From correlation to causation: Data-driven exploration of transcriptional regulation using population genomics

Luijk, R.

From correlation to causation
Data-driven exploration of transcriptional regulation using population genomics

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus Prof. dr. C. J. J. M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 16 oktober 2019
klokke 13:45 uur

door
René Luijk
geboren te Leiden
in 1988

Promotor:	Prof. dr. P. Slagboom	
Co-promotoren:	Dr. B.T. Heijmans	
	Dr. E.W. van Zwet	
Leden promotiecommissie:	Prof. Dr. D. I. Boomsma	Vrije Universiteit Amsterdam
	Prof. Dr. S. C. Cannegieter	
	Prof. Dr. J. J. Goeman	
	Prof. Dr. J. H. Gribnau	Erasmus MC

From correlation to causation
Data-driven exploration of transcriptional regulation using population genomics
R. Luijk, MSc
ISBN: 978-94-6380-536-0

De aanhouder wint

Table of Contents

1 | Introduction

Molecular epidemiology

Epidemiology refers to the study of the distribution of health and disease conditions, their determinants, and their risk factors in variously defined populations. Molecular epidemiology is a subfield of epidemiology that is particularly interested in how changes at the molecular level contribute to these biological traits and disease susceptibilities. Such information is used to predict the individual disease risk in the population, predict the prognosis of patients, or to monitor the effect of interventions in a biomedical or clinical setting. The information gathered at the molecular level includes, among others, genetic variation (inter-individual differences in the DNA sequence), transcriptomic variation (in the expression levels of genes), epigenomic variation (changes in the function of DNA without changing the DNA sequence itself, *e.g.*, through modification of the accessibility of the DNA), metabolomic variation (in the levels of metabolites) and proteomics (in levels and activity of proteins), collectively referred to as omics. Driven by recent technological advances, researchers are able to routinely measure such molecular phenotypes on a genome-wide scale in large numbers of study participants, providing a detailed view of a person's full genomic profile.

The molecular epidemiology field particularly investigates common and complex diseases, the development of which is influenced by multiple genes and environmental risk factors. In fact, twin-and family-based studies have shown that many common traits and diseases are influenced by a significant genetic component, in addition to being affected by environmental factors [Thomas, 2010]. However, these specific types of studies have generally been unable to pinpoint which specific locations on the genome are responsible for a phenotypic trait (reviewed in Botstein and Risch [2003]). Genetic studies have subsequently identified large numbers of specific genetic variants in the genome that associate with quantitative traits, such as serum cholesterol levels, blood pressure, and a range of disease conditions (reviewed in Visscher et al. [2012]). However, the understanding of how the variants identified by these Genome Wide Association Studies (GWAS) affect the trait or onset of disease is often hampered, as the majority (over 90%) of the strongest associated variants do not directly affect the production of a protein [Hindorff et al., 2009]. Because of this, these variants are thought to affect a biological trait through effects on transcriptional regulation. This process determines the degree to which any gene in the genome is switched on or off in a particular cell through transcription from DNA to RNA. This process is orchestrated by transcription factors, an important collection of proteins that initiate transcription, which ultimately determines a phenotypic trait. Understanding transcriptional regulation is crucial, as its dysregulation often forms a key element of disease development [Lee and Young, 2013]. Hence, this thesis aims to better understand how genetic variation influences transcriptional regulation, which is the first step in understanding how genetic variants affect a phenotypic trait.

From correlation to causation

In addition to the aforementioned genetic studies, inter-individual differences in many different aspects of an individual's genomic profile (such as the transcriptome, methylome, and others) are routinely related to phenotypic traits. For example, changes in the expression of several genes could be related to the presence or absence of a specific cardiovascular disease (CVD), potentially providing insight into the genes contributing to CVD development. However, as epidemiological research is observational in nature, making any causal claims about an association between any of the (molecular) determinants and a phenotypic outcome is often impossible. For example, suppose that changes in the expression of a particular gene are indeed strongly correlated to the presence of a CVD in the general population (Figure 1.1). At face value, it may seem reasonable to conclude the changes in gene expression are one of the causal drivers behind developing this CVD, but several important factors hinder us from drawing such conclusions. Firstly, the causal relation could be the other way around, where the presence of a CVD itself induces changes in the expression levels of that gene, a phenomenon known as reverse causation. In addition, other factors could influence both the expression levels of the gene as well as the risk of developing a CVD, thereby introducing a correlation between them (Figure 1.1). This scenario is known as confounding, and is particularly problematic in epidemiological research, as confounding variables may not always be known. Failing to account for such confounding factors could then lead to the false conclusion of a causal relationship between the expression levels and lung cancer (Figure 1.1).

