the MCMC based BDgraph method showed overall stable performance and outperformed the graphical LASSO, it also exhibits much longer run time which needs to be considered in practical applications.

Lastly, we started off our analyses by applying our strategy to whole-blood data for which numerous ChIP-seq experiments have been performed (e.g. in blood-derived cell-lines) for several transcription factors to obtain transcription factor binding sites (TFBS). However, only relatively few biological contexts (e.g. tissues) have been profiled for genome-wide TFBS for a larger amount of transcription factors which could limit our analysis strategy to much fewer applications. As a proof of concept for the adaptability of our framework, we showed an example in GTEx Skeletal Muscle tissue. We predicted TFBS from ENCODE muscle cell-line DNAse-seq data [60, 257] using a deep learning based model [258], and used the obtained TFBS to reconstruct the underlying network for a single genetic *trans* hotspot. These novel developments of deep learning in genomic contexts, especially for sequence and signal based prediction tasks as in this case, have the potential to allow scaling of genomic analyses to more contexts, potentially simplifying interpretation and downstream analyses of results.

6.4. Prior based network inference yields novel insights into disease loci

We systematically applied our Bayesian inference framework to hundreds of *trans* expression and methylation QTL hotspots using human population-scale multi-omics data, advancing previous studies which applied network inference only for model systems [28, 32] or did not curate comprehensive priors for systematic network inference [30, 31, 124]. Detailed investigation of selected hotspots generated novel insights into two trait-associated genetic loci related to schizophrenia (SCZ) and lean body mass (LBM), for which we also recovered known trait-related genes. Overall, we showed that the genes identified in the networks are coherent with the phenotypes previously associated with the genetic loci, indicating the soundness of our strategy. Nevertheless, these results need to be interpreted carefully. For instance, the schizophrenia-related locus was derived from whole-blood data and direct interpretation in the SCZ context is therefore not straight forward as, for instance, gene expression assayed in whole-blood only

moderately correlates with gene expression in brain tissue [329]. Ideally, the analyses can be followed up in brain tissue to corroborate our current findings. Also, one needs to be cautious when trying to interpret the gene candidates in the schizophrenia HLA locus based on *cis*-eQTL alone, due to its difficult haplotype structure. *PBX2*, however, is included in the network independently of the *cis*-eQTL based on its connections to the *trans*-eQTL genes. Our integrated network analysis involving *cis* and *trans* entities prioritizes *PBX2*, which would have been impossible using *cis*-eQTL information alone, therefore expanding similar observations made previously in *trans*-eQTL studies [22].

For the LBM locus, we applied our framework to Skeletal Muscle data which, compared to the SCZ locus, provides a better tissue match for the observed phenotype. Here, too, we recovered known trait-associated genes, and although the functions and interactions of the identified *HDAC1*, *HDAC2*, and *SIN3B* genes have only been described in mice [316], our network indicates a similar functional relationship in humans. The observed interaction of *CREM* and *SYNC* in the network further led us to hypothesize a role of *CREM* in regulating muscle function and LBM, which has not been described before. However, in all cases, careful experimental validation of individual pathways should be executed before moving the purely computational findings to a more applied (e.g. clinical) context, for instance, by performing knockdowns of identified candidate genes and observing, whether and to which extend associated *trans* sites are affected.

6.5. Future perspective of single-cell data in systems biology

An interesting extension of our strategy can be in the context of single-cell data, which have been revolutionizing genomics research in recent years and open up promising new avenues for analyzing regulatory pathways in cellular systems. In contrast to the bulk data used in this thesis, which in essence measure mean levels of molecules over a population of cells, single-cell experiments identify omics levels from individual cells and hence achieve more detailed readouts. Importantly, with functional single-cell data, it is possible to investigate associations between variables under a more favorable statistical setup, where the number of samples N (cells) is the same or even more as the number of variables P (e.g. genes) of interest. Recent studies seek to make use of these favorable statistical properties. One example is the study by Aibar et al. [91], who proposed a framework (*SCENIC*) for single-cell based clustering and network inference, where the authors used common sub-networks to pinpoint stable cell states for individual cells. The authors implemented an extension to *GENIE3* [88], *GRNBoost*, specifically for application on single-cell data and which hence scales well for large datasets. Pliner et al. [330] used single-cell ATAC-seq data (an assay to probe DNA accessibility) with a graphical LASSO based approach (*CICERO*), for the identification of co-accessible genomic regions. Biologically, these regions reflect interacting regulatory elements distal to and at the promoter of target genes. They compared their co-accessible regions to physical DNA interactions derived from promoter-capture Hi-C [231], which showed strong concordance of the observed interactions. In addition, single-cell resolution

enables inference about the dynamic properties of cells based on a single data point. In their study, Ocone et al. [331] utilized this idea and were able to delineate regulatory networks for differentiated cells observed in their data.

Recent approaches allow assaying of multiple omic layers within the same cell, including scNMT-seq (single-cell nucleosome, methylation and transcription sequencing) [332], sciCAR [333], or scCAT-seq [334]. Using the scCAT-seq protocol, which can probe both gene expression and DNA accessibility in individual cells, Liu et al. [334], for example, inferred regulatory interactions between accessible chromatin regions and gene expression measured simultaneously in single cells.

In general, these new developments can improve our understanding of cell regulatory mechanisms by providing more detailed molecular insights and first studies show promising results [91, 330, 334, 335]. However, using data obtained from single cells poses its own challenges. For instance, these data typically have a low coverage per cell and therefore are prone to dropout events and exhibit differing noise properties compared to bulk data [336]. These challenges have to be overcome to make full use of single-cell technology in systems genetics, which necessitates further adaptation of methodological concepts (often designed for bulk data) to single-cell contexts.

6.6. Conclusions

This thesis provides new insights into the genome-wide regulation of epigenetic DNA methylation marks and gene expression through genetic variants.

We comprehensively characterized a novel set of genome-wide meQTL, derived from almost 7,000 individuals, providing new evidence for the functional relevance of *cis*-, *longrange*- and *trans*-acting variants influencing DNA methylation. By demonstrating enrichment of *longrange* and *trans* associated SNP-CpG pairs in topologically associating domains and inter-chromosomal contact regions, we provided evidence for a link between genome-wide promoter-enhancer interactions and meQTL in humans, for example via chromatin loops.

Our detailed network analyses of often trait-associated *trans* hotspots revealed the molecular actors underlying genome regulation and suggested potential mediators of disease mechanisms. So far, mostly direct links between *trans* associated entities have been investigated and a systematic evaluation of indirect connections to explain hotspots through regulatory pathways has been missing. We systematically established candidate genes and their regulatory interactions for numerous hotspots using a random walk based approach, thus providing new mechanistic insights into the genetic and epigenetic control of gene regulation and complex traits. For instance, we generated novel molecular hypotheses underlying the rheumatoid arthritis associated *NFKBIE* locus. Rheumatoid arthritis can be treated with IL-6 targeting drugs and we were the first to propose a molecular mechanism of how the genetic locus affects IL-6 regulation. Notably, formal colocalization of GWAS and QTL signals as well as partial wet-lab validation of our results indicate their reliability and generate confidence for follow up

studies.

We devised a fully integrative approach for de-novo network inference, involving QTL hotspots, multi-omics data, and biologically informed prior information, and were thus able to significantly expand our random walk based strategy, enabling the inference of new regulatory relationships which might have been missed before. The extensive benchmarking study of several state-of-the-art inference methods accentuated the benefit of using prior information for network inference and allowed us to formulate recommendations for network reconstruction at *trans*-acting loci. Application of our inference frameworks to diverse contexts yielded insights into gene regulatory mechanisms involving multiple regulatory steps. We highlighted two trait-associated hotspots related to schizophrenia and lean body mass, for which we could recover known trait-associated genes and generate new regulatory hypotheses. For instance, we generated a novel hypothesis about the involvement of *PBX2* in mediating the genome-wide effects of the schizophrenia-associated locus. The genetic locus is a *cis*-eQTL for *PBX2*, which further shows binding to *SPI1* and both genes have previously been associated with other neurological disorders, making *PBX2* an interesting candidate gene for this locus.

Our comprehensive meQTL resource is published as an easily accessible web interface. Together with our proposed strategy for explaining *trans* hotspots through regulatory networks, we expect that it will enable future genetic and epigenetic studies to leverage information about and facilitate the analysis of important globally acting genetic variants. Finally, by facilitating the detection of novel regulator genes involved in mediating the genome-wide effects of disease-related *trans* hotspots, we hope that this work will ultimately translate to clinical research and improve patient diagnostics and treatment for complex diseases.

Appendices

A. List of Figures

1.1. The first draft of the 'Central Dogma of Biology' as outlined by Francis Crick . . . 7
1.2. Schematic showing the two main marks of epigenetic gene regulation . . 9
1.3. Intuition behind (m)eQTL studies. . . . 13

2.1. Example plot showing the distribution of β values obtained from methylation arrays . . . 28
2.2. Number of samples in GTEx which have genotype and gene expression data available for the distinct tissues. . . . 34

3.1. Example of a Snakemake rule to annotate gene TSSs with transcription factor binding site information. . . . 53
3.2. Example 'rule graph' for the Snakemake pipeline when executing the network inference simulation study. . . . 54

4.1. Schematic of how SNPs could influence CpG methylation in *cis* , *longrange* and *trans* and how their combined effect can alter gene expression. 57
4.2. Analysis plan followed in the meQTL project . . . 60
4.3. Overview of all cosmopolitan methylation quantitative trait loci. . . . 75
4.4. Histograms showing the number of associated sentinel CpGs for all sentinel SNPs in the three meQTL categories *cis*, *longrange* and *trans*. . . . 76
4.5. Comparison of effect sizes for meQTL derived in isolated white cell subsets and the ones derived from whole-blood in our study. . . . 77
4.6. Summary of cross tissue replication of cosmopolitan meQTL. . . . 78
4.7. Histograms showing the amount of meQTL and background SNP-CpG pairs residing in the same regulatory class (enhancer or promoter). . . . 80
4.8. Enrichment of *longrange*- and *trans*-meQTL pairs in HiC data. . . . 81
4.9. Enrichment of meQTL SNPs and CpGs for association with gene expression. 82
4.10. Histograms showing the functional enrichment of *trans* regulators for meQTL. . . . 83
4.11. Schematic illustrating the random walk approach. . . . 84
4.12. Two example networks inferred using the two-step random walk approach on *trans* meQTL hotspots . . . 87
4.13. Relationship between distance of sentinel and EPIC array specific CpGs and their correlation. . . . 88

5.1. Illustration of the concept of partial correlations. . . . 104

5.2. Analysis plan followed in the Bayesian network inference project 113
5.3. Overview of all collected hotspots in the network inference study. 115
5.4. Overview of collected entities per hotspot. 116
5.5. Overview of the priors collected in the network inference study. 117
5.6. Results of the sample size simulation. 118
5.7. Results of the prior based simulation. 119
5.8. SNP-Gene recovery performance in the simulation. 120
5.9. Cross cohort replication performance of network inference methods. . . . 120
5.10. Transcription factor activities versus expression based cohort replication. 121
5.11. Network comparison for the *rs730775* locus. 123
5.12. The two highlight networks inferred in the network inference studies. . . 125
5.13. Colocalization analysis results. 125

C.1. Screenshot of the QTLdb application. 153
C.2. Candidate genes identified for sentinel SNPs in the meQTL study. 154
C.3. Results of the sensitivity analysis for the TFBS enrichment analysis. . . . 155
C.4. Observed and expected proportions of CpG sites that overlap *ZNF333* DNA binding sites, ... 156

B. List of Tables

1.1. ChromHMM states obtained from ChromHMM using the 15 state model model. . . . 10
1.2. Overview on selected resources for molecular interactions and omics datasets. . . . 18

2.1. Overview over the cohorts used in the meQTL study . . . 30
2.2. List of filters applied to the ReMap TFBS data set to obtain only blood related cell-type experiments. . . . 36

4.1. Example for a contingency table derived during the TAD/HiC analysis. . 64
4.2. Results from the comparison of TFBS enrichment in 450k vs EPIC data. . 89
4.3. Table comparing *cis*-meQTL chromHMM analyses between 450k and EPIC arrays. . . . 90

5.1. List of inference methods and their respective implementation used . . . 107
5.2. Comparison of inferred networks from the priors based network inference and the meQTL studies . . . 122

C.1. Sentinels and their annotated *cis* genes removed from analysis due to the genes not being measured on the microarrays. . . . 149
C.2. Individual covariates (technical and biological) used in models for testing associations between SNPs and CpGs. . . . 150
C.3. Table showing results of prior noise based simulation study. . . . 151
C.4. Table showing results of sample size based simulation study. . . . 151
C.5. Table showing mean MCC for cross cohort replication in the network inference study. . . . 152

C. Supplementary Information

A. Data used for meQTL replication

The text in this section was adapted from Hawe, Wilson, Loh, et al. [1] and was provided by our collaboration partners from the meQTL project (Chapter 4). Here, we describe the experimental details for the replication of meQTL in isolated cells and adipose tissue.

A.1. Isolated white blood cell studies

White cell subset samples were collected from 60 individuals comprising 30 obese and 30 normal weight people (Body-Mass-Index $BMI > 35\frac{kg}{m^2}$ and $BMI < 25\frac{kg}{m^2}$, respectively) 12ml whole blood (EDTA) were collected for each subject and samples processed immediately to isolate white blood cell subsets, including monocytes, neutrophils, CD4 and CD8 lymphocytes, via red blood cell lysis in line with the manufacturer's instructions (BioLegend). Staining was performed accordingly (>20min in 50μl; Ca2+-free PBS with 5mM EDTA and 1% human albumin; containing 1μl anti-CD14 PE-Cy7 (Clone-M5E2, BD), anti-CD16 BV510 (Clone-3G8, BioLegend), anti-CD45 BV605 (Clone-HI30, BioLegend), anti-CD8 APC (Clone-SK1, BioLegend); 2μl anti-CD3 PE (Clone-Leu-4, BD), anti-CD4 FITC (Clone-RPA-T4, BioLegend)).

After initial staining, clumped cells (30μm mesh, Miltenyi Biotec) were removed and additional staining added for dead cells (μl Sytox Blue, Life Technologies) [337, 338]. Stained and lysed samples were then sorted using a FACSAria II SORP cell sorter and a flow rate of 6,000–9,000 events/second. Raw data were retrieved with FACSDiva 8 and analysed using FlowJo v10. Controls without the primary labelled antibody of interest (i.e. fluorescence minus one negative controls) were utilized to estimate positive and negative boundaries for each gate. To ensure alignment and parametrization of the cell sorter, we ran Daily Cytometer Set-up and Tracking quality control beads (Anti-Mouse IgK and Negative Control, BSA; Compensation Plus Particles, BD). Live cells were defined based on Sytox Blue (450/50V nm) negative events and FCS-A and SSC-A then employed to separate granulocytes from monocyte and lymphocyte cells. CD14- and CD16+ neutrophils were then separated from other granulocytes. We separated monocytes from lymphocytes in a two-step process as CD14+, CD45+ and CD16- cells. Lastly, CD4+ and CD8+ cells were separated from other lymphocytes based on staining where we defined CD4+ cells: CD3+, CD4+, CD8-, CD14- and CD45+; and CD8+ cells: CD3+, CD4-, CD8+, CD14- and CD45+. The sorted cell subsets were assessed for purity, pelleted and snap-frozen for storage at $-80\,^{\circ}$C and subsequently average purities assessed, yielding values for neutrophils 98.3% ($\pm$1.2% (s.d.)); monocytes 99.2%

(±0.7%); CD4+ lymphocytes 99.6% (±0.4%); CD8+ lymphocytes 97.9% (±2.0%). Data to be used in the genome-wide association analysis were then generated for 57 out of the 60 collected samples. For this, genomic DNA was first isolated (Qiagen QIAshredder; Allprep DNA/RNA Micro) and then quantified using the Qubit double-stranded DNA broad range assay.

A.2. Isolated adipocyte studies

We obtained samples from subcutaneous and visceral adipose tissue intraoperatively in 24 healthy controls ($BMI < 30\frac{kg}{m^2}$) undergoing non-bariatric laparoscopic abdominal surgery and in 24 morbidly obese individuals ($BMI > 40\frac{kg}{m^2}$) undergoing laparoscopic bariatric surgery.

Primary human adipocyte cell populations were immediately isolated from adipose samples using previously described protocols [339]. To minimize adipocyte cell lysis, polypropylene plastic ware was utilized and tissue samples minced into 1-2mm^3 pieces, washed in Hank's buffered salt solution (HBSS) and then digested in a 37 °C water bath shaking at 100 rpm for about 45min using type 1 collagenase ($1mgml^{-1}$, Worthington). Samples were subsequently filtered to remove debris through a 300 μm nylon mesh and centrifuged at 500g and 4 °C for 5 minutes, leaving four distinct layers: oil, mature adipocytes, supernatant and stromovascular pellet. The oil layer was removed and the mature adipocyte layer gathered by pipette, then washed in HBSS (5x volume) and again centrifuged. The adipocyte cell suspension was collected after three washes for snap-freezing and storage (−80 °C). Genomic DNA and RNA were extracted from isolated adipocytes using Qiagen's AllPrep DNA/RNA/miRNA Universal Kit in line with the manufacturer's protocol proposed for lipid-rich samples.

A.3. DNA methylation in adipose tissue

We collected 603 adipose tissue samples from the MuTHER study for further replication. MuTHER contains 856 samples from female individuals of European descent recruited from the TwinsUK Adult Twin Registry. From a relatively photo-protected area inferior and adjacent to the umbilicus, punch biopsies (8mm) were taken and from each biopsy, subcutaneous adipose tissue dissected, weighed and split into multiple pieces. Samples were stored immediately in liquid nitrogen. Genomic DNA was extracted from the collected adipose tissue and DNA methylation profiled using the Illumina Infinium HumanMethylation450 BeadChIP as described previously [208]. Bisulfite conversion using the EZ-96 DNA Methylation Kit (Zymo Research) was performed with 700ng DNA, in line with the manufacturer's protocol.

B. Experimental validation of the ZNF333 *trans* locus

To validate our hypotheses for the *ZNF333* related *trans*-meQTL locus our collaboration partners experimentally identified 1) DNA binding sites of the ZNF333 protein and 2) protein binding partners of ZNF333. To this end, chromatin immunoprecipitation followed by sequencing (ChIP-seq) and immunoprecipitation mass spectrometry (IP-MS) was performed separately for cells in which *ZNF333* expression was induced. The below sections describe the experimental steps undertaken by our experimental partners to generate these data including initial data processing and was adapted from [1].

B.1. ChIP-seq experiment to determine *ZNF333* binding sites

We used chromatin immunoprecipitation followed by sequencing (ChIP-seq) to investigate binding of *ZNF333* to the respective DNA methylation sites of the *trans*-acting locus. We purchased plasmids overexpressing dual-tagged (Myc and FLAG) human *ZNF333* transcript (RC216457) from OriGene Technologies. Plasmids (ZNF333 and control GFP) were transfected into HCT116 cells using JetPrime transfection reagen (Polyplus) corresponding to the manufacturer's instructions in 15-cm tissue culture dishes. After 24 hours culture media was refreshed and cells cultured for another 24 hours. After 48 hours, cell lysates were used to perform chromatin immunoprecipitation. Western blot for Myc and FLAG antibodies was used to confirm *ZNF333* expression. For each of the two tags (Mcy, FLAG) we performed two replicates in the ChIP experiment and additional IP control experiments, executed for GFP transfected samples incubated with the separate tags. In addition to these controls, we performed two input control experiments for *ZNF333* and GFP transfected cells. Transfected HCT116 cells were prepared for ChIP-seq by cross-linking using 1% formaldehyde at room temperature for 10 minutes and subsequent quenching with 0.125 M glycine for another 5 minutes. Next, cells were washed (ice-cold PBS), scraped and pelleted down at 800g at 4 °C for 5 minutes. To facilitate the cell lysis process, the pellet was then resuspended in FA lysis buffer. We pelleted cell nuclei using centrifugation at 3,000 rpm at 4 °C for another 5 minutes and then lysed them in a 1% SDS lysis buffer (1% SDS, 1% Triton X-100, 2 mM EDTA, 50 mM HEPES-KOH (pH 7.5), 0.1% sodium dodecyl sulfate, Roche 1X Complete protease inhibitor). We next isolated the chromatin by applying ultracentrifugation at 20,000 rpm for 30 minutes in 4 °C. After resuspension (300 μl 0.1% SDS lysis buffer), chromatin was fragmented to a mean fragment size of 200-500 base pairs using sonication (Bioruptor Next gen, Diagenode). Immunoprecipitation for anti-Flag (Sigma, #F3165) and anti-Myc (Abcam, #ab9106) antibodies was done overnight for solubilized chromatin. We pulled down the respective antibody-chromatin complexes using Protein G Dynabeads (Invitrogen) and eluted the complexes after washing using a 1% SDS, 10 mM EDTA, 50 mM Tris-HCl (pH 8) elution buffer. The cross-linking was reversed and extract treated with Proteinase K. Using phenol-chloroform, we extracted precipitated DNA, applied ethanol precipitation and treated it with RNAse. Finally, obtained DNA was quantified using Qubit fluorometric quantification (Thermo Fisher Scientific) and

prepared for sequencing using New England Biolabs Ultra II Kit in accordance with the manufacturer's protocols. The Illumina NextSeq High platform was used to sequence the prepared library with 76bp single end reads.