One way to circumvent these issues is by involving the DNA sequence in the analysis, which is formed at conception, and generally does not change over the course of a lifetime - with some exceptions. This implies genetic variation precedes all other phenotypic variation, so a phenotype could not have changed the DNA sequence itself (*i.e.*, reverse causation). As a result, genetic approaches are generally free from confounding or reverse causation, suggesting any association between genetic variation and a disease phenotype or quantitative trait must have started at the DNA sequence level. This principle has been widely applied by GWAS by mapping quantitative trait loci (QTL) through investigating a set of genetic variants distributed across all chromosomes. Using statistical hypothesis testing, each genetic variant is formally tested for an association with a phenotype. Investigating all possible correlations between a set of genetic variants and a phenotypic outcome enables the identification of individual genetic variants associated with that phenotype (reviewed by Visscher et al. [2012]). Given a certain statistical significance threshold, any genetic variant significantly associated with that phenotype is then a possible candidate for further research. Using this approach, GWAS have successfully uncovered numerous genetic loci associated with any of an increasing number of phenotypes studied (reviewed by Visscher et al. [2012, 2017]). Despite the properties of genetics virtually guaranteeing a directed relation between genotype and phenotype, GWAS do have their limitations.

The first drawback pertains to the interpretation of their results. As the

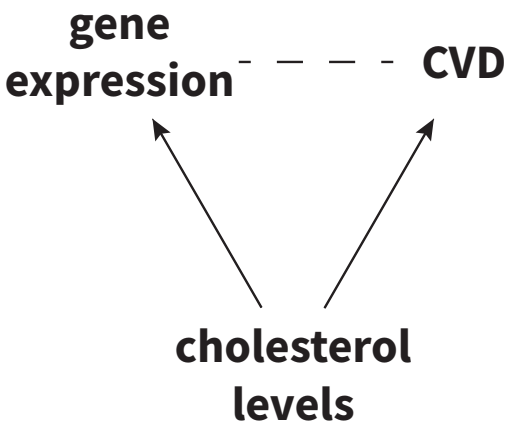

Figure 1.1: Depiction of the relationship between gene expression levels, a specific cardiovascular disease (CVD), and possible confounders, such as cholesterol levels. A correlation between gene expression and the CVD (dashed line) may lead to the conclusion of a possible causal relation between the two. However, a third, confounding variable, cholesterol levels, may causally influence both gene expression and the presence of the CVD (solid arrows), inducing a correlation between them without there being an actual causal relation (confounding).

sample sizes used in GWAS increase, providing more statistical power, so do the number of identified variants. For example, an early GWA study for height identified at least 180 independent variants [Lango Allen et al., 2010], while a recent study in roughly 700,000 individuals identified 3,290 independent genetic variants associated with height, an 18-fold increase in the number of genetic variants identified [Yengo et al., 2018]. Indeed, the number of statistically significant hits for several different phenotypes seem to be ever increasing, are located all over the genome, and are often located near genes with no apparent connection to the phenotype under study [Boyle et al., 2017]. This observation has led to the postulation that many complex phenotypes are not influenced by just several genes, but rather that these phenotypic traits are omnigenic, where almost all genes are in some way related to the phenotypic trait under study [Boyle et al., 2017]. Boyle *et al.* specifically suggest that while a set of core genes are responsible for most of the phenotypic variation, most other genes in the genome contribute as well, even if just marginally. Hence, there would be a limited return on investing more resources towards increasing sample sizes.

Moreover, a large proportion of the identified variants are located in non-coding regions [Visscher et al., 2012] - regions that do not encode for protein sequences, and as a result do not directly alter a protein's function in a cell. Therefore, it is assumed the identified genetic variants are involved in transcriptional (dys)regulation [Hindorff et al., 2009; Manolio, 2010; Edwards

et al., 2013], a key element of disease [Lee and Young, 2013]. This has sparked a field known as molecular quantitative trait loci (QTL) mapping, which uses similar methodology to GWAS, in order to explore the genetic underpinnings of transcriptional regulation. Instead of investigating a single phenotype, molecular QTL mapping investigates many different molecular phenotypes simultaneously. For example, investigating gene expression levels (Westra et al., 2013), DNA methylation levels [Bell et al., 2011], and histone modifications [Pelikan et al., 2018]. While this approach does provide a crucial first step towards better understanding how genetic variation influences transcriptional regulation, it does not allow for the next step: inter-relating different types of non-genetic omics data. For example, the interplay between the expression of different genes, or between genes and DNA methylation at CpG sites is still obstructed by confounding and reverse causation. Ideally, we would also be able to investigate these associations, and even posit causal hypotheses about such associations and the way they influence phenotypic endpoints.