B.2. Pull-down assay to identify ZNF333 binding partners

In addition to the ChIP-seq experiment described above, we employed IP-MS to identify binding partners of the *ZNF333* protein. To this end, we cultured HCT-116 cells using RPMI + 10% FBS medium at 37 °C and 5% CO2. We transfected a plasmid containing the Open Reading Frame (ORF) of ZNF333 locus (OHu29285, GenScript) into HCT-116 cells using Lipfectamine 2000 in T75 flasks as per manufacturer's recommendations. Nuclear and cytoplasmic extracts were subsequently gathered after 24-48 hours from transfected and untransfected cells (NE-PER extraction KIT from Thermo Scientific, in line with the manufacturer's protocols). Similar as for the ChIP-seq experiment above, we confirmed the over-expression of the ORF using Western blotting of the proteomic extracts and an anti-FLAG antibody (anti-DYKDDDDK tag, GenScript). Immunoprecipitation was performed for the nuclear protein fractions from transfected and untransfected cells using anti-FLAG mAb (A00187, GenScript) and anti-ZNF333 (HPA054680, Atlas Antibodies) antibodies, in addition to IgG2b mAb as an isotypic control (Monoclonal Antibody Core Facility, HMGU). After purification of protein complexes, these were subjected to label-free quantitative mass spectrometry (LC-MS/MS, data-dependent) using a QExactive HF mass spectrometer (Thermo Scientific) online coupled to an Ultimate 3000 nano-RSLC (Dionex, part of Thermo Scientific). We performed label-free proteome quantification in Progenesis QI for proteomics as previously suggested [340]. We used the Mascot search engine (Matrix Science) to query the generated MSMS spectra against the Swissprot human database (20235 sequences, Release 2017_02), setting an identification false discovery rate cut-off of 1%. The individual matches were then imported to the Progenesis QI Software and subsequently matched to the previous peptide quantifications. We used normalized protein abundances from the individual samples to calculate IP enrichment values as compared to the performed IgG control.

C. A public browser for quantitative trait loci

For the cosmopolitan meQTL associations identified in the meQTL project we published all results in an easily accessible and user-friendly web interface at `https://qtldb.helmholtz-muenchen.de` (tab 'Hawe et al. 2020 - whole blood'). The interface can be used to browse all meQTL, eQTL and eQTM associations replicated across cohorts. By applying filters, displayed association results can be restricted to specific entities (SNPs, genes or CpGs) and either the full list of significant associations or only the filtered subsets can be downloaded. Supplementary Figure C.1 provides a screenshot for the current version of the interface.

D. Workflow and code availability

The code for all projects was made available via GitHub. All scripts were written in R or standard Unix shell.

Code used to perform the random walk based network analysis can be found at `https://github.com/matthiasheinig/QTLnetwork`.

All analysis code for the network inference project was deposited under `https://github.com/jhawe/bggm`. This repository contains a dockerfile[1], which can be used to recreate the full software environment used in the project for full transparency and reproducibility. A pre-built Docker container for this dockerfile can be found at `https://hub.docker.com/repository/docker/jhawe/r3.5.2_custom`.

[1] `https://github.com/jhawe/bggm/blob/master/Dockerfile`

E. Supplementary Tables

	sentinel	chr	cis_gene	gene_start	gene_end	gene_strand	gene_biotype
1	rs10870226	chr10	TTC40	134621896	134756327	-	protein_coding
2	rs10870226	chr10	RP13-137A17.4	134757471	134778793	-	lincRNA
3	rs10870226	chr10	RP13-137A17.5	134774844	134775741	-	lincRNA
4	rs10870226	chr10	RP13-137A17.6	134779038	134789858	+	lincRNA
5	rs17420384	chr2	AC105393.1	388412	416885	+	lincRNA
6	rs17420384	chr2	AC105393.2	421057	422303	+	lincRNA
7	rs17420384	chr2	AC093326.1	490944	492655	-	lincRNA
8	rs17420384	chr2	AC093326.2	545805	546667	+	lincRNA
9	rs17420384	chr2	AC093326.3	558204	578145	+	lincRNA
10	rs2295981	chr13	LINC00354	112554299	112555490	+	lincRNA
11	rs2295981	chr13	AL136302.1	112563079	112563148	-	miRNA
12	rs2685252	chr2	AC105393.1	388412	416885	+	lincRNA
13	rs2685252	chr2	AC105393.2	421057	422303	+	lincRNA
14	rs2685252	chr2	AC093326.1	490944	492655	-	lincRNA
15	rs2685252	chr2	AC093326.2	545805	546667	+	lincRNA
16	rs2685252	chr2	AC093326.3	558204	578145	+	lincRNA
17	rs57743634	chr5	SDHAP3	1568637	1594735	-	pseudogene
18	rs57743634	chr5	CTD-2012J19.3	1594741	1611582	+	lincRNA
19	rs57743634	chr5	CTD-2012J19.2	1598242	1598362	+	pseudogene
20	rs57743634	chr5	RP11-43F13.1	1599035	1634120	-	pseudogene
21	rs57743634	chr5	CTD-2012J19.1	1614951	1616449	+	pseudogene
22	rs57743634	chr5	MIR4277	1708900	1708983	-	miRNA
23	rs57743634	chr5	CTD-2587M23.1	1725264	1728287	+	lincRNA

Supplementary Table C.1.: Sentinels and their annotated *cis* genes removed from analysis due to the genes not being measured on the microarrays. Sentinels rs1570038 and rs7924137 did not have any *cis* genes annotated. Table taken from [2].

KORA F4 (discovery)	KORA F4 (European replication)	KORA F4 (cross-ethnic replication)	KORA F3	NFBC66	NFBC86	SYS	LOLIPOP discovery	LOLIPOP replication
age, sex, BMI, white blood cell count;	age, sex, BMI, white blood cell count, Houseman imputed WBC subsets, methylation plate;	age, sex, white blood cell count, Houseman imputed WBC subsets, first 10 principal components of the control probes	Houseman imputed WBC subsets, first 10 principal components of the control probes	sex, CD8T, CD4T, NK, Bcell, Mono, Gran, first 30 control probe PCs, PC1-3 of genetic data	sex, CD8T, CD4T, NK, Bcell, Mono, Gran, first 30 control probe PCs, PC1-3 of genetic data	batch effect and blood cell type proportions.	age, sex, Houseman imputed WBC subsets, first 20 control probe PCs	age, sex, Houseman imputed WBC subsets, first 20 control probe PCs

Supplementary Table C.2.: Individual covariates (technical and biological) used in models for testing associations between SNPs and CpGs.

method	R=0	R=0.1	R=0.2	R=0.3	R=0.4	R=0.5	R=0.6	R=0.7	R=0.8	R=0.9	R=1	R=rbinom
bdgraph (priors)	**0.93**	**0.91**	**0.87**	0.83	0.80	0.77	0.72	0.69	0.64	0.60	0.55	**0.83**
glasso (priors)	0.87	0.81	0.74	0.66	0.60	0.53	0.46	0.41	0.34	0.27	0.21	0.42
bdgraph (empty)	0.84	0.84	0.84	**0.84**	**0.84**	**0.84**	**0.84**	**0.85**	**0.85**	**0.85**	**0.85**	**0.83**
bdgraph (full)	0.84	0.84	0.84	**0.84**	**0.84**	**0.84**	**0.84**	**0.85**	**0.84**	**0.84**	**0.85**	**0.83**
genenet	0.65	0.66	0.65	0.65	0.65	0.66	0.66	0.67	0.67	0.67	0.67	0.65
irafnet	0.45	0.43	0.42	0.39	0.37	0.36	0.34	0.31	0.28	0.24	0.20	0.42
glasso	0.43	0.43	0.43	0.43	0.43	0.43	0.44	0.44	0.44	0.45	0.46	0.42
genie3	0.38	0.37	0.37	0.37	0.37	0.38	0.38	0.39	0.39	0.38	0.40	0.35

Supplementary Table C.3.: Table giving an overview on the performance (mean MCC) in the simulation study for each inference method for all prior noise scenarios, sorted by first column. Highest mean MCC for each scenario is indicated in bold. Table taken from Hawe, Saha, Waldenberger, et al. [2].

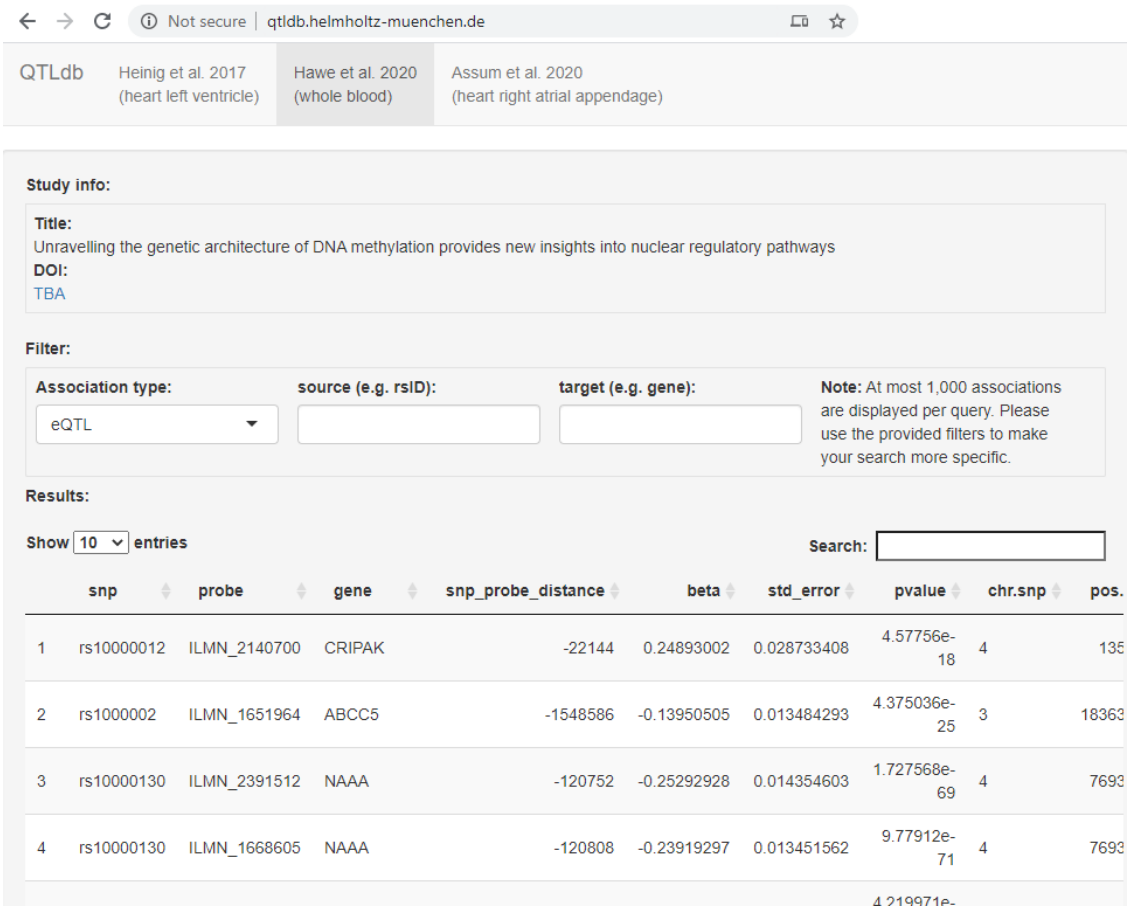
method	N= 50	100	150	200	250	300	350	400	450	500	550	600
bdgraph (priors)	**0.86**	**0.86**	**0.87**	**0.89**	**0.90**	**0.91**	**0.92**	**0.92**	**0.92**	**0.93**	**0.93**	**0.93**
glasso (priors)	0.83	0.85	0.86	0.86	0.87	0.87	0.87	0.87	0.88	0.88	0.88	0.88
bdgraph (empty)	0.43	0.57	0.64	0.69	0.74	0.76	0.78	0.80	0.81	0.82	0.83	0.84
bdgraph (full)	0.42	0.56	0.64	0.69	0.74	0.76	0.78	0.80	0.81	0.82	0.83	0.84
irafnet	0.32	0.38	0.41	0.42	0.43	0.44	0.44	0.44	0.45	0.45	0.45	0.45
genenet	0.29	0.43	0.50	0.54	0.57	0.60	0.61	0.63	0.63	0.65	0.65	0.66
genie3	0.26	0.30	0.32	0.33	0.34	0.34	0.35	0.35	0.36	0.36	0.36	0.36
glasso	0.20	0.27	0.31	0.34	0.37	0.38	0.39	0.40	0.41	0.42	0.42	0.43

Supplementary Table C.4.: Table giving an overview on the performance (mean MCC) in the simulation study for each inference method for different sub samplings of simulated data (increasing from left to right), sorted by first column. Highest mean MCC for each scenario is indicated in bold. Table taken from Hawe, Saha, Waldenberger, et al. [2].

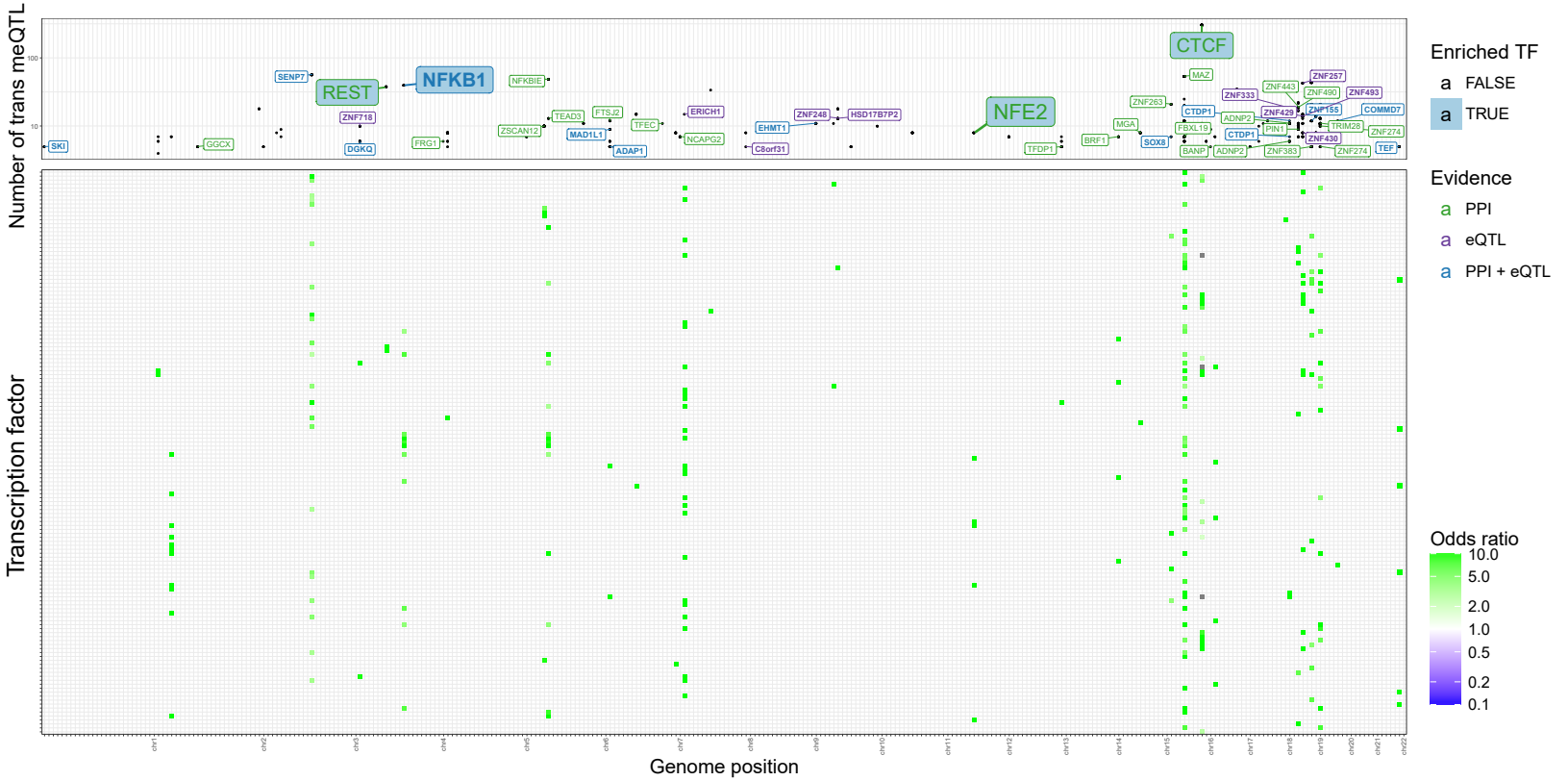
	method	Expression	TF activities
1	glasso (priors)	0.74 (0.18)	**0.75 (0.18)**
2	bdgraph (priors)	0.46 (0.1)	**0.49 (0.1)**
3	irafnet	0.36 (0.23)	**0.43 (0.21)**
4	genenet	0.31 (0.12)	**0.33 (0.12)**
5	bdgraph (empty)	0.29 (0.11)	**0.3 (0.12)**
6	glasso	0.25 (0.21)	**0.28 (0.22)**
7	genie3	**0.2 (0.23)**	**0.2 (0.19)**

Supplementary Table C.5.: Table indicates the mean cross cohort replication MCC for analyses based on expression and TF activity over all methods. Numbers in parentheses indicate standard deviations. Bold numbers indicate the higher mean MCC per method (TF activities versus expression). Table taken from Hawe, Saha, Waldenberger, et al. [2].