Causality is not easily demonstrated, however, and hindered by the confounding factors mentioned earlier (Figure 1.1). In principle, experimental manipulation of one variable, while keeping everything else fixed, is necessary to prove the causal effect of that variable on another variable. Sometimes, it is possible to conduct lab experiments to achieve this, either by altering the expression of genes using different techniques. For example, using techniques involving the breakdown of RNA molecules, or by genome-editing using CRISPR-Cas to make targeted changes in the DNA sequence itself. Regrettably, using these techniques to investigate all the possible leads generated by omics-wide association studies is not always an option. Either the sheer number of possible genes to investigate makes it infeasible to do so, or ethical considerations prevent researchers from applying these experimental techniques in human subjects.

Fortunately, it is possible to use observational omics data to at least provide some evidence of causality. The aforementioned method of molecular quantitative trait loci (QTL) mapping may provide a starting point through its investigation of genetic effects on different molecular phenotypes, forming the basis for so-called Mendelian Randomization (MR)-type methods [Davey Smith and Hemani, 2014]. Given that several key assumptions are met [Burgess et al., 2016], MR makes it possible to find causal evidence for an association between two genes that would normally be plagued by confounding and reverse causation.

In this thesis, we aim to take a step towards finding such causal evidence regarding transcriptional regulation by applying different data-analytical approaches to several large-scale, population-based omics datasets. Specifically, we first systematically investigate the effects of genetic variation on gene expression and DNA methylation. Next, we use these results as a springboard to move beyond associations between genetics and gene expression and DNA methylation, and posit and evaluate causal hypotheses about underlying mechanisms and transcriptional networks.

Before we do so, we will first go into more detail about the concepts and

terminology necessary for the remainder of this thesis, starting with transcriptional regulation, and how the epigenome plays an important role in that process. Next, we discuss how molecular QTL mapping could be used as a first step in investigating the genetic underpinnings of transcriptional regulation, followed by a description of Mendelian Randomization-type methods aiming to move beyond the associations generated by molecular QTL mapping. Lastly, we underscore the importance of big data in all of these methods, and how combined efforts play a key role in this.

Transcriptional regulation

Transcriptional regulation describes how sequences of DNA are transcribed into RNA, which are subsequently translated into the proteins that perform a wide variety of functions in the cell, and is imperative in keeping an organism healthy [Lee and Young, 2013]. Transcription is accomplished by RNA polymerases, enzymes moving along the DNA, translating it into RNA. While several types of RNA molecules exist, many are the product of protein-coding genes, yielding messenger RNA (mRNA). Another key element in transcription are so-called transcription factors, proteins that orchestrate transcriptional regulation by binding to specific DNA sequences (motifs) within *cis*-regulatory regions (close to the gene of interest, *e.g.*, within the target gene promoter, Figure 1.2), initiating transcription. The ability of a transcription factor to influence the transcriptional activity of the target gene depends on several aspects. For example, genetic variation could change the motif, altering the binding potential of the transcription factor. Alternatively, epigenomic modifications may also influence binding by altering the accessibility of the chromatin to RNA polymerases through DNA methylation or histone modifications, either enhancing or decreasing binding potential [Hu et al., 2013].

Epigenome

The epigenome encompasses different molecular components, including histone modifications, and non-coding RNAs, all influencing transcription in a different way (Figure 1.2). In this thesis, we focus on DNA methylation, a widely studied and well-characterized epigenomic process. The relatively stable nature of methylation over time, and the advancements in high-throughput arrays targeting methylation make it a useful mark to study. In mammals, the most common form of DNA methylation is the addition of a specific molecule, a methyl group, to a single base of the DNA, a cytosine in the dinucleotide CG. Commonly referred to as CpG, where the 'p' represents the phosphate backbone of the DNA, an estimated 60% to 80% of the roughly 28 million CpGs that are part of the human genome are methylated, typically repressing gene expression [Smith and Meissner, 2013].