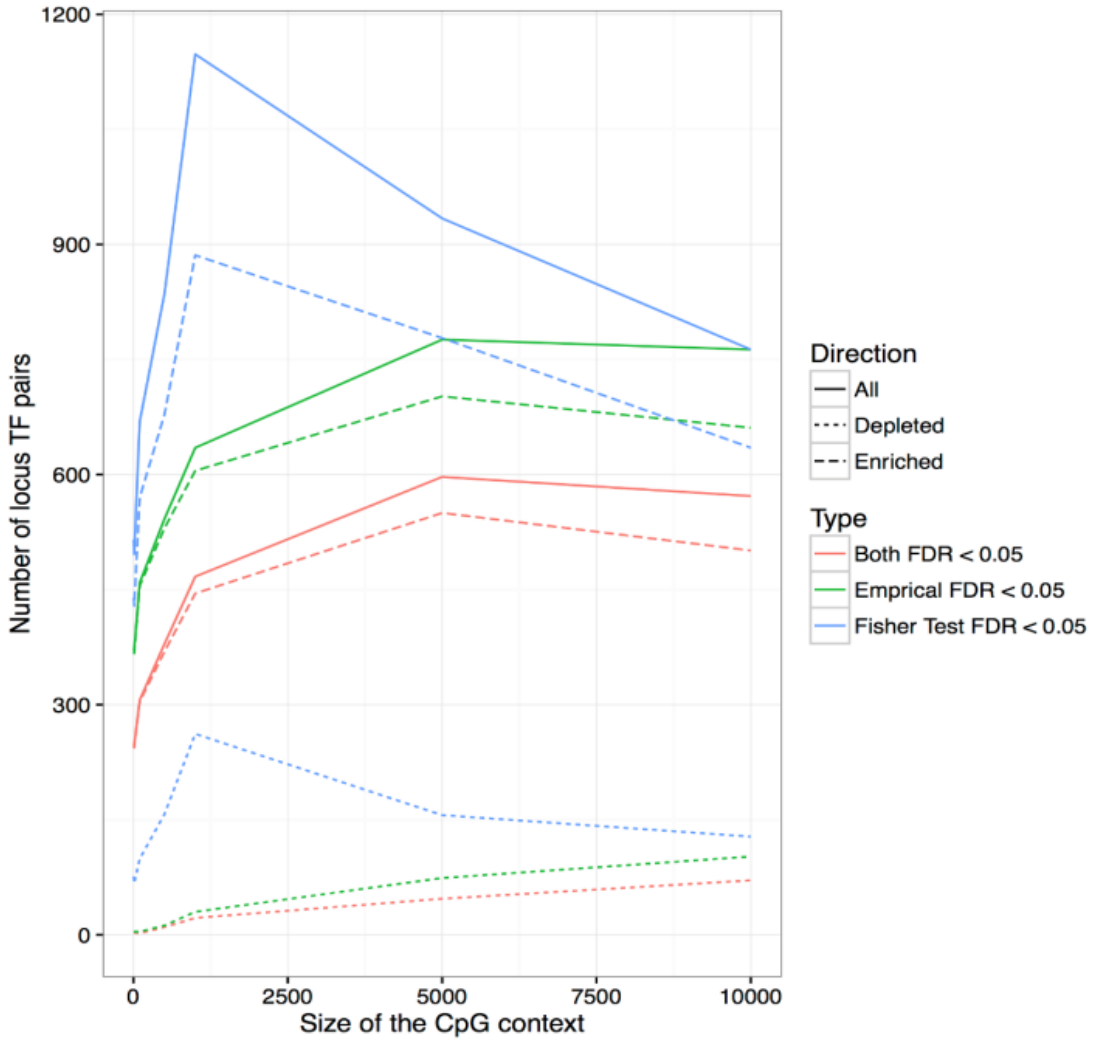
F. Supplementary Figures



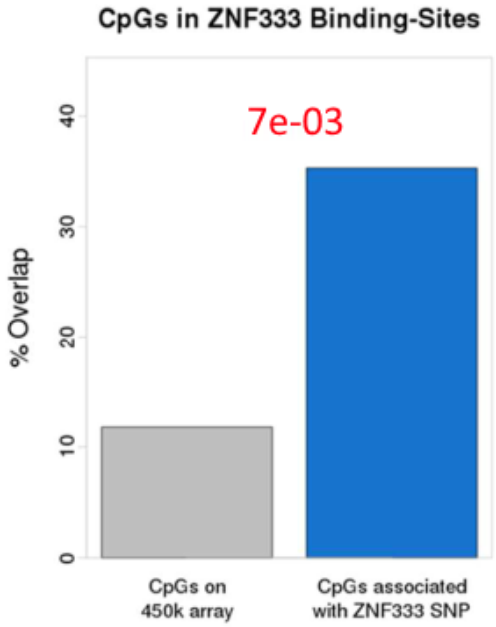
Supplementary Figure C.1.: Screenshot of the QTLdb application (`http://qtldb.helmholtz-muenchen.de/`) we built to provide the results from the meQTL project (currently selected) and other QTL projects.



Supplementary Figure C.2.: Candidate genes identified for sentinel SNPs in the meQTL study for which *trans* associated CpGs overlap with TFBS. Upper panel indicates observed TF enrichment for TFs encoded in *cis* with *trans* associated CpG sites, genes selected by the random walk analysis ('PPI') and genes that are *cis*-eQTL for individual sentinels. The lower panel shows a heatmap of enrichment or depletion of binding of TFs (y-axis) at the associated CpG sites of each sentinel (x-axis). Colors indicate Odds Ratios (ORs) comparing the overlap frequency at associated CpGs with background CpGs. ORs above 1 indicate enrichment, ORs below 1 indicate depletion. For improved readability, ORs greater than 10 or less than 0.1 have been set to 10 or 0.1, respectively. Figure adapted from Hawe, Wilson, Loh, et al. [1] by courtesy of Dr. Matthias Heinig.



Supplementary Figure C.3.: Sensitivity analysis for the transcription factor binding site (TFBS) enrichment analysis. X-axis shows interval sizes around the respective CpGs. For each interval, we test the 255 transcription factors for overlap with the *trans* CpG signatures of the 115 sentinel SNPs associated with ≥ 5 CpGs in *trans*, and present the total number of significant associations between a sentinel SNP and a TF (y-axis), where the TFBS overlaps the location of the *trans* associated CpGs more often than expected by chance. Significance of overlap between TFs and CpG signatures was determined using Fisher's exact test with all CpG sites on the array as background (blue), by re-sampling of CpG sites matched for mean and standard deviation of methylation levels (green), or both approaches at the same time (red). Figure adapted from Hawe, Wilson, Loh, et al. [1] by courtesy of Dr. Matthias Heinig.



Supplementary Figure C.4.: Proportions (observed and expected) of CpG sites that overlap *ZNF333* DNA binding regions. Expected proportion is estimated through permutation testing (see Methods), significance of enrichment is assessed using Fisher's exact test. Figure adapted from Hawe, Wilson, Loh, et al. [1] and provided by courtesy of Tan Lek Wen Wilson.

Bibliography

[1] J. S. Hawe, R. Wilson, M. Loh, K. Schmid, B. C. Lehne, B. Kühnel, Z. Li, W. R. Scott, M. Wielscher, I. Cebola, C. Baumbach, F. L. Tai, D. P. Lee, M. Kalra, E. Marouli, M. Bernard, L. Pfeiffer, P. Matías-García, M. I. Autio, S. Bourgeois, C. Herder, V. Karhunen, T. Meitinger, M. A. Pouladi, H. Prokisch, W. Rathmann, M. Roden, S. Sebert, J. Shin, B. M. Sim, K. Strauch, W. Zhang, W. L. Tan, S. M. Hauck, J. Merl-Pham, H. Grallert, X. Xu, R. Zeng, E. G. Barbosa, MuTHER Consortium, T. Illig, E. Reischl, A. Peters, T. Paus, Z. Pausova, P. Deloukas, J. Scott, J. Ferrer, R. S. Foo, M.-R. Jarvelin, J. S. Kooner, M. Heinig, C. Gieger, M. Waldenberger, and J. C. Chambers. "Genetic variation influencing DNA methylation provides new insights into the molecular pathways regulating genomic function." In: *Nature Genetics (manunscript under review)* (July 2020).

[2] J. S. Hawe, A. Saha, M. Waldenberger, S. Kunze, S. Wahl, M. Müller-Nurasyid, H. Prokisch, C. Herder, A. Peters, F. J. Theis, C. Gieger, J. Chambers, A. Battle, and M. Heinig. "Network reconstruction for trans acting genetic loci using multi-omics data and prior information". In: *bioRxiv* (May 2020), p. 2020.05.19.101592. DOI: 10.1101/2020.05.19.101592.

[3] J. S. Hawe, F. J. Theis, and M. Heinig. "Inferring Interaction Networks From Multi-Omics Data". In: *Frontiers in Genetics* 10 (June 2019), p. 535. DOI: 10.3389/fgene.2019.00535.

[4] F. Quagliarini, A. A. Mir, K. Balazs, M. Wierer, K. A. Dyar, C. Jouffe, K. Makris, J. Hawe, M. Heinig, F. V. Filipp, G. D. Barish, and N. H. Uhlenhaut. "Cistromic Reprogramming of the Diurnal Glucocorticoid Hormone Response by High-Fat Diet". In: *Molecular Cell* (Nov. 2019). DOI: 10.1016/j.molcel.2019.10.007.

[5] D. Altshuler, M. J. Daly, and E. S. Lander. "Genetic Mapping in Human Disease". In: *Science (New York, N.Y.)* 322.5903 (2008), p. 881. DOI: 10.1126/SCIENCE.1156409.

[6] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. "Five years of GWAS discovery." In: *American journal of human genetics* 90.1 (Jan. 2012), pp. 7–24. DOI: 10.1016/j.ajhg.2011.11.029.

[7] M. D. Gallagher and A. S. Chen-Plotkin. "The Post-GWAS Era: From Association to Function". In: *American Journal of Human Genetics* 102.5 (2018), p. 717. DOI: 10.1016/J.AJHG.2018.04.002.

[8] M. C. Mills and C. Rahal. "A scientometric review of genome-wide association studies". In: *Communications Biology* 2.1 (Dec. 2019), p. 9. DOI: 10.1038/s42003-018-0261-x.

[9] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. "Systematic localization of common disease-associated variation in regulatory DNA." In: *Science (New York, N.Y.)* 337.6099 (Sept. 2012), pp. 1190–5. DOI: 10.1126/science.1222794.

[10] Y. Hasin, M. Seldin, and A. Lusis. "Multi-omics approaches to disease". en. In: *Genome Biology* 18.1 (May 2017), p. 83.

[11] M. Civelek and A. J. Lusis. "Systems genetics approaches to understand complex traits." In: *Nature reviews. Genetics* 15.1 (2014), pp. 34–48. DOI: 10.1038/nrg3575. eprint: NIHMS150003.

[12] A. Bird. "Perceptions of epigenetics". In: *Nature* 447.7143 (2007), pp. 396–398. DOI: 10.1038/nature05913.

[13] J. Van Dongen, M. G. Nivard, G. Willemsen, J. J. Hottenga, Q. Helmer, C. V. Dolan, E. A. Ehli, G. E. Davies, M. Van Iterson, C. E. Breeze, S. Beck, H. E. Suchiman, R. Jansen, J. B. Van Meurs, B. T. Heijmans, P. E. Slagboom, and D. I. Boomsma. "Genetic and environmental influences interact with age and sex in shaping the human methylome". In: *Nature Communications* 7 (2016), pp. 1–13. DOI: 10.1038/ncomms11115.

[14] R. Joehanes, A. C. Just, R. E. Marioni, L. C. Pilling, L. M. Reynolds, P. R. Mandaviya, W. Guan, T. Xu, C. E. Elks, S. Aslibekyan, H. Moreno-Macias, J. A. Smith, J. A. Brody, R. Dhingra, P. Yousefi, J. S. Pankow, S. Kunze, S. H. Shah, A. F. McRae, K. Lohman, J. Sha, D. M. Absher, L. Ferrucci, W. Zhao, E. W. Demerath, J. Bressler, M. L. Grove, T. Huan, C. Liu, M. M. Mendelson, C. Yao, D. P. Kiel, A. Peters, R. Wang-Sattler, P. M. Visscher, N. R. Wray, J. M. Starr, J. Ding, C. J. Rodriguez, N. J. Wareham, M. R. Irvin, D. Zhi, M. Barrdahl, P. Vineis, S. Ambatipudi, A. G. Uitterlinden, A. Hofman, J. Schwartz, E. Colicino, L. Hou, P. S. Vokonas, D. G. Hernandez, A. B. Singleton, S. Bandinelli, S. T. Turner, E. B. Ware, A. K. Smith, T. Klengel, E. B. Binder, B. M. Psaty, K. D. Taylor, S. A. Gharib, B. R. Swenson, L. Liang, D. L. DeMeo, G. T. O'Connor, Z. Herceg, K. J. Ressler, K. N. Conneely, N. Sotoodehnia, S. L. R. Kardia, D. Melzer, A. A. Baccarelli, J. B. J. van Meurs, I. Romieu, D. K. Arnett, K. K. Ong, Y. Liu, M. Waldenberger, I. J. Deary, M. Fornage, D. Levy, and S. J. London. "Epigenetic Signatures of Cigarette Smoking". In: *Circulation: Cardiovascular Genetics* 9.5 (Oct. 2016), pp. 436–447. DOI: 10.1161/CIRCGENETICS.116.001506.

[15] É. Héberlé and A. F. Bardet. "Sensitivity of transcription factors to DNA methylation". In: *Essays in Biochemistry* 63.6 (2019), p. 727. DOI: 10.1042/EBC20190033.

[16] P. W. Laird. "Principles and challenges of genome-wide DNA methylation analysis". In: *Nature Reviews Genetics* 11.3 (Mar. 2010), p. 191. DOI: 10.1038/nrg2732.

[17] A. J. Lea, C. M. Vockley, R. A. Johnston, C. A. Del Carpio, L. B. Barreiro, T. E. Reddy, and J. Tung. "Genome-wide quantification of the effects of DNA methylation on human gene regulation". In: *eLife* 7 (Dec. 2018). DOI: 10.7554/eLife.37513.

[18] T. Huan, R. Joehanes, C. Song, F. Peng, Y. Guo, M. Mendelson, C. Yao, C. Liu, J. Ma, M. Richard, G. Agha, W. Guan, L. M. Almli, K. N. Conneely, J. Keefe, S. J. Hwang, A. D. Johnson, M. Fornage, L. Liang, and D. Levy. "Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease". In: *Nature Communications* 10.1 (2019), pp. 1–14. DOI: 10.1038/s41467-019-12228-z.

[19] F. W. Albert and L. Kruglyak. "The role of regulatory variation in complex traits and disease". In: *Nat. Rev. Genet.* 16.4 (Apr. 2015), pp. 197–212. DOI: 10.1038/nrg3891.

[20] Y. Gilad, S. A. Rifkin, and J. K. Pritchard. "Revealing the architecture of gene regulation: the promise of eQTL studies". In: *Trends in Genetics* 24.8 (Aug. 2008), pp. 408–415. DOI: 10.1016/J.TIG.2008.06.001.

[21] R. Joehanes, X. Zhang, T. Huan, C. Yao, S.-X. Ying, Q. T. Nguyen, C. Y. Demirkale, M. L. Feolo, N. R. Sharopova, A. Sturcke, A. A. Schäffer, N. Heard-Costa, H. Chen, P.-C. Liu, R. Wang, K. A. Woodhouse, K. Tanriverdi, J. E. Freedman, N. Raghavachari, J. Dupuis, A. D. Johnson, C. J. O'Donnell, D. Levy, and P. J. Munson. "Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies". en. In: *Genome Biology* 18 (2017).

[22] U. Võsa, A. Claringbould, H.-j. Westra, M. J. Bonder, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Kasela, I. Alvaes, M.-j. Fave, M. Agbessi, M. Christiansen, J. Verlouw, H. Yaghootkar, R. Sönmez, A. Brown, V. Kukushkina, A. Kalnapenkis, S. Rüeger, E. Porcu, J. Kronberg-, J. Kettunen, J. Powell, B. Lee, F. Zhang, F. Beutner, B. Consortium, H. Brugge, M. Kähönen, Y. Kim, J. C. Knight, P. Kovacs, K. Krohn, O. Stegle, A. Battle, J. Yang, P. M. Visscher, and M. Scholz. "Unraveling the polygenic architecture of complex traits using blood eQTL meta- analysis". In: (2018), pp. 1–57. DOI: 10.1101/447367.

[23] M. J. Bonder, R. Luijk, D. V. Zhernakova, M. Moed, P. Deelen, M. Vermaat, M. van Iterson, F. van Dijk, M. van Galen, J. Bot, R. C. Slieker, P. M. Jhamai, M. Verbiest, H. E. D. Suchiman, M. Verkerk, R. van der Breggen, J. van Rooij, N. Lakenberg, W. Arindrarto, S. M. Kielbasa, I. Jonkers, P. van 't Hof, I. Nooren, M. Beekman, J. Deelen, D. van Heemst, A. Zhernakova, E. F. Tigchelaar, M. A.

Swertz, A. Hofman, A. G. Uitterlinden, R. Pool, J. van Dongen, J. J. Hottenga, C. D. A. Stehouwer, C. J. H. van der Kallen, C. G. Schalkwijk, L. H. van den Berg, E. W. van Zwet, H. Mei, Y. Li, M. Lemire, T. J. Hudson, P. E. Slagboom, C. Wijmenga, J. H. Veldink, M. M. J. van Greevenbroek, C. M. van Duijn, D. I. Boomsma, A. Isaacs, R. Jansen, J. B. J. van Meurs, P. A. C. 't Hoen, L. Franke, B. T. Heijmans, and B. T. Heijmans. "Disease variants alter transcription factor levels and methylation of their binding sites". In: *Nature Genetics* 49.1 (Jan. 2017), pp. 131–138. DOI: 10.1038/ng.3721.

[24] L. T. Husquin, M. Rotival, M. Fagny, H. Quach, N. Zidane, L. M. McEwen, J. L. MacIsaac, M. S. Kobor, H. Aschard, E. Patin, and L. Quintana-Murci. "Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation". In: *Genome Biology* 19.1 (Dec. 2018), p. 222. DOI: 10.1186/s13059-018-1601-3.

[25] M. Heinig, E. Petretto, C. Wallace, L. Bottolo, M. Rotival, H. Lu, Y. Li, R. Sarwar, S. R. Langley, A. Bauerfeind, O. Hummel, Y.-A. Lee, S. Paskas, C. Rintisch, K. Saar, J. Cooper, R. Buchan, E. E. Gray, J. G. Cyster, C. Consortium, J. Erdmann, C. Hengstenberg, S. Maouche, W. H. Ouwehand, C. M. Rice, N. J. Samani, H. Schunkert, A. H. Goodall, H. Schulz, H. Roider, M. Vingron, S. Blankenberg, T. Münzel, T. Zeller, S. Szymczak, A. Ziegler, L. Tiret, D. J. Smyth, M. Pravenec, T. J. Aitman, F. Cambien, D. Clayton, J. A. Todd, N. Hubner, and S. A. Cook. "A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk". In: *Nature* 467.7314 (2010), p. 460. DOI: 10.1038/NATURE09386.

[26] H.-J. Westra, M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, J. E. Powell, A. Zhernakova, D. V. Zhernakova, J. H. Veldink, L. H. Van den Berg, J. Karjalainen, S. Withoff, A. G. Uitterlinden, A. Hofman, F. Rivadeneira, P. A. C. 't Hoen, E. Reinmaa, K. Fischer, M. Nelis, L. Milani, D. Melzer, L. Ferrucci, A. B. Singleton, D. G. Hernandez, M. A. Nalls, G. Homuth, M. Nauck, D. Radke, U. Völker, M. Perola, V. Salomaa, J. Brody, A. Suchy-Dicey, S. A. Gharib, D. A. Enquobahrie, T. Lumley, G. W. Montgomery, S. Makino, H. Prokisch, C. Herder, M. Roden, H. Grallert, T. Meitinger, K. Strauch, Y. Li, R. C. Jansen, P. M. Visscher, J. C. Knight, B. M. Psaty, S. Ripatti, A. Teumer, T. M. Frayling, A. Metspalu, J. B. J. van Meurs, and L. Franke. "Systematic identification of trans eQTLs as putative drivers of known disease associations". In: *Nature Genetics* 45.10 (Sept. 2013), pp. 1238–1243. DOI: 10.1038/ng.2756.

[27] X. Liu, Y. I. Li, and J. K. Pritchard. "Trans Effects on Gene Expression Can Drive Omnigenic Inheritance". In: *Cell* 177.4 (2019), 1022–1034.e6. DOI: 10.1016/j.cell.2019.04.014.

[28] J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt. "Integrating large-scale functional genomic data to dissect

the complexity of yeast regulatory networks". In: *Nature Genetics* 40.7 (2008), pp. 854–861. DOI: 10.1038/ng.167.

[29] K. Suhre, M. Arnold, A. M. Bhagwat, R. J. Cotton, R. Engelke, J. Raffler, H. Sarwath, G. Thareja, A. Wahl, R. K. DeLisle, L. Gold, M. Pezer, G. Lauc, M. A. El-Din Selim, D. O. Mook-Kanamori, E. K. Al-Dous, Y. A. Mohamoud, J. Malek, K. Strauch, H. Grallert, A. Peters, G. Kastenmüller, C. Gieger, and J. Graumann. "Connecting genetic risk to disease end points through the human blood plasma proteome". In: *Nature Communications* 8 (Feb. 2017), p. 14357. DOI: 10.1038/ncomms14357.

[30] Z. Wang, W. Xu, F. A. S. Lucas, and Y. Liu. "Incorporating prior knowledge into Gene Network Study". In: *Bioinformatics* 29.20 (2013), p. 26332640. DOI: 10.1093/bioinformatics/btt443.

[31] Y. Li and S. A. Jackson. "Gene Network Reconstruction by Integration of Prior Biological Knowledge". In: *G3 (Bethesda)* 5.6 (Mar. 2015), pp. 1075–1079.

[32] F. Petralia, P. Wang, J. Yang, and Z. Tu. "Integrative random forest for gene regulatory network inference". In: *Bioinformatics* 31.12 (June 2015), pp. i197–i205. DOI: 10.1093/bioinformatics/btv268.

[33] D. V. Manatakis, V. K. Raghu, and P. V. Benos. "piMGM: incorporating multi-source priors in mixed graphical models for learning disease networks". In: *Bioinformatics* 34.17 (Sept. 2018), pp. i848–i856. DOI: 10.1093/bioinformatics/bty591.