DNA methylation plays a vital role in development [Smith and Meissner, 2013] and is further implicated in several well-known processes, including genomic imprinting [Ferguson-Smith, 2011], carcinogenesis [Laird, 2005]. Most

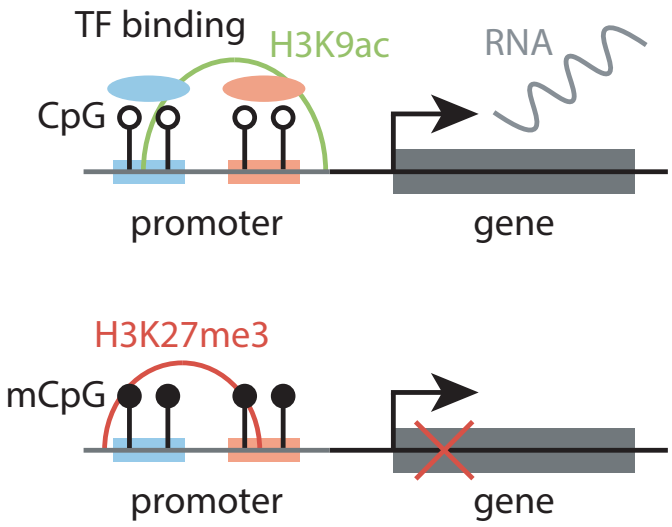

Figure 1.2: Simplified schematic overview of the interplay between epigenomic marks, transcription factor binding, and transcription. In the top panel, unmethylated CpGs and active histone marks (histone mark H3K9ac, in green) typically allow for the binding of transcription factors on their respective binding sites (indicated by blue and red) in the promoter of the target gene. This initiates transcription, where a RNA molecule is formed. Methylated CpGs and repressive histone marks, however, change the structure of the chromatin, packing it tightly. This makes the DNA inaccessible for transcription factors, not allowing them to bind. As a result, the target gene is not expressed.

notably, DNA methylation is a key element of X-chromosome inactivation (XCI; Yasukochi et al. [2010]; Sharp et al. [2011]), a process where one of two copies of the X-chromosome present in females is silenced [Lyon, 1961]. XCI can be considered a paradigm to study epigenomic regulation [Morey and Avner, 2010] through the identification of autosomal genetic variants affecting X-chromosome DNA methylation.

Methylation has traditionally been described as a repressive mechanism [Kass et al., 1997; Miranda and Jones, 2007], where hypermethylation of mostly CpG dense regions (CpG islands), often located in gene promoters [Illingworth et al., 2010], are associated with repressed chromatin and transcriptional repression [Smith and Meissner, 2013]. However, the relation between methylation and transcriptional activity seems to be more complex than previously appreciated [Battle et al., 2014], as a positive association between the two appear to be fairly common [Gutierrez-Arcelus et al., 2013].

Lastly, while DNA methylation does not alter the DNA itself, methylation levels are indeed dependent on sequence variation [Heijmans et al., 2007; Bell et al., 2011]. Identifying genetic variants associated with DNA methylation is therefore an often-used strategy to investigate the genetic underpinnings of transcriptional regulation, often referred to as molecular quantitative trait loci (QTL) mapping, and used throughout this thesis.

Starting from the bottom: using genetics to investigate transcriptional regulation

As mentioned earlier, the generally immutable character of DNA [Belshaw et al., 2004] virtually guarantees that genetic variation is the causal driver behind an association with any other molecular phenotype, such as DNA methylation. Hence, if one aims to understand the molecular mechanisms behind transcriptional regulation, identifying genetic loci associated with other molecular phenotypes pertaining to transcriptional regulation (*e.g.*, DNA methylation or gene expression) provides a good starting point. Molecular quantitative trait loci (QTL) mapping is an often-used method to achieve this, and is one of the strategies used in this thesis.

Design and objectives

The study design of methylation QTL (meQTL) and expression QTL (eQTL) mapping in the current context is similar to that of a genome-wide association study focused on identifying disease susceptibility or quantitative trait loci (which I will call GWAS here), where a large number of unrelated individuals are genotyped, and each measured genetic variant across the whole genome is tested separately for an association with a molecular trait. Within molecular QTL mapping, a distinction is often made between *cis* (local, typically within 250Kb to 1Mb), and trans (distal, usually outside 1Mb) molecular QTL mapping, where only genetic variants local or distal to a gene or CpG site are taken into consideration.