[34] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim. "Methods of integrating data to uncover genotype–phenotype interactions". In: *Nature Reviews Genetics* 16.2 (Feb. 2015), pp. 85–97. DOI: 10.1038/nrg3868.

[35] F. H. Crick. "On protein synthesis". In: *Symp. Soc. Exp. Biol.* 12 (1958), pp. 138–163.

[36] J. D. Watson. *Molecular biology of the gene*. New York 10016 (One Park Avenue): W. A. Benjamin, Inc., 1965.

[37] D. Baltimore. "Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses". In: *Nature* 226.5252 (1970), pp. 1209–1211.

[38] W. Reik. "Stability and flexibility of epigenetic gene regulation in mammalian development". In: *Nature* 447.7143 (May 2007), pp. 425–432. DOI: 10.1038/nature05918.

[39] Y.-H. Zhang, Y. Hu, Y. Zhang, L.-D. Hu, and X. Kong. "Distinguishing three subtypes of hematopoietic cells based on gene expression profiles using a support vector machine". In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1864.6 (June 2018), pp. 2255–2265. DOI: 10.1016/J.BBADIS.2017.12.003.

[40] G. J. Hannon. "RNA interference". In: *Nature* 418.6894 (July 2002), pp. 244–251. DOI: 10.1038/418244a.

[41] S. Saurabh, A. S. Vidyarthi, and D. Prasad. "RNA interference: concept to reality in crop improvement." In: *Planta* 239.3 (Mar. 2014), pp. 543–64. DOI: 10.1007/s00425-013-2019-5.

[42] B. Roy, L. M. Haupt, and L. R. Griffiths. "Review: Alternative Splicing (AS) of Genes As An Approach for Generating Protein Complexity." In: *Current genomics* 14.3 (May 2013), pp. 182–94. DOI: 10.2174/1389202911314030004.

[43] C. T. Walsh, S. Garneau-Tsodikova, and G. J. Gatto. "Protein posttranslational modifications: The chemistry of proteome diversifications". In: *Angewandte Chemie - International Edition* 44.45 (2005), pp. 7342–7372. DOI: 10.1002/anie.200501023.

[44] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. "High-Resolution Profiling of Histone Methylations in the Human Genome". In: *Cell* 129.4 (May 2007), pp. 823–837. DOI: 10.1016/J.CELL.2007.05.009.

[45] S. L. Klemm, Z. Shipony, and W. J. Greenleaf. "Chromatin accessibility and the regulatory epigenome". In: *Nature Reviews Genetics* 20.4 (Apr. 2019), pp. 207–220. DOI: 10.1038/s41576-018-0089-8.

[46] C. D. Allis and T. Jenuwein. "The molecular hallmarks of epigenetic control". In: *Nature Reviews Genetics* 17.8 (Aug. 2016), pp. 487–500. DOI: 10.1038/nrg.2016.59.

[47] T. Jenuwein and C. D. Allis. "Translating the histone code". In: *Science* 293.5532 (2001), pp. 1074–1080. DOI: 10.1126/science.1063127.

[48] Roadmap Epigenomics Consortium. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (2015), pp. 317–330. DOI: 10.1038/nature14248.

[49] J. Ernst and M. Kellis. "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature Methods* 9.3 (Feb. 2012), pp. 215–216. DOI: 10.1038/nmeth.1906.

[50] J. C. Chambers, M. Loh, B. Lehne, A. Drong, J. Kriebel, V. Motta, S. Wahl, H. R. Elliott, F. Rota, W. R. Scott, W. Zhang, S.-T. Tan, G. Campanella, M. Chadeau-Hyam, L. Yengo, R. C. Richmond, M. Adamowicz-Brice, U. Afzal, K. Bozaoglu, Z. Y. Mok, H. K. Ng, F. Pattou, H. Prokisch, M. A. Rozario, L. Tarantini, J. Abbott, M. Ala-Korpela, B. Albetti, O. Ammerpohl, P. A. Bertazzi, C. Blancher, R. Caiazzo, J. Danesh, T. R. Gaunt, S. de Lusignan, C. Gieger, T. Illig, S. Jha, S. Jones, J. Jowett, A. J. Kangas, A. Kasturiratne, N. Kato, N. Kotea, S. Kowlessur, J. Pitkäniemi, P. Punjabi, D. Saleheen, C. Schafmayer, P. Soininen, E.-S. Tai, B. Thorand, J. Tuomilehto, A. R. Wickremasinghe, S. A. Kyrtopoulos, T. J. Aitman, C. Herder, J. Hampe, S. Cauchi, C. L. Relton, P. Froguel, R. Soong, P. Vineis, M.-R. Jarvelin, J. Scott, H. Grallert, V. Bollati, P. Elliott, M. I. McCarthy, and J. S. Kooner. "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 & endocrinology* 3.7 (July 2015), pp. 526–534. DOI: 10.1016/S2213-8587(15)00127-8.

[51] R. E. Marioni, S. Shah, A. F. McRae, B. H. Chen, E. Colicino, S. E. Harris, J. Gibson, A. K. Henders, P. Redmond, S. R. Cox, A. Pattie, J. Corley, L. Murphy, N. G. Martin, G. W. Montgomery, A. P. Feinberg, M. D. Fallin, M. L. Multhaup, A. E. Jaffe, R. Joehanes, J. Schwartz, A. C. Just, K. L. Lunetta, J. M. Murabito, J. M. Starr, S. Horvath, A. A. Baccarelli, D. Levy, P. M. Visscher, N. R. Wray, and I. J. Deary. "DNA methylation age of blood predicts all-cause mortality in later life". In: *Genome Biology* 16.1 (2015), pp. 1–12. DOI: 10.1186/s13059-015-0584-6.

[52] Y. Zhang, R. Wilson, J. Heiss, L. P. Breitling, K.-U. Saum, B. Schöttker, B. Holleczek, M. Waldenberger, A. Peters, and H. Brenner. "DNA methylation signatures in peripheral blood strongly predict all-cause mortality". In: *Nature Communications* 8.1 (Apr. 2017), p. 14617. DOI: 10.1038/ncomms14617.

[53] P. van der Harst, L. J. de Windt, and J. C. Chambers. "Translational Perspective on Epigenetics in Cardiovascular Disease". In: *Journal of the American College of Cardiology* 70.5 (2017), pp. 590–606. DOI: 10.1016/j.jacc.2017.05.067.

[54] 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. DOI: 10.1038/nature20784.Epigenome-wide.

[55] S. Ambatipudi, C. Cuenin, H. Hernandez-Vargas, A. Ghantous, F. Le Calvez-Kelm, R. Kaaks, M. Barrdahl, H. Boeing, K. Aleksandrova, A. Trichopoulou, P. Lagiou, A. Naska, D. Palli, V. Krogh, S. Polidoro, R. Tumino, S. Panico, B. Bueno-de-Mesquita, P. H. Peeters, J. R. Quirós, C. Navarro, E. Ardanaz, M. Dorronsoro, T. Key, P. Vineis, N. Murphy, E. Riboli, I. Romieu, and Z. Herceg. "Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study". In: *Epigenomics* 8.5 (May 2016), pp. 599–618. DOI: 10.2217/epi-2016-0001.

[56] T. R. Gaunt, H. A. Shihab, G. Hemani, J. L. Min, G. Woodward, O. Lyttleton, J. Zheng, A. Duggirala, W. L. McArdle, K. Ho, S. M. Ring, D. M. Evans, G. Davey Smith, and C. L. Relton. "Systematic identification of genetic influences on methylation across the human life course". In: *Genome biology* 17.1 (Dec. 2016), p. 61. DOI: 10.1186/s13059-016-0926-z.

[57] T. G. Richardson, P. C. Haycock, J. Zheng, N. J. Timpson, T. R. Gaunt, G. Davey Smith, C. L. Relton, and G. Hemani. "Systematic Mendelian randomization framework elucidates hundreds of CpG sites which may mediate the influence of genetic variants on disease". In: *Human Molecular Genetics* 27.18 (Sept. 2018), pp. 3293–3304. DOI: 10.1093/hmg/ddy210.

[58] B. L. Tremblay, F. Guénard, B. Lamarche, L. Pérusse, and M. C. Vohl. "Familial resemblances in blood leukocyte DNA methylation levels". In: *Epigenetics* 11.11 (2016), pp. 831–838. DOI: 10.1080/15592294.2016.1232234.

[59] E. Hannon, T. J. Gorrie-Stone, M. C. Smart, J. Burrage, A. Hughes, Y. Bao, M. Kumari, L. C. Schalkwyk, and J. Mill. "Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits". In: *American Journal of Human Genetics* 103.5 (2018), pp. 654–665. DOI: 10.1016/j.ajhg.2018.09.007.

[60] The ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome". en. In: *Nature* 489.7414 (Sept. 2012), p. 57.

[61] D. Jjingo, A. B. Conley, S. V. Yi, V. V. Lunyak, and I. King Jordan. "On the presence and role of human gene-body DNA methylation". In: *Oncotarget* 3.4 (2012), pp. 462–474. DOI: 10.18632/oncotarget.497.

[62] M. P. Ball, J. B. Li, Y. Gao, J.-H. Lee, E. Leproust, I.-H. Park, B. Xie, G. Q. Daley, and G. M. Church. "Targeted and genome-scale methylomics reveals gene body signatures in human cell lines". In: *Nature Biotechnology* 27.4 (2009), pp. 361–368. DOI: 10.1038/nbt.1533.Targeted.

[63] G. Lev Maor, A. Yearim, and G. Ast. "The alternative role of DNA methylation in splicing regulation". In: *Trends in Genetics* 31.5 (2015), pp. 274–280. DOI: 10.1016/j.tig.2015.03.002.

[64] T. Ideker, T. Galitski, and L. Hood. "A new approach to decoding life : Systems Biology". In: *Annu. Rev. Genomics Hum. Genet.* (2001).

[65] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell. "HITS-CLIP yields genome-wide insights into brain alternative RNA processing". In: *Nature* 456.7221 (Nov. 2008), pp. 464–469.

[66] A. Brueckner, C. Polge, N. Lentze, D. Auerbach, and U. Schlattner. "Yeast Two-Hybrid, a Powerful Tool for Systems Biology". In: *International Journal of Molecular Sciences* 10.6 (2009), pp. 2763–2788. DOI: 10.3390/ijms10062763.

[67] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. "Genome-wide mapping of in vivo protein-DNA interactions". In: *Science* 316.5830 (June 2007), pp. 1497–1502.

[68] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo. "Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)". In: *Nat. Methods* 13.6 (June 2016), pp. 508–514. DOI: 10.1038/nmeth.3810.

[69] J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, and Y. Zhan. "Hi–C: A comprehensive technique to capture the conformation of genomes". In: *Methods* 58.3 (Nov. 2012), pp. 268–276. DOI: 10.1016/J.YMETH.2012.05.001.

[70] J. M. Engreitz, K. Sirokman, P. McDonel, A. A. Shishkin, C. Surka, P. Russell, S. R. Grossman, A. Y. Chow, M. Guttman, and E. S. Lander. "RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites." In: *Cell* 159.1 (Sept. 2014), pp. 188–199. DOI: 10.1016/j.cell.2014.08.018.

[71] J. Chèneby, M. Gheorghe, M. Artufel, A. Mathelier, and B. Ballester. "ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments". In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D267–D275. DOI: 10.1093/nar/gkx1092.

[72] "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." In: *Science (New York, N.Y.)* 326.5950 (Oct. 2009), pp. 289–93. DOI: 10.1126/science.1181369.

[73] "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C". In: *Nature Genetics* 47.6 (June 2015), pp. 598–606. DOI: 10.1038/ng.3286.

[74] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, N. Ahmadiyeh, M. Pomerantz, C. Grisanzio, P. Herman, L. Jia, V. Almendro, H. He, M. Brown, X. Liu, B. Abbas, A. Classen, B. van Steensel, B. Vogelstein, D. Pardoll, D. Coffey, X. Xie, T. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis, E. Lander, E. Yaffe, A. Tanay, B. Zehnbauer, and B. Vogelstein. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping". In: *Cell* 159.7 (Dec. 2014), pp. 1665–1680. DOI: 10.1016/j.cell.2014.11.021.

[75] B. M. Javierre, S. Sewitz, J. Cairns, S. W. Wingett, C. Vï¿rnai, M. J. Thiecke, P. Freire-Pritchett, M. Spivakov, P. Fraser, O. S. Burren, A. J. Cutler, J. A. Todd, C. Wallace, S. P. Wilder, R. Kreuzhuber, M. Kostadima, D. R. Zerbino, O. Stegle, R. Kreuzhuber, F. Burden, S. Farrow, K. Rehnstrï¿m, K. Downes, L. Grassi, M. Kostadima, W. H. Ouwehand, M. Frontini, R. Kreuzhuber, F. Burden, S. Farrow, K. Rehnstrï¿m, K. Downes, M. Kostadima, W. H. Ouwehand, M. Frontini, S. M. Hill, F. Wang, H. G. Stunnenberg, W. H. Ouwehand, M. Frontini, W. H. Ouwehand, J. H. Martens, B. Kim, N. Sharifi, E. M. Janssen-Megens, M. L. Yaspo, M. Linser, A. Kovacsovics, L. Clarke, D. Richardson, A. Datta, and P. Flicek. "Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters". In: *Cell* 167.5 (2016), 1369–1384.e19. DOI: 10.1016/j.cell.2016.09.037.

[76] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Y. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar, S. Alizadeh, S. Bahr, R. L. Brost, Y. Chen, M. Cokol, C. Nislow, O. G. Troyanskaya, H. Bussey, G. D. Bader, A.-C. Gingras, Q. D. Morris, P. M. Kim, C. A. Kaiser, C. L. Myers, B. J. Andrews, and C. Boone. "The genetic landscape of a cell." In: *Science* 327.5964 (Jan. 2010), pp. 425–31. DOI: 10.1126/science.1180823.

[77] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. "Detecting Protein Function and Protein-Protein Interactions from Genome Sequences". In: *Science* 285.5428 (1999), pp. 751–753. DOI: 10.1126/science.285.5428.751.

[78] S. De Bodt, G. Theissen, and Y. Van de Peer. "Promoter Analysis of MADS-Box Genes in Eudicots Through Phylogenetic Footprinting". In: *Molecular Biology and Evolution* 23.6 (June 2006), pp. 1293–1303. DOI: 10.1093/molbev/msk016.

[79] M. Heinig, M. E. Adriaens, S. Schafer, H. W. M. van Deutekom, E. M. Lodder, J. S. Ware, V. Schneider, L. E. Felkin, E. E. Creemers, B. Meder, H. A. Katus, F. Rühle, M. Stoll, F. Cambien, E. Villard, P. Charron, A. Varro, N. H. Bishopric, A. L. George, C. dos Remedios, A. Moreno-Moral, F. Pesce, A. Bauerfeind, F. Rüschendorf, C. Rintisch, E. Petretto, P. J. Barton, S. A. Cook, Y. M. Pinto, C. R. Bezzina, and N. Hubner. "Natural genetic variation of the cardiac transcriptome in non-diseased donors and patients with dilated cardiomyopathy". In: *Genome Biology* 18.1 (Dec. 2017), p. 170. DOI: 10.1186/s13059-017-1286-z.

[80] L. J. Scott, M. R. Erdos, J. R. Huyghe, R. P. Welch, A. T. Beck, B. N. Wolford, P. S. Chines, J. P. Didion, N. Narisu, H. M. Stringham, D. L. Taylor, A. U. Jackson, S. Vadlamudi, L. L. Bonnycastle, L. Kinnunen, J. Saramies, J. Sundvall, R. D. Albanus, A. Kiseleva, J. Hensley, G. E. Crawford, H. Jiang, X. Wen, R. M. Watanabe, T. A. Lakka, K. L. Mohlke, M. Laakso, J. Tuomilehto, H. A. Koistinen, M. Boehnke, F. S. Collins, and S. C. J. Parker. "The genetic regulatory signature of type 2 diabetes in human skeletal muscle". In: *Nature Communications* 7.1 (Sept. 2016), p. 11764. DOI: 10.1038/ncomms11764.

[81] K. Schramm, C. Marzi, C. Schurmann, M. Carstensen, E. Reinmaa, R. Biffar, G. Eckstein, C. Gieger, H.-J. Grabe, G. Homuth, G. Kastenmüller, R. Mägi, A. Metspalu, E. Mihailov, A. Peters, A. Petersmann, M. Roden, K. Strauch, K. Suhre, A. Teumer, U. Völker, H. Völzke, R. Wang-Sattler, M. Waldenberger, T. Meitinger, T. Illig, C. Herder, H. Grallert, and H. Prokisch. "Mapping the genetic architecture of gene regulation in whole blood." In: *PloS one* 9.4 (Jan. 2014), e93844. DOI: 10.1371/journal.pone.0093844.

[82] F. Aguet, A. A. Brown, S. E. Castel, J. R. Davis, Y. He, B. Jo, P. Mohammadi, Y. S. Park, P. Parsana, A. V. Segrè, B. J. Strober, Z. Zappala, P. Guan, S. Koester, A. R. Little, S. J. Trevanion, D. R. Zerbino, B. Craft, M. Goldman, M. Haeussler, W. J. Kent, C. M. Lee, B. Paten, K. R. Rosenbloom, J. Vivian, and J. Zhu. "Genetic effects on gene expression across human tissues". In: *Nature* 550.7675 (Oct. 2017), pp. 204–213. DOI: 10.1038/nature24277.

[83] R. Breitling, Y. Li, B. M. Tesson, J. Fu, C. Wu, T. Wiltshire, A. Gerrits, L. V. Bystrykh, G. de Haan, A. I. Su, and R. C. Jansen. "Genetical genomics: spotlight on QTL hotspots." In: *PLoS genetics* 4.10 (Oct. 2008), e1000232. DOI: 10.1371/journal.pgen.1000232.

[84] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. J. Theis. "Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data." In: *BMC Syst. Biol.* 5 (Jan. 2011), p. 21. DOI: 10.1186/1752-0509-5-21.

[85] A. Saha, Y. Kim, A. D. H. Gewirtz, B. Jo, C. Gao, I. C. McDowell, T. G. GTEx Consortium, B. E. Engelhardt, and A. Battle. "Co-expression networks reveal the tissue-specific regulation of transcription and splicing." In: *Genome Res.* 27.11 (Nov. 2017), pp. 1843–1858. DOI: 10.1101/gr.216721.116.

[86] N. Meinshausen and P. Bühlmann. "High-dimensional graphs and variable selection with the Lasso". In: *Ann. Stat.* 34.3 (June 2006), pp. 1436–1462. DOI: 10.1214/009053606000000281.

[87] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (July 2008), pp. 432–441. DOI: 10.1093/biostatistics/kxm045.

[88] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods". In: *PLoS ONE* 5.9 (Sept. 2010), e12776. DOI: 10.1371/journal.pone.0012776.

[89] A. Greenfield, A. Madar, H. Ostrer, and R. Bonneau. "DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models". In: *PloS one* 5.10 (2010), e13397.

[90] D. Marbach, J. C. Costello, R. Küffner, N. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, the DREAM5 Consortium, M. Kellis, J. J. Collins, and G. Stolovitzky. "Wisdom of crowds for robust gene network inference". In: *Nat. Methods* 9.8 (2012), p. 796. DOI: 10.1038/NMETH.2016.

[91] S. Aibar, C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, Z. K. Atak, J. Wouters, and S. Aerts. "SCENIC: single-cell regulatory network inference and clustering". In: *Nat. Methods* 14.11 (Oct. 2017), pp. 1083–1086. DOI: 10.1038/nmeth.4463.

[92] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. "Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation". In: (2013), pp. 1–9. DOI: 10.4209/aaqr.2014.11.0299. arXiv: 1306.3212.

[93] The GTEx Consortium. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348 (2015), pp. 648–660. DOI: 10.1126/science.1262110.

[94] J. Lee and T. Hastie. "Structure Learning of Mixed Graphical Models". In: *Aistats 16* 31.5 (2013), pp. 388–396. DOI: 10.1080/10618600.2014.900500. arXiv: 1205.5012.

[95] J. M. Haslbeck and L. J. Waldorp. "mgm: Structure Estimation for time-varying mixed graphical models in high-dimensional data". In: *J Stat Softw* (2016).

[96] B. Fellinghauer, P. Bühlmann, M. Ryffel, M. Von Rhein, and J. D. Reinhardt. "Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables". In: *Comput. Stat. Data Anal.* 64 (2013), pp. 132–152. DOI: 10.1016/j.csda.2013.02.022. arXiv: 1109.0152.

[97] J. Zierer, T. Pallister, P.-C. Tsai, J. Krumsiek, J. T. Bell, G. Lauc, T. D. Spector, C. Menni, and G. Kastenmüller. "Exploring the molecular basis of age-related disease comorbidities using a multi-omics graphical model". In: *Sci. Rep.* 6 (Nov. 2016), p. 37646. DOI: 10.1038/srep37646.

[98] A. Mohammadi, F. Abegaz, E. V. D. Heuvel, and E. C. Wit. "Bayesian Gaussian Copula Graphical Modeling for Dupuytren Disease". In: 1 (2015), pp. 1–22. arXiv: 1501.04849.

[99] S. Huang, K. Chaudhary, and L. X. Garmire. "More Is Better: Recent Progress in Multi-Omics Data Integration Methods." In: *Front. Genet.* 8 (2017), p. 84. DOI: 10.3389/fgene.2017.00084.

[100] B. Palsson and K. Zengler. "The challenges of integrating multi-omic data sets". In: *Nature Chemical Biology* 6.11 (Nov. 2010), pp. 787–789. DOI: 10.1038/nchembio.462.

[101] The Genotype Tissue Expression Consortium. "The GTEx Consortium atlas of genetic regulatory effects across human tissues The Genotype Tissue Expression Consortium". In: (2019). DOI: 10.1101/787903.

[102] Y. Zuo, Y. Cui, G. Yu, R. Li, and H. W. Ressom. "Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO". In: *BMC Bioinformatics* 18.1 (Dec. 2017), p. 99. DOI: 10.1186/s12859-017-1515-1.

[103] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering. "STRING v10: protein–protein interaction networks, integrated over the tree of life". In: *Nucleic Acids Research* 43.D1 (Jan. 2015), pp. D447–D452. DOI: 10.1093/nar/gku1003.

[104] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. "STRING: a database of predicted functional associations between proteins". In: *Nucleic Acids Res.* 31.1 (Jan. 2003), pp. 258–261.

[105] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. "BioGRID: a general repository for interaction datasets". In: *Nucleic Acids Res.* 34.Database issue (Jan. 2006), pp. D535–539.

[106] T. Li, R. Wernersson, R. B. Hansen, H. Horn, J. Mercer, G. Slodkowicz, C. T. Workman, O. Rigina, K. Rapacki, H. H. Stærfeldt, S. Brunak, T. S. Jensen, and K. Lage. "A scored human protein-protein interaction network to catalyze genomic interpretation." In: *Nature methods* 14.1 (2017), pp. 61–64. DOI: 10.1038/nmeth.4083.

[107] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, and H. Parkinson. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)." In: *Nucleic acids research* 45.D1 (2017), pp. D896–D901. DOI: 10.1093/nar/gkw1133.

[108] K. Watanabe, S. Stringer, O. Frei, M. Umićević Mirkov, C. de Leeuw, T. J. C. Polderman, S. van der Sluis, O. A. Andreassen, B. M. Neale, and D. Posthuma. "A global overview of pleiotropy and genetic architecture in complex traits." In: *Nature genetics* 51.9 (2019), pp. 1339–1348. DOI: 10.1038/s41588-019-0481-0.

[109] M. A. Kamat, J. A. Blackshaw, R. Young, P. Surendran, S. Burgess, J. Danesh, A. S. Butterworth, and J. R. Staley. "PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations." In: *Bioinformatics (Oxford, England)* 35.22 (Nov. 2019), pp. 4851–4853. DOI: 10.1093/bioinformatics/btz469.

[110] M. Kanehisa and S. Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Res.* 28.1 (Jan. 2000), pp. 27–30. DOI: 10.1093/nar/28.1.27.

[111] D. Alonso-Lopez, M. A. Gutierrez, K. P. Lopes, C. Prieto, R. Santamaria, and J. De Las Rivas. "APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks". In: *Nucleic Acids Res.* 44.W1 (July 2016), W529–535.

[112] K. Blin, C. Dieterich, R. Wurmus, N. Rajewsky, M. Landthaler, and A. Akalin. "DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation." In: *Nucleic acids research* 43.Database issue (Jan. 2015), pp. D160–7. DOI: 10.1093/nar/gku1180.

[113] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. Del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, and H. Hermjakob. "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases". In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D358–D363. DOI: 10.1093/nar/gkt1115.

[114] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. "Pathway Commons, a web resource for biological pathway data". In: *Nucleic Acids Research* 39.Database (Jan. 2011), pp. D685–D690. DOI: 10.1093/nar/gkq1039.

[115] A. Yilmaz, M. K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch, and E. Grotewold. "AGRIS: the Arabidopsis Gene Regulatory Information Server, an update." In: *Nucleic acids research* 39.Database issue (Jan. 2011), pp. D1118–22. DOI: 10.1093/nar/gkq1120.

[116] S. E. Celniker, L. A. L. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. MacAlpine, G. Micklem, F. Piano, M. Snyder, L. Stein, K. P. White, R. H. Waterston, and modENCODE Consortium. "Unlocking the secrets of the genome." In: *Nature* 459.7249 (June 2009), pp. 927–30. DOI: 10.1038/459927a.

[117] R. Edgar, M. Domrachev, and A. E. Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic Acids Res.* 30.1 (Jan. 2002), pp. 207–210.

[118] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. "NCBI GEO: archive for functional genomics data sets–update". In: *Nucleic Acids Res.* 41.Database issue (Jan. 2013), pp. D991–995.

[119] A. Lachmann, D. Torre, A. B. Keenan, K. M. Jagodnik, H. J. Lee, L. Wang, M. C. Silverstein, and A. Ma'ayan. "Massive mining of publicly available RNA-seq data from human and mouse". In: *Nature Communications* 9.1 (Dec. 2018), p. 1366. DOI: 10.1038/s41467-018-03751-6.

[120] P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, Å. Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C. Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. von Feilitzen, K. S. Lilley, M. Uhlén, and E. Lundberg. "A subcellular map of the human proteome." In: *Science (New York, N.Y.)* 356.6340 (May 2017), eaal3321. DOI: 10.1126/science.aal3321.

[121] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J. L. Griffin, and C. Steinbeck. "MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data". In: *Nucleic Acids Research* 41.D1 (Jan. 2013), pp. D781–D786. DOI: 10.1093/nar/gks1004.

[122] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, K. Chang, C. J. Creighton, C. Davis, L. Donehower, J. Drummond, D. Wheeler, A. Ally, M. Balasundaram, I. Birol, S. N. Butterfield, D. Wolf, L. Van 't Veer, E. A. Collisson, D. Anastassiou, T. H. Ou Yang, N. Lopez-Bigas, A. Gonzalez-Perez, D. Tamborero, Z. Xia, W. Li, D. Y. Cho, T. Przytycka, M. Hamilton, S. McGuire, S. Nelander, P. Johansson, R. Jornsten, T. Kling, and J. Sanchez. "The Cancer Genome Atlas Pan-Cancer analysis project". In: *Nat. Genet.* 45.10 (Oct. 2013), pp. 1113–1120.

[123] F. Hosp, H. Vossfeldt, M. Heinig, D. Vasiljevic, A. Arumughan, E. Wyler, M. Landthaler, N. Hubner, E. E. Wanker, L. Lannfelt, M. Ingelsson, M. Lalowski, A. Voigt, and M. Selbach. "Quantitative Interaction Proteomics of Neurodegenerative Disease Proteins". In: *Cell Rep.* 11.7 (May 2015), pp. 1134–1146. DOI: 10.1016/J.CELREP.2015.04.030.

[124] A. Mohammadi and E. C. Wit. "Bayesian Structure Learning in Sparse Gaussian Graphical Models". In: *Bayesian Anal.* 10.1 (Mar. 2015), pp. 109–138. DOI: 10.1214/14-BA889.

[125] A. J. Sedgewick, K. Buschur, I. Shi, J. D. Ramsey, V. K. Raghu, D. V. Manatakis, Y. Zhang, J. Bon, D. Chandra, C. Karoleski, F. C. Sciurba, P. Spirtes, C. Glymour, and P. V. Benos. "Mixed Graphical Models for Integrative Causal Analysis with Application to Chronic Lung Disease Diagnosis and Prognosis". In: *Bioinformatics* (Sept. 2018). Ed. by J. Wren. DOI: 10.1093/bioinformatics/bty769.

[126] A. J. Sedgewick, I. Shi, R. M. Donovan, and P. V. Benos. "Learning mixed graphical models with separate sparsity parameters and stability-based model selection". In: *BMC Bioinformatics* 17.S5 (Dec. 2016), S175. DOI: 10.1186/s12859-016-1039-0.

[127] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. "Using Bayesian networks to analyze expression data". In: *Proceedings of the fourth annual international conference on Computational molecular biology - RECOMB '00* 7 (2000), pp. 127–135. DOI: 10.1145/332306.332355.

[128] A. Greenfield, C. Hafemeister, and R. Bonneau. "Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks". In: *Bioinformatics* 29.8 (2013), pp. 1060–1067. DOI: 10.1093/bioinformatics/btt099.

[129] M. E. Studham, A. Tjärnberg, T. E. Nordling, S. Nelander, and E. L. Sonnhammer. "Functional association networks as priors for gene regulatory network inference". In: *Bioinformatics* 30.12 (2014), pp. 130–138. DOI: 10.1093/bioinformatics/btu285.

[130] K. Y. Lam, Z. M. Westrick, C. L. Müller, L. Christiaen, and R. Bonneau. "Fused Regression for Multi-source Gene Regulatory Network Inference". In: *PLoS Computational Biology* 12.12 (2016), pp. 1–23. DOI: 10.1371/journal.pcbi.1005157.

[131] M. Gustafsson and M. Hörnquist. "Gene expression prediction by soft integration and the elastic net - Best performance of the DREAM3 gene expression challenge". In: *PLoS ONE* 5.2 (2010). DOI: 10.1371/journal.pone.0009134.

[132] A. F. Siahpirani and S. Roy. "A prior-based integrative framework for functional transcriptional regulatory network inference". In: *Nucleic Acids Research* 45.4 (2017), pp. 1–22. DOI: 10.1093/nar/gkw963.

[133] B. Pei and D. G. Shin. "Reconstruction of biological networks by incorporating prior knowledge into bayesian network models". In: *Journal of Computational Biology* 19.12 (2012), pp. 1324–1334. DOI: 10.1089/cmb.2011.0194.

[134] I. Piazza, K. Kochanowski, V. Cappelletti, T. Fuhrer, E. Noor, U. Sauer, and P. Picotti. "A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication." In: *Cell* 172.1-2 (Jan. 2018), 358–372.e23. DOI: 10.1016/j.cell.2017.12.006.

[135] J. K. Harris, K. J. Johnson, B. J. Carothers, X. Wang, T. Combs, and D. A. Luke. "Examining and improving reproducible research practices in public health". In: *Online Journal of Public Health Informatics* 11.1 (May 2019). DOI: 10.5210/ojphi.v11i1.9750.

[136] C. Boettiger and Carl. "An introduction to Docker for reproducible research". In: *ACM SIGOPS Operating Systems Review* 49.1 (Jan. 2015), pp. 71–79. DOI: 10.1145/2723872.2723882.

[137] C. Gandrud. *Reproducible Research with R and RStudio*. New York: Chapman and Hall/CRC, Sept. 2018. DOI: 10.1201/9781315382548.

[138] S. A. Iqbal, J. D. Wallach, M. J. Khoury, S. D. Schully, and J. P. A. Ioannidis. "Reproducible Research Practices and Transparency across the Biomedical Literature". In: *PLOS Biology* 14.1 (Jan. 2016). Ed. by D. L. Vaux, e1002333. DOI: 10.1371/journal.pbio.1002333.

[139] A. Rule, A. Birmingham, C. Zuniga, I. Altintas, S.-C. Huang, R. Knight, N. Moshiri, M. H. Nguyen, S. B. Rosenthal, F. Pérez, and P. W. Rose. "Ten Simple Rules for Reproducible Research in Jupyter Notebooks". In: (Oct. 2018). arXiv: 1810.08055.

[140] F. Markowetz. "Five selfish reasons to work reproducibly". In: *Genome Biology* 16.1 (2015), pp. 1–4. DOI: 10.1186/s13059-015-0850-7.

[141] C. G. Begley and L. M. Ellis. "Raise standards for preclinical cancer research". In: *Nature* 483.7391 (Mar. 2012), pp. 531–533. DOI: 10.1038/483531a.

[142] B. P. Cabral, M. d. G. D. Fonseca, and F. B. Mota. "The recent landscape of cancer research worldwide: A bibliometric and network analysis". In: *Oncotarget* 9.55 (2018), pp. 30474–30484. DOI: 10.18632/oncotarget.25730.

[143] Open Science Collaboration. "PSYCHOLOGY. Estimating the reproducibility of psychological science". In: *Science (New York, N.Y.)* 349.6251 (2015), aac4716. DOI: 10.1126/science.aac4716.

[144] F. Prinz, T. Schlange, and K. Asadullah. "Believe it or not: how much can we rely on published data on potential drug targets?" In: *Nature Reviews Drug Discovery* 10.9 (Sept. 2011), pp. 712–712. DOI: 10.1038/nrd3439-c1.

[145] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Lupia, P. Mabry, T. Madon, N. Malhotra, M. Mcnutt, E. A. Miguel, L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. Vandenbos,

E. J. Wagenmakers, R. Wilson, and T. Yarkoni. "Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility". In: *Science* 348.6242 (2015), pp. 1422–1425. DOI: 10.1126/science.aab2374.Promoting.

[146] M. Baker and D. Penny. "Is there a reproducibility crisis?" In: *Nature* 533.7604 (2016), pp. 452–454. DOI: 10.1038/533452A.

[147] F. K. Dankar, A. Ptitsyn, and S. K. Dankar. "The development of large-scale de-identified biomedical databases in the age of genomics-principles and challenges". In: *Human Genomics* 12.1 (2018), pp. 1–15. DOI: 10.1186/s40246-018-0147-5.

[148] C. A. Cassa, B. Schmidt, I. S. Kohane, and K. D. Mandl. "My sister's keeper?: genomic research and the identifiability of siblings". In: *BMC Medical Genomics* 1.1 (2008), pp. 1–11. DOI: 10.1186/1755-8794-1-32.

[149] J. Köster and S. Rahmann. "Snakemake-a scalable bioinformatics workflow engine". In: *Bioinformatics* 28.19 (Oct. 2012), pp. 2520–2522. DOI: 10.1093/bioinformatics/bts480.

[150] R. Priedhorsky and T. Randles. "Charliecloud: Unprivileged Containers for User-Defined Software Stacks in HPC". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. SC '17. Denver, Colorado: Association for Computing Machinery, 2017. DOI: 10.1145/3126908.3126925.

[151] J. Goecks, A. Nekrutenko, J. Taylor, and T. Galaxy Team. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences". In: *Genome Biology* 11.8 (Aug. 2010), R86. DOI: 10.1186/gb-2010-11-8-r86.

[152] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. "KNIME: The Konstanz Information Miner". In: *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[153] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. "Nextflow enables reproducible computational workflows". In: *Nature Biotechnology* 35.4 (Apr. 2017), pp. 316–319. DOI: 10.1038/nbt.3820.

[154] G. M. Kurtzer, V. Sochat, and M. W. Bauer. "Singularity: Scientific containers for mobility of compute". In: *PLoS ONE* 12.5 (2017), pp. 1–20. DOI: 10.1371/journal.pone.0177459.

[155] MAQC Consortium, L. Shi, L. H. Reid, et al. "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements". In: *Nature Biotechnology* 24.9 (Sept. 2006), p. 1151. DOI: 10.1038/NBT1239.

[156] SEQC Consortium. "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control consortium". In: *Nature biotechnology* 32.9 (2014), p. 903. DOI: 10.1038/NBT.2957.

[157] S. Das, G. R. Abecasis, and B. L. Browning. "Genotype Imputation from Large Reference Panels". In: *Annual Review of Genomics and Human Genetics* 19.1 (2018), pp. 73–96. DOI: 10.1146/annurev-genom-083117-021602.

[158] T. J. Hoffmann, M. N. Kvale, S. E. Hesselson, Y. Zhan, C. Aquino, Y. Cao, S. Cawley, E. Chung, S. Connell, J. Eshragh, M. Ewing, J. Gollub, M. Henderson, E. Hubbell, C. Iribarren, J. Kaufman, R. Z. Lao, Y. Lu, D. Ludwig, G. K. Mathauda, W. McGuire, G. Mei, S. Miles, M. M. Purdy, C. Quesenberry, D. Ranatunga, S. Rowell, M. Sadler, M. H. Shapero, L. Shen, T. R. Shenoy, D. Smethurst, S. K. Van den Eeden, L. Walter, E. Wan, R. Wearley, T. Webster, C. C. Wen, L. Weng, R. A. Whitmer, A. Williams, S. C. Wong, C. Zau, A. Finn, C. Schaefer, P.-Y. Kwok, and N. Risch. "Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array." In: *Genomics* 98.2 (Aug. 2011), pp. 79–89. DOI: 10.1016/j.ygeno.2011.04.005.

[159] T. LaFramboise. "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances." In: *Nucleic acids research* 37.13 (July 2009), pp. 4181–93. DOI: 10.1093/nar/gkp552.

[160] 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. "A global reference for human genetic variation." In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. DOI: 10.1038/nature15393.

[161] M. H. Kowalski, H. Qian, Z. Hou, J. D. Rosen, A. L. Tapia, Y. Shan, D. Jain, M. Argos, D. K. Arnett, C. Avery, K. C. Barnes, L. C. Becker, S. A. Bien, J. C. Bis, J. Blangero, E. Boerwinkle, D. W. Bowden, S. Buyske, J. Cai, M. H. Cho, S. H. Choi, H. Choquet, L. A. Cupples, M. Cushman, M. Daya, P. S. de Vries, P. T. Ellinor, N. Faraday, M. Fornage, S. Gabriel, S. K. Ganesh, M. Graff, N. Gupta, J. He, S. R. Heckbert, B. Hidalgo, C. J. Hodonsky, M. R. Irvin, A. D. Johnson, E. Jorgenson, R. Kaplan, S. L. R. Kardia, T. N. Kelly, C. Kooperberg, J. A. Lasky-Su, R. J. F. Loos, S. A. Lubitz, R. A. Mathias, C. P. McHugh, C. Montgomery, J.-Y. Moon, A. C. Morrison, N. D. Palmer, N. Pankratz, G. J. Papanicolaou, J. M. Peralta, P. A. Peyser, S. S. Rich, J. I. Rotter, E. K. Silverman, J. A. Smith, N. L. Smith, K. D. Taylor, T. A. Thornton, H. K. Tiwari, R. P. Tracy, T. Wang, S. T. Weiss, L.-C. Weng, K. L. Wiggins, J. G. Wilson, L. R. Yanek, S. Zöllner, K. E. North, P. L. Auer, L. M. Raffield, A. P. Reiner, Y. Li, A. P. Reiner, and Y. Li. "Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations". In: *PLOS Genetics* 15.12 (Dec. 2019). Ed. by G. S. Barsh, e1008500. DOI: 10.1371/journal.pgen.1008500.

[162] The International HapMap Consortium, K. A. Frazer, D. G. Ballinger, et al. "A second generation human haplotype map of over 3.1 million SNPs." In: *Nature* 449.7164 (Oct. 2007), pp. 851–61. DOI: 10.1038/nature06258.

[163] B. L. Browning and S. R. Browning. "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals." In: *American journal of human genetics* 84.2 (Feb. 2009), pp. 210–23. DOI: 10.1016/j.ajhg.2009.01.005.

[164] B. N. Howie, P. Donnelly, and J. Marchini. "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies". In: *PLoS Genetics* 5.6 (June 2009). Ed. by N. J. Schork, e1000529. DOI: 10.1371/journal.pgen.1000529.

[165] B. Howie, J. Marchini, and M. Stephens. "Genotype imputation with thousands of genomes." In: *G3 (Bethesda, Md.)* 1.6 (Nov. 2011), pp. 457–70. DOI: 10.1534/g3.111.001198.

[166] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. "Fast and accurate genotype imputation in genome-wide association studies through prephasing". In: *Nature Genetics* 44.8 (Aug. 2012), pp. 955–959. DOI: 10.1038/ng.2354.