Its objectives, however, are different from a GWAS, as molecular QTL mapping is often utilized to investigate two distinct issues. The first is to support the interpretation of genetic variants identified by GWAS. Relating these genetic variants to the expression of nearby or distant genes or methylation of relevant CpG sites possibly yields a better understanding of the underlying mechanisms leading to the phenotype under study. The other objective of molecular QTL mapping is more basic, and relates to the fundamental question of the contexts in which genetics influence expression and methylation levels, providing a general understanding of the genetic influences on transcriptional regulation. By context we mean, for example, the tissue-specificity of genome regulation, or its response to environmental stimuli.

Early molecular QTL studies

While early molecular QTL studies have related genetic variation to gene expression, these were often hypothesis-driven studies investigating one, or a limited set of specific effects [Heijmans et al., 2007]. With the advent of array-based technologies, genome-wide, hypothesis-free eQTL and meQTL studies became feasible (*e.g.*, Bell et al. [2011]). However, testing all genetic variants for an association with all measured gene expression or methylation at CpG sites results in a massively increased multiple testing burden, also compared to a GWAS for a single phenotype, requiring increasingly larger sample sizes to overcome. Despite the technological advancements, early studies often employed relatively small sample sizes, limiting the statistical power. As a result, early studies limited the number of tests performed by restricting the analysis to genetic variants and genes or CpGs in close proximity (*in cis*), usually within 1Mb of either a gene's transcription start site or a CpG. These studies showed great potential, as many genes and CpGs were found to harbor *cis*-eQTLs and *cis*-meQTLs [Bell et al., 2011; Grundberg et al., 2012; Westra et al., 2013; Shi et al., 2014]. As array technologies became cheaper, individual large cohorts and meta-analyses allowed for the investigation of long-range (*trans*) effects, usually farther than 1 Mb [Westra et al., 2013; Lemire et al., 2015]. In addition, later studies showed eQTLs and meQTLs may have tissue-, and population-specific effects [Grundberg et al., 2012; Smith et al., 2014; GTEx Consortium, 2017; Yang et al., 2017]. This finding is particularly relevant for diseases, which may develop in specific tissues [Nica and Dermitzakis, 2008].

Limitations

Despite these early successes and the utility of molecular QTL mapping, molecular QTL studies do suffer from a number of limitations, some of which are similar to those in GWAS. As briefly mentioned above, early molecular QTL studies were plagued by smaller sample sizes, restricting statistical power. While technological advancements have made it possible to utilize larger sample sizes, they are also responsible for the increasing dimensionality of data. For example, commonly used methylation arrays have gone from investigating 27,000 CpGs to over

850,000 CpGs, a 31.5-fold increase in the number of CpGs interrogated. Next-generation sequencing will only add to the number of variables measured.

In addition, the data itself are often influenced by undesirable variation due to technical factors, as well as usually not well understood variation due to biological and environmental context. For example, measuring the subjects in different batches introduces systematic differences in the measurements, adding noise to the data, and possibly even confounding the analysis [Buhule et al., 2014; van Iterson et al., 2017]. Biological influences of poorly understood measurable variation include those resulting from differences in age [Garagnani et al., 2012] or sex [McCarthy et al., 2009; Hall et al., 2014] or smoking [Zeilinger et al., 2013], while environmental factors or lifestyle differences often add poorly understood and unmeasured variation, *e.g.*, caused by diet [Heijmans et al., 2008].

Lastly, linkage disequilibrium (LD) causes many neighboring genetic variants to all be associated to the same gene expression or CpG site methylation levels. Pinpointing the causal variant amidst strongly correlated, but non-causal variants is therefore challenging, and even impossible on the basis of statistical evidence alone. To complicate things even further, a genetic variant may be associated with multiple nearby genes or CpGs [Bell et al., 2011], hampering the annotation of this variant with a single gene.