[167] J. Sandoval, H. Heyn, S. Moran, J. Serra-Musach, M. A. Pujana, M. Bibikova, and M. Esteller. "Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome". In: *Epigenetics* 6.6 (June 2011), pp. 692–702. DOI: 10.4161/epi.6.6.16196.

[168] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, and R. Shen. "High density DNA methylation array with single CpG site resolution". In: *Genomics* 98.4 (Oct. 2011), pp. 288–295. DOI: 10.1016/J.YGENO.2011.07.007.

[169] C. S. Wilhelm-Benartzi, D. C. Koestler, M. R. Karagas, J. M. Flanagan, B. C. Christensen, K. T. Kelsey, C. J. Marsit, E. A. Houseman, and R. Brown. "Review of processing and analysis methods for DNA methylation array data". In: *British Journal of Cancer* 109.6 (Sept. 2013), pp. 1394–1402. DOI: 10.1038/bjc.2013.496.

[170] J. A. Heiss and A. C. Just. "Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses". In: *Clinical Epigenetics* 11.1 (Dec. 2019), p. 15. DOI: 10.1186/s13148-019-0615-3.

[171] B. Lehne, A. W. Drong, M. Loh, W. Zhang, W. R. Scott, S.-T. Tan, U. Afzal, J. Scott, M.-R. Jarvelin, P. Elliott, M. I. McCarthy, J. S. Kooner, and J. C. Chambers. "A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies". In: *Genome Biology* 16.1 (Dec. 2015), p. 37. DOI: 10.1186/s13059-015-0600-x.

[172] E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey. "DNA methylation arrays as surrogate measures of cell mixture distribution." In: *BMC bioinformatics* 13 (May 2012), p. 86. DOI: 10.1186/1471-2105-13-86.

[173] I. Barbulovic-Nad, M. Lucente, Y. Sun, M. Zhang, A. R. Wheeler, and M. Bussmann. "Bio-Microarray Fabrication Techniques—A Review". In: *Critical Reviews in Biotechnology* 26.4 (Jan. 2006), pp. 237–259. DOI: 10.1080/07388550600978358.

[174] H. Auer, S. Lyianarachchi, D. Newsom, M. I. Klisovic, uido Marcucci, and K. Kornacker. "Chipping away at the chip bias: RNA degradation in microarray analysis". In: *Nature Genetics* 35.4 (Dec. 2003), pp. 292–293. DOI: 10.1038/ng1203-292.

[175] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, S. Lightfoot, W. Menzel, M. Granzow, and T. Ragg. "The RIN: an RNA integrity number for assigning integrity values to RNA measurements". In: *BMC Molecular Biology* 7.1 (Jan. 2006), p. 3. DOI: 10.1186/1471-2199-7-3.

[176] R. Holle, M. Happich, H. Löwel, and H. E. Wichmann. "KORA - A research platform for population based health research". In: *Gesundheitswesen* 67.SUPPL. 1 (2005). DOI: 10.1055/s-2005-858235.

[177] A. Döring, C. Gieger, D. Mehta, H. Gohlke, H. Prokisch, S. Coassin, G. Fischer, K. Henke, N. Klopp, F. Kronenberg, B. Paulweber, A. Pfeufer, D. Rosskopf, H. Völzke, T. Illig, T. Meitinger, H.-E. Wichmann, and C. Meisinger. "SLC2A9 influences uric acid concentrations with pronounced sex-specific effects". In: *Nature Genetics* 40.4 (Apr. 2008), pp. 430–436. DOI: 10.1038/ng.107.

[178] M. Steffens, C. Lamina, I. T, B. T, V. R, E. P, S. EK, T. MR, K. N, C. A, K. IR, K. K, L. J, D. L. A, F. R, L. P, Z. A, W. A, K. M, N. P, H. J, S. S, M. T, W. HE, R. K, W. TF, and B. MP. "SNP-based Analysis of Genetic Substructure in the German Population". In: *Human heredity* 62.1 (2006). DOI: 10.1159/000095850.

[179] S. Zeilinger, B. Kühnel, N. Klopp, H. Baurecht, A. Kleinschmidt, C. Gieger, S. Weidinger, E. Lattka, J. Adamski, A. Peters, K. Strauch, M. Waldenberger, and T. Illig. "Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation". In: *PLoS ONE* 8.5 (May 2013). Ed. by A. Chen, e63812. DOI: 10.1371/journal.pone.0063812.

[180] D. Mehta, K. Heim, C. Herder, M. Carstensen, G. Eckstein, C. Schurmann, G. Homuth, M. Nauck, U. Völker, M. Roden, T. Illig, C. Gieger, T. Meitinger, and H. Prokisch. "Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood." In: *European journal of human genetics : EJHG* 21.1 (Jan. 2013), pp. 48–54. DOI: 10.1038/ejhg.2012.106.

[181] P. Rantakallio. "The longitudinal study of the Northern Finland birth cohort of 1966". In: *Paediatric and Perinatal Epidemiology* 2.1 (Jan. 1988), pp. 59–88. DOI: 10.1111/j.1365-3016.1988.tb00180.x.

[182] U. Sovio, A. J. Bennett, I. Y. Millwood, J. Molitor, P. F. O'Reilly, N. J. Timpson, M. Kaakinen, J. Laitinen, J. Haukka, D. Pillas, I. Tzoulaki, J. Molitor, C. Hoggart, L. J. M. Coin, J. Whittaker, A. Pouta, A.-L. Hartikainen, N. B. Freimer, E. Widen, L. Peltonen, P. Elliott, M. I. McCarthy, and M.-R. Jarvelin. "Genetic Determinants of Height Growth Assessed Longitudinally from Infancy to Adulthood in the Northern Finland Birth Cohort 1966". In: *PLoS Genetics* 5.3 (2009). DOI: 10.1371/JOURNAL.PGEN.1000409.

[183] M. R. Järvelin, A.-L. Hartikainen-Sorri, and P. Rantakallio. "Labour induction policy in hospitals of different levels of specialisation". In: *BJOG: An International Journal of Obstetrics & Gynaecology* 100.4 (Apr. 1993), pp. 310–315. DOI: 10.1111/j.1471-0528.1993.tb12971.x.

[184] J. A, S. U, K. M, K. M, S. MJ, F. P, C. S, J. MR, and L. J. "Meal Frequencies Modify the Effect of Common Genetic Variants on Body Mass Index in Adolescents of the Northern Finland Birth Cohort 1986". In: *PloS one* 8.9 (2013). DOI: 10.1371/JOURNAL.PONE.0073802.

[185] Z. Pausova, T. Paus, M. Abrahamowicz, M. Bernard, D. Gaudet, G. Leonard, M. Peron, G. B. Pike, L. Richer, J. R. Séguin, and S. Veillette. "Cohort Profile: The Saguenay Youth Study (SYS)". In: *International Journal of Epidemiology* 46.2 (Mar. 2016), e19–e19. DOI: 10.1093/ije/dyw023. eprint: https://academic.oup.com/ije/article-pdf/46/2/e19/24171838/dyw023.pdf.

[186] T. Paus, Z. Pausova, M. Abrahamowicz, D. Gaudet, G. Leonard, G. B. Pike, and L. Richer. "Saguenay Youth Study: A multi-generational approach to studying virtual trajectories of the brain and cardio-metabolic health". In: *Developmental Cognitive Neuroscience* 11 (2015), pp. 129–144. DOI: 10.1016/j.dcn.2014.10.003.

[187] R. Leinonen, H. Sugawara, M. Shumway, and International Nucleotide Sequence Database Collaboration. "The sequence read archive." In: *Nucleic acids research* 39.Database issue (Jan. 2011), pp. D19–21. DOI: 10.1093/nar/gkq1019.

[188] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, F. Zhang, S. Dolma, A. Willems, J. Coulombe-Huntington, A. Chatr-Aryamontri, K. Dolinski, and M. Tyers. "The BioGRID interaction database: 2019 update." In: *Nucleic acids research* 47.D1 (Jan. 2019), pp. D529–D541. DOI: 10.1093/nar/gky1079.

[189] A. Athar, A. Füllgrabe, N. George, H. Iqbal, L. Huerta, A. Ali, C. Snow, N. A. Fonseca, R. Petryszak, I. Papatheodorou, U. Sarkans, and A. Brazma. "ArrayExpress update – from bulk to single-cell expression data". In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D711–D715. DOI: 10.1093/nar/gky964.

[190] P. J. Park. "ChIP-seq: advantages and challenges of a maturing technology." In: *Nature reviews. Genetics* 10.10 (2009), pp. 669–80. DOI: 10.1038/nrg2641. eprint: NIHMS150003.

[191] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[192] L. Fahrmeir, R. Künstler, I. Pigeot, and G. Tutz. *Statistik: der Weg zur Datenanalyse*. Berlin: Springer, 2011.

[193] Y. R. Wang and H. Huang. "Review on statistical methods for gene network reconstruction using expression data". In: *Journal of Theoretical Biology* 362 (Dec. 2014), pp. 53–61. DOI: 10.1016/J.JTBI.2014.03.040.

[194] J. J. Faraway. *Linear models with R*. CRC press, 2014.

[195] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, NY, USA: 2001, p. 72.

[196] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[197] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2017.

[198] R. Tibshirani. "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x.

[199] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. "Pathwise coordinate optimization". In: (Aug. 2007). DOI: 10.1214/07-AOAS131. arXiv: 0708.1485.

[200] E. Zeggini and J. P. Ioannidis. "Meta-analysis in genome-wide association studies". In: *Pharmacogenomics* 10.2 (Feb. 2009), p. 191. DOI: 10.2217/14622416.10.2.191.

[201] K. L. Lunetta. "Methods for Meta-Analysis of Genetic Data". In: *Current Protocols in Human Genetics* 77.1 (Apr. 2013), pp. 1.24.1–1.24.8. DOI: 10.1002/0471142905.hg0124s77.

[202] R. A. Fisher. "Statistical methods for research workers". In: *Breakthroughs in statistics*. Springer, 1992, pp. 66–70.

[203] 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 (1995), pp. 289–300.

[204] H. Zou, T. Hastie, and R. Tibshirani. "On the "degrees of freedom" of the lasso". In: *Annals of Statistics* 35.5 (2007), pp. 2173–2192. DOI: 10.1214/009053607000000127.

[205] I. Murray, Z. Ghahramani, and D. MacKay. *MCMC for doubly-intractable distributions*. 2012. arXiv: 1206.6848 [stat.CO].

[206] O. Delaneau, J.-F. Zagury, and J. Marchini. "Improved whole-chromosome phasing for disease and population genetic studies". In: *Nature Methods* 10.1 (Jan. 2013), pp. 5–6. DOI: 10.1038/nmeth.2307.

[207] L. Pfeiffer, S. Wahl, L. C. Pilling, E. Reischl, J. K. Sandling, S. Kunze, L. M. Holdt, A. Kretschmer, K. Schramm, J. Adamski, N. Klopp, T. Illig, Å. K. Hedman, M. Roden, D. G. Hernandez, A. B. Singleton, W. E. Thasler, H. Grallert, C. Gieger, C. Herder, D. Teupser, C. Meisinger, T. D. Spector, F. Kronenberg, H. Prokisch, D. Melzer, A. Peters, P. Deloukas, L. Ferrucci, and M. Waldenberger. "DNA methylation of lipid-related genes affects blood lipid levels." In: *Circulation. Cardiovascular genetics* 8.2 (Apr. 2015), pp. 334–42. DOI: 10.1161/CIRCGENETICS.114.000804.

[208] E. Grundberg, E. Meduri, J. K. Sandling, Å. K. Hedman, S. Keildson, A. Buil, S. Busche, W. Yuan, J. Nisbet, M. Sekowska, A. Wilk, A. Barrett, K. S. Small, B. Ge, M. Caron, S. Y. Shin, M. Lathrop, E. T. Dermitzakis, M. I. McCarthy, T. D. Spector, J. T. Bell, and P. Deloukas. "Global analysis of dna methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements". In: *American Journal of Human Genetics* 93.5 (2013), pp. 876–890. DOI: 10.1016/j.ajhg.2013.10.004.

[209] K. D. Robertson. "DNA methylation and human disease". In: *Nature Reviews Genetics* 6.8 (Aug. 2005), pp. 597–610. DOI: 10.1038/nrg1655.

[210] M. J. Bonder, S. Kasela, M. Kals, R. Tamm, K. Lokk, I. Barragan, W. A. Buurman, P. Deelen, J.-W. Greve, M. Ivanov, S. S. Rensen, J. V. van Vliet-Ostaptchouk, M. G. Wolfs, J. Fu, M. H. Hofker, C. Wijmenga, A. Zhernakova, M. Ingelman-Sundberg, L. Franke, and L. Milani. "Genetic and epigenetic regulation of gene expression in fetal and adult human livers." In: *BMC genomics* 15.1 (Oct. 2014), p. 860. DOI: 10.1186/1471-2164-15-860.

[211] M. Lemire, S. H. Zaidi, M. Ban, B. Ge, D. Aïssi, M. Germain, I. Kassam, M. Wang, B. W. Zanke, F. Gagnon, P.-E. Morange, D.-A. Trégouët, P. S. Wells, S. Sawcer, S. Gallinger, T. Pastinen, and T. J. Hudson. "Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci". In: *Nature Communications* 6.1 (Dec. 2015), p. 6326. DOI: 10.1038/ncomms7326.

[212] M. Gutierrez-Arcelus, H. Ongen, T. Lappalainen, S. B. Montgomery, A. Buil, A. Yurovsky, J. Bryois, I. Padioleau, L. Romano, A. Planchon, E. Falconnet, D. Bielser, M. Gagnebin, T. Giger, C. Borel, A. Letourneau, P. Makrythanasis, M. Guipponi, C. Gehrig, S. E. Antonarakis, and E. T. Dermitzakis. "Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing". In: *PLoS Genet.* 11.1 (Jan. 2015), e1004958.

[213] J. R. Gibbs, M. P. van der Brug, D. G. Hernandez, B. J. Traynor, M. A. Nalls, S. L. Lai, S. Arepalli, A. Dillman, I. P. Rafferty, J. Troncoso, R. Johnson, H. R. Zielke, L. Ferrucci, D. L. Longo, M. R. Cookson, and A. B. Singleton. "Abundant quantitative trait loci exist for DNA methylation and gene expression in Human Brain". In: *PLoS Genetics* 6.5 (2010), p. 29. DOI: 10.1371/journal.pgen.1000952.

[214] G. B. Ehret, D. Lamparter, C. J. Hoggart, J. C. Whittaker, J. S. Beckmann, and Z. Kutalik. "A multi-SNP locus-association method reveals a substantial fraction of the missing heritability". In: *American Journal of Human Genetics* 91.5 (2012), pp. 863–871. DOI: 10.1016/j.ajhg.2012.09.013.

[215] Z. Kutalik, T. Johnson, M. Bochud, V. Mooser, P. Vollenweider, G. Waeber, D. Waterworth, J. S. Beckmann, and S. Bergmann. "Methods for testing association between uncertain genotypes and quantitative traits". In: *Biostatistics* 12.1 (2011), pp. 1–17. DOI: 10.1093/biostatistics/kxq039.

[216] C. J. Hoggart, G. Venturini, M. Mangino, et al. "Novel approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of-origin effect on body mass index". In: *PLoS Genetics* 10.7 (2014), pp. 1–12. DOI: 10.1371/journal.pgen.1004508.

[217] A. A. Shabalin. "Matrix eQTL: ultra fast eQTL analysis via large matrix operations". In: *Bioinformatics* 28.10 (May 2012), pp. 1353–1358. DOI: 10.1093/bioinformatics/bts163.

[218] C. J. Willer, Y. Li, and G. R. Abecasis. "METAL: Fast and efficient meta-analysis of genomewide association scans". In: *Bioinformatics* 26.17 (2010), pp. 2190–2191. DOI: 10.1093/bioinformatics/btq340.

[219] X. Zhou and M. Stephens. "Genome-wide efficient mixed-model analysis for association studies". In: *Nature Genetics* 44.7 (2012), pp. 821–824. DOI: 10.1038/ng.2310.

[220] C. G. Bell, F. Gao, W. Yuan, L. Roos, R. J. Acton, Y. Xia, J. Bell, K. Ward, M. Mangino, P. G. Hysi, J. Wang, and T. D. Spector. "Obligatory and facilitative allelic variation in the DNA methylome within common disease-associated loci". In: *Nature Communications* 9.1 (2018). DOI: 10.1038/s41467-017-01586-1.

[221] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. "Topological domains in mammalian genomes identified by analysis of chromatin interactions". In: *Nature* 485.7398 (2012), pp. 376–380. DOI: 10.1038/nature11082.

[222] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. "A census of human transcription factors: function, expression and evolution". In: *Nature Reviews Genetics* 10.4 (Apr. 2009), pp. 252–263. DOI: 10.1038/nrg2538.

[223] S. Suthram, A. Beyer, R. M. Karp, Y. Eldar, and T. Ideker. "eQED: an efficient method for interpreting eQTL associations using protein networks." In: *Molecular systems biology* 4 (2008), p. 162. DOI: 10.1038/msb.2008.4.

[224] Z. Tu, L. Wang, M. N. Arbeitman, T. Chen, and F. Sun. "An integrative approach for causal gene identification and gene regulatory pathway inference". In: *Bioinformatics* 22.14 (July 2006), e489–e496. DOI: 10.1093/bioinformatics/btl234.

[225] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis. "Diffusion pseudotime robustly reconstructs lineage branching". In: *Nature Methods* 13.10 (Oct. 2016), pp. 845–848. DOI: 10.1038/nmeth.3971.

[226] L. Haghverdi, F. Buettner, and F. J. Theis. "Diffusion maps for high-dimensional single-cell analysis of differentiation data". In: *Bioinformatics* 31.18 (Sept. 2015), pp. 2989–2998. DOI: 10.1093/bioinformatics/btv325.

[227] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows Wheeler transform". In: *Bioinformatics (Oxford, England)* 25.14 (July 2009), pp. 1754–60. DOI: 10.1093/bioinformatics/btp324.

[228] V. Kumar, M. Muratani, N. A. Rayan, P. Kraus, T. Lufkin, H. H. Ng, and S. Prabhakar. "Uniform, optimal signal processing of mapped deep-sequencing data". In: *Nature Biotechnology* 31.7 (July 2013), pp. 615–622. DOI: 10.1038/nbt.2596.

[229] M. L. Sinkus, C. E. Adams, J. Logel, R. Freedman, and S. Leonard. "Expression of immune genes on chromosome 6p21.3–22.1 in schizophrenia". In: *Brain, Behavior, and Immunity* 32 (Aug. 2013), pp. 51–62. DOI: 10.1016/j.bbi.2013.01.087.

[230] H. G. Stunnenberg, M. Hirst, S. Abrignani, D. Adams, M. de Almeida, L. Altucci, V. Amin, I. Amit, S. E. Antonarakis, S. Aparicio, T. Arima, L. Arrigoni, R. Arts, V. Asnafi, M. Esteller, J.-B. Bae, K. Bassler, S. Beck, B. Berkman, P. Südbeck, H. Sun, N. Suzuki, Y. Suzuki, A. Tanay, D. Torrents, F. L. Tyson, T. Ulas, S. Ullrich, T. Ushijima, A. Valencia, E. Vellenga, M. Vingron, C. Wallace, S. Wallner, J. Walter, H. Wang, S. Weber, N. Weiler, A. Weller, A. Weng, S. Wilder, S. M. Wiseman, A. R. Wu, Z. Wu, J. Xiong, Y. Yamashita, X. Yang, D. Y. Yap, K. Y. Yip, S. Yip, J.-I. Yoo, D. Zerbino, and G. Zipprich. "The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery". In: *Cell* 167.5 (Nov. 2016), pp. 1145–1149. DOI: 10.1016/j.cell.2016.11.007.

[231] J. Cairns, P. Freire-Pritchett, S. W. Wingett, C. Várnai, A. Dimond, V. Plagnol, D. Zerbino, S. Schoenfelder, B.-M. Javierre, C. Osborne, P. Fraser, and M. Spivakov. "CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data". In: *Genome Biology* 17.1 (Dec. 2016), p. 127. DOI: 10.1186/s13059-016-0992-2.

[232] E. C. Schofield, T. Carver, P. Achuthan, P. Freire-Pritchett, M. Spivakov, J. A. Todd, and O. S. Burren. "CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets". In: *Bioinformatics* 32.16 (Aug. 2016), pp. 2511–2513. DOI: 10.1093/bioinformatics/btw173.

[233] Q. Szabo, F. Bantignies, and G. Cavalli. "Principles of genome folding into topologically associating domains". In: *Science Advances* 5.4 (Apr. 2019), eaaw1668. DOI: 10.1126/sciadv.aaw1668.

[234] Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, S. Yoshida, R. R. Graham, A. Manoharan, W. Ortmann, T. Bhangale, J. C. Denny, R. J. Carroll, A. E. Eyler, J. D. Greenberg, J. M. Kremer, D. A. Pappas, L. Jiang, J. Yin, L. Ye, D. F. Su, J. Yang, G. Xie, E. Keystone, H. J. Westra, T. Esko, A. Metspalu, X. Zhou, N. Gupta, D. Mirel, E. A. Stahl, D. Diogo, J. Cui, K. Liao, M. H. Guo, K. Myouzen, T. Kawaguchi, M. J. Coenen, P. L. Van Riel, M. A. Van

De Laar, H. J. Guchelaar, T. W. Huizinga, P. Dieudé, X. Mariette, S. L. Bridges, A. Zhernakova, R. E. Toes, P. P. Tak, C. Miceli-Richard, S. Y. Bang, H. S. Lee, J. Martin, M. A. Gonzalez-Gay, L. Rodriguez-Rodriguez, S. Rantapää-Dahlqvist, L. Ärlestig, H. K. Choi, Y. Kamatani, P. Galan, M. Lathrop, S. Eyre, J. Bowes, A. Barton, N. De Vries, L. W. Moreland, L. A. Criswell, E. W. Karlson, A. Taniguchi, R. Yamada, M. Kubo, J. S. Liu, S. C. Bae, J. Worthington, L. Padyukov, L. Klareskog, P. K. Gregersen, S. Raychaudhuri, B. E. Stranger, P. L. De Jager, L. Franke, P. M. Visscher, M. A. Brown, H. Yamanaka, T. Mimori, A. Takahashi, H. Xu, T. W. Behrens, K. A. Siminovitch, S. Momohara, F. Matsuda, K. Yamamoto, and R. M. Plenge. "Genetics of rheumatoid arthritis contributes to biology and drug discovery". In: *Nature* 506.7488 (2014), pp. 376–381. DOI: 10.1038/nature12873.

[235] P. Emery, E. Keystone, H. P. Tony, A. Cantagrel, R. van Vollenhoven, A. Sanchez, E. Alecock, J. Lee, and J. Kremer. "IL-6 receptor inhibition with tocilizumab improves treatment outcomes in patients with rheumatoid arthritis refractory to anti-tumour necrosis factor biologicals: results from a 24-week multicentre randomised placebo-controlled trial". In: *Annals of the Rheumatic Diseases* 67.11 (2008), pp. 1516–1523. DOI: 10.1136/ard.2008.092932. eprint: https://ard.bmj.com/content/67/11/1516.full.pdf.

[236] I. Navarro-Millán, J. A. Singh, and J. R. Curtis. "Systematic Review of Tocilizumab for Rheumatoid Arthritis: A New Biologic Agent Targeting the Interleukin-6 Receptor". In: *Clinical Therapeutics* 34.4 (2012), 788–802.e3. DOI: https://doi.org/10.1016/j.clinthera.2012.02.014.

[237] X. Wen, R. Pique-Regi, and F. Luca. "Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization". In: *PLOS Genetics* 13.3 (Mar. 2017). Ed. by B. Li, e1006646. DOI: 10.1371/journal.pgen.1006646.

[238] M. Pividori, P. S. Rajagopal, A. N. Barbeira, Y. Liang, O. Melia, L. Bastarache, Y. Park, T. G. Consortium, X. Wen, and H. K. Im. "PhenomeXcan: Mapping the genome to the phenome through the transcriptome". In: *bioRxiv* (2019), p. 833210. DOI: 10.1101/833210.

[239] T. Lappalainen, M. Sammeth, M. R. Friedländer, P. a. C. 't Hoen, J. Monlong, M. a. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis. "Transcriptome and genome

sequencing uncovers functional variation in humans." In: *Nature* 501.7468 (2013), pp. 506–11. DOI: 10.1038/nature12531. eprint: NIHMS150003.

[240] J. D. Storey and R. Tibshirani. "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences of the United States of America* 100.16 (2003), pp. 9440–9445. DOI: 10.1073/pnas.1530509100.

[241] R. R. Miles, D. K. Crockett, M. S. Lim, and K. S. Elenitoba-Johnson. "Analysis of BCL6-interacting proteins by tandem mass spectrometry". In: *Molecular & Cellular Proteomics* 4.12 (2005), pp. 1898–1909.

[242] P. Dhordain, O. Albagli, N. Honore, F. Guidez, D. Lantoine, M. Schmid, H. De The, A. Zelent, and M. H. Koken. "Colocalization and heteromerization between the two human oncogene POZ/zinc finger proteins, LAZ3 (BCL6) and PLZF". In: *Oncogene* 19.54 (2000), pp. 6240–6250. DOI: 10.1038/sj.onc.1203976.

[243] D. K. Lee, D. Suh, H. J. Edenberg, and M. W. Hur. "POZ domain transcription factor, FBI-1, represses transcription of ADH5/FDH by interacting with the zinc finger and interfering with DNA binding activity of Sp1". In: *Journal of Biological Chemistry* 277.30 (2002), pp. 26761–26768. DOI: 10.1074/jbc.M202078200.

[244] F. Wei, K. Zaprazna, J. Wang, and M. L. Atchison. "PU.1 Can Recruit BCL6 to DNA To Repress Gene Expression in Germinal Center B Cells". In: *Molecular and Cellular Biology* 29.17 (2009), pp. 4612–4622. DOI: 10.1128/mcb.00234-09.

[245] R. Luijk, K. F. Dekkers, M. van Iterson, W. Arindrarto, A. Claringbould, P. Hop, D. I. Boomsma, C. M. van Duijn, M. M. J. van Greevenbroek, J. H. Veldink, C. Wijmenga, L. Franke, P. A. C. 't Hoen, R. Jansen, J. van Meurs, H. Mei, P. E. Slagboom, B. T. Heijmans, and E. W. van Zwet. "Genome-wide identification of directed gene networks using large-scale population genomics data". In: *Nature Communications* 9.1 (Dec. 2018), p. 3097. DOI: 10.1038/s41467-018-05452-6.

[246] I. Ahel, D. Ahel, T. Matsusaka, A. J. Clark, J. Pines, S. J. Boulton, and S. C. West. "Poly(ADP-ribose)-binding zinc finger motifs in DNA repair/checkpoint proteins". In: *Nature* 451.7174 (2008), pp. 81–85. DOI: 10.1038/nature06420.

[247] A. J. Garvin, R. M. Densham, S. A. Blair-Reid, K. M. Pratt, H. R. Stone, D. Weekes, K. J. Lawrence, and J. R. Morris. "The deSUMOylase SENP7 promotes chromatin relaxation for homologous recombination DNA repair". In: *EMBO Reports* 14.11 (2013), pp. 975–983. DOI: 10.1038/embor.2013.141.

[248] W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, and B. E. Engelhardt. "Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements". In: *Genome Biology* 16.1 (Dec. 2015), p. 14. DOI: 10.1186/s13059-015-0581-9.

[249] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson. "The inferelator: An algorithn for learning parsimonious regulatory networks from systems-biology data sets de novo". In: *Genome Biology* 7.5 (2006). DOI: 10.1186/gb-2006-7-5-r36.

[250] J. C. Castro, I. Valdés, L. N. Gonzalez-García, G. Danies, S. Cañas, F. V. Winck, C. E. Ñústez, S. Restrepo, and D. M. Riaño-Pachón. "Gene regulatory networks on transfer entropy (GRNTE): A novel approach to reconstruct gene regulatory interactions applied to a case study for the plant pathogen Phytophthora infestans". In: *Theoretical Biology and Medical Modelling* 16.1 (2019), pp. 1–15. DOI: 10.1186/s12976-019-0103-7.

[251] J. Carrera, G. Rodrigo, A. Jaramillo, and S. F. Elena. "Reverse-engineering the Arabidopsis thaliana transcriptional network under changing environmental conditions". In: *Genome Biology* 10.9 (Sept. 2009), R96. DOI: 10.1186/gb-2009-10-9-r96.

[252] E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis. "An integrative genomics approach to infer causal associations between gene expression and disease." In: *Nature genetics* 37.7 (2005), pp. 710–7. DOI: 10.1038/ng1589.

[253] J. J. Keurentjes, J. Fu, I. R. Terpstra, J. M. Garcia, G. Van Den Ackerveken, L. B. Snoek, A. J. Peeters, D. Vreugdenhil, M. Koornneef, and R. C. Jansen. "Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.5 (2007), pp. 1708–1713. DOI: 10.1073/pnas.0610429104.

[254] K. L. Mine, N. Shulzhenko, A. Yambartsev, M. Rochman, G. F. O. Sanson, M. Lando, S. Varma, J. Skinner, N. Volfovsky, T. Deng, S. M. F. Brenna, C. R. N. Carvalho, J. C. L. Ribalta, M. Bustin, P. Matzinger, Ismael D C, H. Lyng, M. Gerbase-DeLima, and A. Morgun. "Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer". en. In: *Nat. Commun.* 4 (May 2013), p. 1806.

[255] A. Kamoun, A. Idbaih, C. Dehais, N. Elarouci, C. Carpentier, E. Letouzé, C. Colin, K. Mokhtari, A. Jouvet, E. Uro-Coste, N. Martin-Duverneuil, M. Sanson, J.-Y. Delattre, D. Figarella-Branger, A. de Reyniès, F. Ducray, POLA network, C. Adam, M. Andraud, M.-H. Aubriot-Lorton, L. Bauchet, P. Beauchesne, F. Bielle, C. Blechet, M. Campone, A. F. Carpentier, I. Carpiuc, D. Cazals-Hatem, M.-P. Chenard, D. Chiforeanu, O. Chinot, E. Cohen-Moyal, P. Colin, P. Dam-Hieu, C. Desenclos, N. Desse, F. Dhermain, M.-D. Diebold, S. Eimer, T. Faillot, M. Fesneau, D. Fontaine, S. Gaillard, G. Gauchotte, C. Gaultier, F. Ghiringhelli, J. Godard, E. M. Gueye, J. S. Guillamo, S. Hamdi-Elouadhani, J. Honnorat, J. L. Kemeny, T. Khallil, F. Labrousse, O. Langlois, A. Laquerriere, D. Larrieu-Ciron, E. Lechapt-Zalcman, C. Le Guérinel, P.-M. Levillain, H. Loiseau, D. Loussouarn, C.-A. Maurage, P. Menei, M. J. M. Fotso, G. Noel, F. Parker, M. Peoc'h, M. Polivka, I. Quintin-Roué, C. Ramirez, D. Ricard, P. Richard, V. Rigau, A. Rousseau, G. Runavot, H. Sevestre,

M. C. Tortel, F. Vandenbos, E. Vauleon, G. Viennet, and C. Villa. "Integrated multi-omics analysis of oligodendroglial tumours identifies three subgroups of 1p/19q co-deleted gliomas". en. In: *Nat. Commun.* 7 (Apr. 2016), p. 11263.

[256] S. Christley, Q. Nie, and X. Xie. "Incorporating Existing Network Information into Gene Network Inference". In: *PLoS ONE* 4.8 (Aug. 2009). Ed. by C. Seoighe, e6799. DOI: 10.1371/journal.pone.0006799.

[257] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, and J. M. Cherry. "The Encyclopedia of DNA elements (ENCODE): data portal update." In: *Nucleic acids research* 46.D1 (2018), pp. D794–D801. DOI: 10.1093/nar/gkx1081.

[258] D. Quang and X. Xie. "FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data". In: *Methods* November 2018 (2019), pp. 1–8. DOI: 10.1016/j.ymeth.2019.03.020.

[259] J. Siek, A. Lumsdaine, and L.-Q. Lee. *The boost graph library: user guide and reference manual*. Addison-Wesley, 2002.

[260] B. Efron et al. "Microarrays, empirical Bayes and the two-groups model". In: *Statistical science* 23.1 (2008), pp. 1–22.

[261] B. Efron. *Local false discovery rates*. 2005.

[262] L. J. Scott, M. R. Erdos, J. R. Huyghe, R. P. Welch, A. T. Beck, B. N. Wolford, P. S. Chines, J. P. Didion, N. Narisu, H. M. Stringham, D. L. Taylor, A. U. Jackson, S. Vadlamudi, L. L. Bonnycastle, L. Kinnunen, J. Saramies, J. Sundvall, R. D. Albanus, A. Kiseleva, J. Hensley, G. E. Crawford, H. Jiang, X. Wen, R. M. Watanabe, T. A. Lakka, K. L. Mohlke, M. Laakso, J. Tuomilehto, H. A. Koistinen, M. Boehnke, F. S. Collins, and S. C. J. Parker. "The genetic regulatory signature of type 2 diabetes in human skeletal muscle". In: *Nature Communications* 7.1 (Sept. 2016), p. 11764. DOI: 10.1038/ncomms11764.

[263] P. Parsana, C. Ruberman, A. E. Jaffe, M. C. Schatz, A. Battle, and J. T. Leek. "Addressing confounding artifacts in reconstruction of gene co-expression networks". In: *Genome Biology* 20.1 (Dec. 2019), p. 94. DOI: 10.1186/s13059-019-1700-9.

[264] B. W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *BBA - Protein Structure* (1975). DOI: 10.1016/0005-2795(75)90109-9.

[265] D. Chicco and G. Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC Genomics* 21.1 (2020), pp. 1–13. DOI: 10.1186/s12864-019-6413-7.

[266] A.-L. Boulesteix and K. Strimmer. "Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach." In: *Theoretical biology & medical modelling* 2 (June 2005), p. 23. DOI: 10.1186/1742-4682-2-23.

[267] M. L. Arrieta-Ortiz, C. Hafemeister, A. R. Bate, T. Chu, A. Greenfield, B. Shuster, S. N. Barry, M. Gallitto, B. Liu, T. Kacmarczyk, F. Santoriello, J. Chen, C. D. A. Rodrigues, T. Sato, D. Z. Rudner, A. Driks, R. Bonneau, and P. Eichenberger. "An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network." In: *Molecular systems biology* 11.11 (Nov. 2015), p. 839. DOI: 10.15252/msb.20156236.

[268] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context". In: *BMC Bioinformatics* 7 Suppl 1 (Mar. 2006), S7.

[269] A. Lachmann, F. M. Giorgi, G. Lopez, and A. Califano. "ARACNe-AP: Gene network reverse engineering through adaptive partitioning inference of mutual information". In: *Bioinformatics* 32.14 (2016), pp. 2233–2235. DOI: 10.1093/bioinformatics/btw216.

[270] B. Zhang and S. Horvath. "A General Framework for Weighted Gene Co-Expression Network Analysis". In: *Statistical Applications in Genetics and Molecular Biology* 4.1 (2005). DOI: 10.2202/1544-6115.1128. eprint: arXiv:1403.6652v2.

[271] P. Langfelder and S. Horvath. "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9.1 (Dec. 2008), p. 559. DOI: 10.1186/1471-2105-9-559.

[272] L. Song, P. Langfelder, and S. Horvath. "Comparison of co-expression measures: mutual information, correlation, and model based indices". In: *BMC Bioinformatics* 13.1 (Dec. 2012), p. 328. DOI: 10.1186/1471-2105-13-328.

[273] J. Schäfer and K. Strimmer. "An empirical Bayes approach to inferring large-scale gene association networks". In: *Bioinformatics* 21.6 (2005), pp. 754–764. DOI: 10.1093/bioinformatics/bti062. eprint: /oup/backfile/content_public/journal/bioinformatics/21/6/10.1093/bioinformatics/bti062/2/bti062.pdf.

[274] R. Opgen-Rhein and K. Strimmer. "From correlation to causation networks: A simple approximate learning algorithm and its application to high-dimensional plant gene expression data". In: *BMC Systems Biology* 1.1 (Aug. 2007), p. 37. DOI: 10.1186/1752-0509-1-37.

[275] S. L. Lauritzen. *Graphical models*. Vol. 17. Clarendon Press, 1996.

[276] J. Krumsiek, K. Suhre, A. M. Evans, M. W. Mitchell, R. P. Mohney, M. V. Milburn, B. Wägele, W. Römisch-Margl, T. Illig, J. Adamski, C. Gieger, F. J. Theis, and G. Kastenmüller. "Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information". In: *PLoS Genetics* 8.10 (Oct. 2012). Ed. by M. I. McCarthy, e1003005. DOI: 10.1371/journal.pgen.1003005.

[277] L. Breiman. "Bagging predictors". In: *Machine Learning* 24.2 (Aug. 1996), pp. 123–140. DOI: 10.1007/BF00058655.

[278] O. Banerjee, L. E. Ghaoui, and A. D'Aspremont. "Model Selection Through Sparse Maximum Likelihood Estimation". In: 9 (2007), pp. 485–516. DOI: 10.1093/rfs/hht062. arXiv: 0707.0704.

[279] L. Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[280] A. Hapfelmeier and K. Ulm. "A new variable selection approach using random forests". In: *Computational Statistics & Data Analysis* 60 (2013), pp. 50–69.

[281] N. Meinshausen and P. Bühlmann. "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (July 2010), pp. 417–473. DOI: 10.1111/j.1467-9868.2010.00740.x.

[282] A. Dobra and A. Lenkoski. "Copula Gaussian graphical models and their application to modeling functional disability data". In: *Ann. Appl. Stat.* 5.2 A (2011), pp. 969–993. DOI: 10.1214/10-AOAS397. arXiv: 1108.1680.

[283] A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D1005–D1012. DOI: 10.1093/nar/gky1120.

[284] M. Arnold, J. Raffler, A. Pfeufer, K. Suhre, and G. Kastenmüller. "SNiPA: An interactive, genetic variant-centered annotation browser". In: *Bioinformatics* 31.8 (2015), pp. 1334–1336. DOI: 10.1093/bioinformatics/btu779.

[285] K. Watanabe, S. Stringer, O. Frei, M. Umićević Mirkov, C. de Leeuw, T. J. Polderman, S. van der Sluis, O. A. Andreassen, B. M. Neale, and D. Posthuma. "A global overview of pleiotropy and genetic architecture in complex traits". In: *Nature Genetics* 51.9 (2019), pp. 1339–1348. DOI: 10.1038/s41588-019-0481-0.

[286] X. Wen, Y. Lee, F. Luca, and R. Pique-Regi. "Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors". In: *American Journal of Human Genetics* 98.6 (2016), pp. 1114–1129. DOI: 10.1016/j.ajhg.2016.03.029.

[287] Y. Lee, F. Luca, R. Pique-Regi, and X. Wen. "Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics". In: *bioRxiv* (2018), pp. 1–46. DOI: 10.1101/316471.

[288] X. Wen. "Effective QTL Discovery Incorporating Genomic Annotations". In: *bioRxiv* (2015), p. 032003. DOI: 10.1101/032003.

[289] T. Berisa and J. K. Pickrell. "Approximately independent linkage disequilibrium blocks in human populations". In: *Bioinformatics* 32.2 (2016), pp. 283–285. DOI: 10.1093/bioinformatics/btv546.

[290] F. S. Goes, J. McGrath, D. Avramopoulos, P. Wolyniec, M. Pirooznia, I. Ruczinski, G. Nestadt, E. E. Kenny, V. Vacic, I. Peters, T. Lencz, A. Darvasi, J. G. Mulle, S. T. Warren, and A. E. Pulver. "Genome-wide association study of schizophrenia in Ashkenazi Jews". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 168.8 (Dec. 2015), pp. 649–659. DOI: 10.1002/ajmg.b.32349.

[291] S. de Jong, K. R. van Eijk, D. W. L. H. Zeegers, E. Strengman, E. Janson, J. H. Veldink, L. H. van den Berg, W. Cahn, R. S. Kahn, M. P. M. Boks, R. A. Ophoff, and T. P. S. (PGC Schizophrenia (GWAS) Consortium. "Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes." In: *European journal of human genetics : EJHG* 20.9 (Sept. 2012), pp. 1004–8. DOI: 10.1038/ejhg.2012.38.

[292] T. Kanazawa, C. A. Bousman, C. Liu, and I. P. Everall. "Schizophrenia genetics in the genome-wide era: A review of Japanese studies". In: *npj Schizophrenia* 3.1 (2017), pp. 2–7. DOI: 10.1038/s41537-017-0028-2.

[293] T. Saito, M. Ikeda, T. Mushiroda, T. Ozeki, K. Kondo, A. Shimasaki, K. Kawase, S. Hashimoto, H. Yamamori, Y. Yasuda, M. Fujimoto, K. Ohi, M. Takeda, Y. Kamatani, S. Numata, T. Ohmori, S. ichi Ueno, M. Makinodan, Y. Nishihata, M. Kubota, T. Kimura, N. Kanahara, N. Hashimoto, K. Fujita, K. Nemoto, T. Fukao, T. Suwa, T. Noda, Y. Yada, M. Takaki, N. Kida, T. Otsuru, M. Murakami, A. Takahashi, M. Kubo, R. Hashimoto, and N. Iwata. "Pharmacogenomic Study of Clozapine-Induced Agranulocytosis/Granulocytopenia in a Japanese Population". In: *Biological Psychiatry* 80.8 (2016), pp. 636–642. DOI: 10.1016/j.biopsych.2015.12.006.

[294] J. Rustenhoven, A. M. Smith, L. C. Smyth, D. Jansson, E. L. Scotter, M. E. V. Swanson, M. Aalderink, N. Coppieters, P. Narayan, R. Handley, C. Overall, T. I. H. Park, P. Schweder, P. Heppner, M. A. Curtis, R. L. M. Faull, and M. Dragunow. "PU.1 regulates Alzheimer's disease-associated genes in primary human microglia." In: *Molecular neurodegeneration* 13.1 (2018), p. 44. DOI: 10.1186/s13024-018-0277-1.

[295] Z. Hu, X. Gu, K. Baraoidan, V. Ibanez, A. Sharma, S. Kadkol, R. Munker, S. Ackerman, G. Nucifora, and Y. Sauntharajah. "RUNX1 regulates corepressor

interactions of PU.1." In: *Blood* 117.24 (June 2011), pp. 6498–508. DOI: 10.1182/blood-2010-10-312512.

[296] Y. Watanabe, A. Nunokawa, N. Kaneko, T. Muratake, T. Arinami, H. Ujike, T. Inada, N. Iwata, H. Kunugi, M. Itokawa, T. Otowa, N. Ozaki, and T. Someya. "Two-stage case–control association study of polymorphisms in rheumatoid arthritis susceptibility genes with schizophrenia". In: *Journal of Human Genetics* 54.1 (Jan. 2009), pp. 62–65. DOI: 10.1038/jhg.2008.4.

[297] T. S. P. G.-W. A. S. (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. "Genome-wide association study identifies five new schizophrenia loci." In: *Nature genetics* 43.10 (Sept. 2011), pp. 969–76. DOI: 10.1038/ng.940.

[298] J. Shi, D. F. Levinson, J. Duan, A. R. Sanders, Y. Zheng, I. Pe'er, F. Dudbridge, P. A. Holmans, A. S. Whittemore, B. J. Mowry, A. Olincy, F. Amin, C. R. Cloninger, J. M. Silverman, N. G. Buccola, W. F. Byerley, D. W. Black, R. R. Crowe, J. R. Oksenberg, D. B. Mirel, K. S. Kendler, R. Freedman, and P. V. Gejman. "Common variants on chromosome 6p22.1 are associated with schizophrenia." In: *Nature* 460.7256 (Aug. 2009), pp. 753–7. DOI: 10.1038/nature08192.

[299] I. S. International Schizophrenia Consortium, S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, and P. Sklar. "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." In: *Nature* 460.7256 (Aug. 2009), pp. 748–52. DOI: 10.1038/nature08185.

[300] H. Stefansson, R. A. Ophoff, S. Steinberg, O. A. Andreassen, S. Cichon, D. Rujescu, T. Werge, O. P. H. Pietiläinen, O. Mors, P. B. Mortensen, E. Sigurdsson, O. Gustafsson, M. Nyegaard, A. Tuulio-Henriksson, A. Ingason, T. Hansen, J. Suvisaari, J. Lonnqvist, T. Paunio, A. D. Børglum, A. Hartmann, A. Fink-Jensen, M. Nordentoft, D. Hougaard, B. Norgaard-Pedersen, Y. Böttcher, J. Olesen, R. Breuer, H.-J. Möller, I. Giegling, H. B. Rasmussen, S. Timm, M. Mattheisen, I. Bitter, J. M. Réthelyi, B. B. Magnusdottir, T. Sigmundsson, P. Olason, G. Masson, J. R. Gulcher, M. Haraldsson, R. Fossdal, T. E. Thorgeirsson, U. Thorsteinsdottir, M. Ruggeri, S. Tosato, B. Franke, E. Strengman, L. A. Kiemeney, Genetic Risk and Outcome in Psychosis (GROUP), I. Melle, S. Djurovic, L. Abramova, V. Kaleda, J. Sanjuan, R. de Frutos, E. Bramon, E. Vassos, G. Fraser, U. Ettinger, M. Picchioni, N. Walker, T. Toulopoulou, A. C. Need, D. Ge, J. L. Yoon, K. V. Shianna, N. B. Freimer, R. M. Cantor, R. Murray, A. Kong, V. Golimbet, A. Carracedo, C. Arango, J. Costas, E. G. Jönsson, L. Terenius, I. Agartz, H. Petursson, M. M. Nöthen, M. Rietschel, P. M. Matthews, P. Muglia, L. Peltonen, D. St Clair, D. B. Goldstein, K. Stefansson, and D. A. Collier. "Common variants conferring risk of schizophrenia." In: *Nature* 460.7256 (Aug. 2009), pp. 744–7. DOI: 10.1038/nature08186.

[301] B. B. Quednow, J. Brinkmeyer, A. Mobascher, M. Nothnagel, F. Musso, G. Gründer, N. Savary, N. Petrovsky, I. Frommann, L. Lennertz, K. N. Spreckelmeyer, T. F. Wienker, N. Dahmen, N. Thuerauf, M. Clepce, F. Kiefer, T. Majic, R. Mössner, W. Maier, J. Gallinat, A. Diaz-Lacava, M. R. Toliat, H. Thiele, P. Nürnberg, M. Wagner,

and G. Winterer. "Schizophrenia risk polymorphisms in the TCF4 gene interact with smoking in the modulation of auditory sensory gating." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.16 (Apr. 2012), pp. 6271–6. DOI: 10.1073/pnas.1118051109.

[302] C. Zweier, M. M. Peippo, J. Hoyer, S. Sousa, A. Bottani, J. Clayton-Smith, W. Reardon, J. Saraiva, A. Cabral, I. Göhring, K. Devriendt, T. de Ravel, E. K. Bijlsma, R. C. Hennekam, A. Orrico, M. Cohen, A. Dreweke, A. Reis, P. Nürnberg, and A. Rauch. "Haploinsufficiency of TCF4 Causes Syndromal Mental Retardation with Intermittent Hyperventilation (Pitt-Hopkins Syndrome)". In: *The American Journal of Human Genetics* 80.5 (May 2007), pp. 994–1001. DOI: 10.1086/515583.

[303] M. Jung, B. M. Häberle, T. Tschaikowsky, M.-T. Wittmann, E.-A. Balta, V.-C. Stadler, C. Zweier, A. Dörfler, C. J. Gloeckner, and D. C. Lie. "Analysis of the expression pattern of the schizophrenia-risk and intellectual disability gene TCF4 in the developing and adult brain suggests a role in development and plasticity of cortical and hippocampal neurons." In: *Molecular autism* 9 (2018), p. 20. DOI: 10.1186/s13229-018-0200-1.

[304] Y. Huo, S. Li, J. Liu, X. Li, and X.-J. Luo. "Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk". In: *Nature Communications* 10.1 (Dec. 2019), p. 670. DOI: 10.1038/s41467-019-08666-4.

[305] P. Roussos, P. Katsel, K. L. Davis, S. G. Giakoumaki, T. Lencz, A. K. Malhotra, L. J. Siever, P. Bitsios, and V. Haroutunian. "Convergent findings for abnormalities of the NF-κB signaling pathway in schizophrenia." In: *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 38.3 (Feb. 2013), pp. 533–9. DOI: 10.1038/npp.2012.215.

[306] E. Bagyinszky, Y. C. Youn, S. S. A. An, and S. Kim. "The genetics of Alzheimer's disease." In: *Clinical interventions in aging* 9 (2014), pp. 535–51. DOI: 10.2147/CIA.S51571.

[307] W. E. Dowdle, J. F. Robinson, A. Kneist, M. S. Sirerol-Piquer, S. G. Frints, K. C. Corbit, N. A. Zaghloul, G. van Lijnschoten, L. Mulders, D. E. Verver, K. Zerres, R. R. Reed, T. Attié-Bitach, C. A. Johnson, J. M. García-Verdugo, N. Katsanis, C. Bergmann, J. F. Reiter, and J. F. Reiter. "Disruption of a Ciliary B9 Protein Complex Causes Meckel Syndrome". In: *The American Journal of Human Genetics* 89.1 (July 2011), pp. 94–110. DOI: 10.1016/j.ajhg.2011.06.003.

[308] E. Sanchez, H. Darvish, R. Mesias, S. Taghavi, S. G. Firouzabadi, R. H. Walker, A. Tafakhori, and C. Paisán-Ruiz. "Identification of a Large DNAJB2 Deletion in a Family with Spinal Muscular Atrophy and Parkinsonism." In: *Human mutation* 37.11 (2016), pp. 1180–1189. DOI: 10.1002/humu.23055.

[309] M. J. Stuart, G. Singhal, and B. T. Baune. "Systematic review of the neurobiological relevance of chemokines to psychiatric disorders". In: *Frontiers in Cellular Neuroscience* 9.September (2015), pp. 1–15. DOI: 10.3389/fncel.2015.00357.

[310] V. M. Saia-Cereda, J. S. Cassoli, A. Schmitt, P. Falkai, J. M. Nascimento, and D. Martins-de-Souza. "Proteomics of the corpus callosum unravel pivotal players in the dysfunction of cell signaling, structure, and myelination in schizophrenia brains". In: *European Archives of Psychiatry and Clinical Neuroscience* 265.7 (Oct. 2015), pp. 601–612. DOI: 10.1007/s00406-015-0621-1.

[311] M. S. Rodriguez, I. Egaña, F. Lopitz-Otsoa, F. Aillet, M. P. Lopez-Mato, A. Dorronroso, S. Lobato-Gil, J. D. Sutherland, R. Barrio, C. Trigueros, and V. Lang. "The RING ubiquitin E3 RNF114 interacts with A20 and modulates NF-κB activity and T-cell activation". In: *Cell Death and Disease* 5.8 (2014). DOI: 10.1038/cddis.2014.366.

[312] A. F. Pardiñas, P. Holmans, A. J. Pocklington, et al. "Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection". In: *Nature Genetics* 50.3 (2018), pp. 381–389. DOI: 10.1038/s41588-018-0059-2.

[313] A. N. Singh and B. Gasman. "Disentangling the genetics of sarcopenia: prioritization of NUDT3 and KLF5 as genes for lean mass & HLA-DQB1-AS1 for hand grip strength with the associated enhancing SNPs & a scoring system". In: *BMC Medical Genetics* 21.1 (Dec. 2020), p. 40. DOI: 10.1186/s12881-020-0977-6.

[314] Y. Oishi, I. Manabe, K. Tobe, M. Ohsugi, T. Kubota, K. Fujiu, K. Maemura, N. Kubota, T. Kadowaki, and R. Nagai. "SUMOylation of Krüppel-like transcription factor 5 acts as a molecular switch in transcriptional programs of lipid metabolism involving PPAR-δ". In: *Nature Medicine* 14.6 (June 2008), pp. 656–666. DOI: 10.1038/nm1756.

[315] V. Moresi, M. Carrer, C. E. Grueter, O. F. Rifki, J. M. Shelton, J. A. Richardson, R. Bassel-Duby, and E. N. Olson. "Histone deacetylases 1 and 2 regulate autophagy flux and skeletal muscle homeostasis in mice." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.5 (Jan. 2012), pp. 1649–54. DOI: 10.1073/pnas.1121159109.

[316] R. A. Silverstein and K. Ekwall. "Sin3: a flexible regulator of global gene expression and genome stability". In: *Current Genetics* 47.1 (Jan. 2005), pp. 1–17. DOI: 10.1007/s00294-004-0541-5.

[317] T. I. Lee and R. A. Young. "Transcription of Eukaryotic Protein-Coding Genes". In: *Annual Review of Genetics* 34.1 (Dec. 2000), pp. 77–137. DOI: 10.1146/annurev.genet.34.1.77.

[318] G. Wang, S. Padmanabhan, E. Miyamoto-Mikami, N. Fuku, M. Tanaka, M. Miyachi, H. Murakami, Y.-C. Cheng, B. D. Mitchell, K. G. Austin, and Y. P. Pitsiladis. "GWAS of Elite Jamaican, African American and Japanese Sprint Athletes: 2254 May 30, 945 AM - 1000 AM". In: *Medicine & Science in Sports & Exercise* 46.5S (2014).

[319] J. Zhang, M.-L. Bang, D. S. Gokhin, Y. Lu, L. Cui, X. Li, Y. Gu, N. D. Dalton, M. C. Scimia, K. L. Peterson, R. L. Lieber, and J. Chen. "Syncoilin is required for generating maximum isometric stress in skeletal muscle but dispensable for muscle cytoarchitecture." In: *American journal of physiology. Cell physiology* 294.5 (May 2008), pp. C1175–82. DOI: 10.1152/ajpcell.00049.2008.

[320] S. C. Brown, S. Torelli, I. Ugo, F. De Biasia, E. V. Howman, E. Poon, J. Britton, K. E. Davies, and F. Muntoni. "Syncoilin upregulation in muscle of patients with neuromuscular disease". In: *Muscle & Nerve* 32.6 (Dec. 2005), pp. 715–725. DOI: 10.1002/mus.20431.

[321] I. Seim, P. L. Jeffery, and L. K. Chopin. "Gene expression profiling of The Cancer Genome Atlas supports an inverse association between body mass index (BMI) and major oesophageal tumour subtypes". In: *bioRxiv* (Sept. 2018), p. 378778. DOI: 10.1101/378778.

[322] K. Oldknow, N. M. Morton, M. Yadav, S. Rajoanah, C. Huesa, L. Bunger, M. Ferron, G. Karsenty, V. MacRae, J. L. Milan, and C. Farquharson. "An emerging role of phospho1 in the regulation of energy metabolism". In: *Bone Abstracts* (May 2013). DOI: 10.1530/boneabs.01.OC6.6.

[323] 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. DOI: 10.1038/nature20784.Epigenome-wide.

[324] A. Pietrobelli, R. C. Lee, E. Capristo, R. J. Deckelbaum, and S. B. Heymsfield. "An independent, inverse association of high-density-lipoprotein-cholesterol concentration with nonadipose body mass". In: *The American Journal of Clinical Nutrition* 69.4 (Apr. 1999), pp. 614–620. DOI: 10.1093/ajcn/69.4.614.

[325] A. F. McRae, R. E. Marioni, S. Shah, J. Yang, J. E. Powell, S. E. Harris, J. Gibson, A. K. Henders, L. Bowdler, J. N. Painter, L. Murphy, N. G. Martin, J. M. Starr, N. R. Wray, I. J. Deary, P. M. Visscher, and G. W. Montgomery. "Identification of 55,000 Replicated DNA Methylation QTL". In: *Scientific Reports* 8.1 (Dec. 2018), p. 17605. DOI: 10.1038/s41598-018-35871-w.

[326] C. Marchal and B. Miotto. "Emerging concept in DNA methylation: Role of transcription factors in shaping DNA methylation patterns". In: *Journal of Cellular Physiology* 230.4 (2015), pp. 743–751. DOI: 10.1002/jcp.24836.

[327] M. Bartkuhn and R. Renkawitz. "Long range chromatin interactions involved in gene regulation". In: *Biochimica et Biophysica Acta - Molecular Cell Research* 1783.11 (2008), pp. 2161–2166. DOI: 10.1016/j.bbamcr.2008.07.011.

[328] A. K. Smith, V. Kilaru, M. Kocak, L. M. Almli, K. B. Mercer, K. J. Ressler, F. A. Tylavsky, and K. N. Conneely. "Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type". In: *BMC Genomics* 15.1 (Feb. 2014), p. 145. DOI: 10.1186/1471-2164-15-145.

[329] C. Cai, P. Langfelder, T. F. Fuller, M. C. Oldham, R. Luo, L. H. van den Berg, R. A. Ophoff, and S. Horvath. "Is human blood a good surrogate for brain tissue in transcriptional studies?" In: *BMC Genomics* 11 (2010), p. 589. DOI: 10.1186/1471-2164-11-589.

[330] H. A. Pliner, J. S. Packer, J. L. McFaline-Figueroa, D. A. Cusanovich, R. M. Daza, D. Aghamirzaie, S. Srivatsan, X. Qiu, D. Jackson, A. Minkina, A. C. Adey, F. J. Steemers, J. Shendure, and C. Trapnell. "Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data". In: *Molecular Cell* 71.5 (Sept. 2018), 858–871.e8. DOI: 10.1016/J.MOLCEL.2018.06.044.

[331] A. Ocone, L. Haghverdi, N. Mueller, and F. Theis. "Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data." In: *Bioinformatics (Oxford, England)* 31.12 (June 2015), pp. i89–96. DOI: 10.1093/bioinformatics/btv257.

[332] S. J. Clark, R. Argelaguet, C. A. Kapourani, T. M. Stubbs, H. J. Lee, C. Alda-Catalinas, F. Krueger, G. Sanguinetti, G. Kelsey, J. C. Marioni, O. Stegle, and W. Reik. "scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells". In: *Nat Commun* 9.1 (Feb. 2018), p. 781.

[333] J. Cao, D. A. Cusanovich, V. Ramani, D. Aghamirzaie, H. A. Pliner, A. J. Hill, R. M. Daza, J. L. McFaline-Figueroa, J. S. Packer, L. Christiansen, F. J. Steemers, A. C. Adey, C. Trapnell, and J. Shendure. "Joint profiling of chromatin accessibility and gene expression in thousands of single cells". In: *Science* 361.6409 (Sept. 2018), pp. 1380–1385. DOI: 10.1126/science.aau0730.

[334] L. Liu, C. Liu, A. Quintero, L. Wu, Y. Yuan, M. Wang, M. Cheng, L. Leng, L. Xu, G. Dong, R. Li, Y. Liu, X. Wei, J. Xu, X. Chen, H. Lu, D. Chen, Q. Wang, Q. Zhou, X. Lin, G. Li, S. Liu, Q. Wang, H. Wang, J. L. Fink, Z. Gao, X. Liu, Y. Hou, S. Zhu, H. Yang, Y. Ye, G. Lin, F. Chen, C. Herrmann, R. Eils, Z. Shang, and X. Xu. "Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity". In: *Nature Communications* 10.1 (Jan. 2019), p. 470. DOI: 10.1038/s41467-018-08205-7.

[335] V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, S.-I. Nishikawa, N. Piterman, V. Kouskoff, F. J. Theis, J. Fisher, and B. Göttgens. "Decoding the regulatory network of early blood development from single-cell gene expression measurements". In: *Nature Biotechnology* 33.3 (Mar. 2015), pp. 269–276. DOI: 10.1038/nbt.3154.

[336] M. Colomé-Tatché and F. J. Theis. "Statistical single cell multi-omics integration". In: *Curr. Opin. Syst. Biol.* 7 (2018), pp. 54–59. DOI: 10.1016/j.coisb.2018.01.003.

[337] A. Lyons and C. R. Parish. "Determination of lymphocyte division by flow cytometry". In: *Journal of Immunological Methods* 171.1 (May 1994), pp. 131–137. DOI: 10.1016/0022-1759(94)90236-4.

[338] P. J. Park. "ChIP-seq: advantages and challenges of a maturing technology." In: *Nature reviews. Genetics* 10.10 (2009), pp. 669–80. DOI: 10.1038/nrg2641. eprint: NIHMS150003.

[339] K. L. Spalding, E. Arner, P. O. Westermark, S. Bernard, B. A. Buchholz, O. Bergmann, L. Blomqvist, J. Hoffstedt, E. Näslund, T. Britton, H. Concha, M. Hassan, M. Rydén, J. Frisén, and P. Arner. "Dynamics of fat cell turnover in humans". In: *Nature* 453.7196 (June 2008), pp. 783–787. DOI: 10.1038/nature06902.

[340] J. Frik, M.-P. J, P. N, M. N, K. J, K. J, G. RM, H. SM, S. S, and G. M. "Cross-talk Between Monocyte Invasion and Astrocyte Proliferation Regulates Scarring in Brain Injury". In: *EMBO reports* 19.5 (2018). DOI: 10.15252/EMBR.201745294.