Beyond molecular QTL mapping: moving towards causality

Molecular QTL mapping is an important first step in investigating how genetic variation influences transcriptional regulation, as it aids in the interpretation of GWA studies, and is a precursor to Mendelian Randomization-type approaches. However, molecular QTL mapping itself does not go beyond the link between genetic variation and other molecular phenotypes. Inter-relating different omics data – *e.g.*, gene expression data - may provide additional information on gene function, regulatory networks [Stuart et al., 2003; de la Fuente, 2010], and its dysregulation in disease [Lee and Young, 2013]. Ideally, one would be able to explore how a genetic variant affects a phenotype through different omics layers, such as through a gene network. Similar to the lung cancer example (Figure 1.1), confounding factors induce strong correlations among the gene expression levels, despite the possible lack of a causal relationship between them. In addition, they do not indicate any directionality of the association.

In the face of these limitations, Mendelian Randomization (MR) provides an approach that uses genetics as a causal anchor to investigate causal hypotheses [Davey Smith and Hemani, 2014]. The fundamental idea behind MR is that genetic variation can mimic the effects of an exposure. The analogy with a treatment in a clinical randomized controlled trial (RCT) often helps, where subjects are randomized over different treatment arms by the researcher. This randomization ensures any association between the treatment and a clinical outcome is not confounded. As alleles are passed down randomly from parents to offspring, the resulting genotype is analogous to the experimental treatment

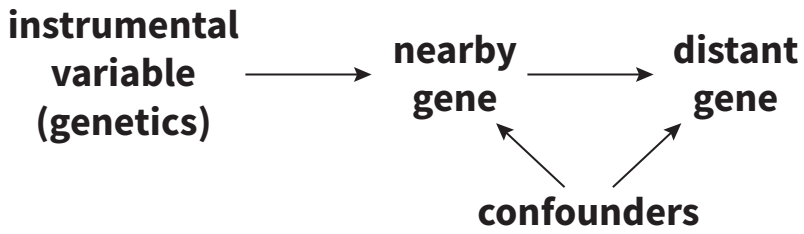

Figure 1.3: Underlying principle of Mendelian Randomization in genome research. A relationship between two genes (nearby gene, distant gene) is confounded by several factors. An instrumental variable is used as a way to manipulate the expression levels of a nearby gene, generating a proxy for its expression levels. This proxy is associated with the expression levels of the distant gene. This approach assumes the instrumental variable is not affected by any of the confounders inducing a correlation between the nearby and distant genes. Furthermore, it is assumed the instrumental variable does not directly influence the distant gene, but that any effect of the instrumental variable on the distant gene is mediated only by the nearby gene.

given in a RCT. Instead of being a proxy for a clinical treatment, it is often used as a way to "manipulate" other molecular phenotypes, such as gene expression. Associating this proxy, often called an instrumental variable, with other omics data may provide a way of detecting causal relationships (Figure 1.3).

The efficacy of MR depends on the quality of the proxy, *i.e.*, how well it mimics changes in the exposure. In the case of gene networks (Figure 1.3), this means how well the expression values can be explained by the instrumental variable. More data often helps in improving its predictive ability, but may not be available to any individual researcher. That is why combined efforts are key in performing this type of research.

The need for more data: combining large-scale efforts

It should come as no surprise that, similar to molecular QTL mapping, Mendelian Randomization needs large datasets, in two different ways. Firstly, increased sample sizes are needed to reliably detect any associations when using a genome-wide approach. Secondly, multiple layers of molecular information are needed to perform such analyses. A single study, whether molecular QTL mapping or Mendelian Randomization, may entail performing tens of millions of statistical hypothesis tests, which is especially true when relating different types of omics data with each other. Each dataset may contain up to millions of variables, and the detected effect sizes are often small. Therefore, a prerequisite of these genome-wide studies, particularly of the types described here, is the availability of thousands of samples. Ideally, all individual data would be stored in one place,

allowing the researcher to explore the full dataset, without being limited to only performing set meta-analyses.

The Biobank-based Integrative Omics Studies (BIOS) Consortium consists of several biobanks from all over The Netherlands. A biobank collects and stores data from a large collection of individuals from specific populations for several research purposes. The gathered data serves as a resource for many researchers to investigate different epidemiological research questions. Each of the biobanks from the BIOS Consortium has gathered lots of phenotypic and molecular information on different Dutch populations, aiming to investigate specific phenotypes. Despite their differences in specific research questions, they have shared their resources to be able to better understand how molecular phenotypes influence phenotypic endpoints. Specifically, genome-wide, individual level data on the genotypes, methylation levels (Illumina HumanMethylation450), and gene expression levels (RNA-sequencing) measured in over 4,000 healthy individuals from the Dutch population are now available to the community. Combining all the raw data has allowed us, and others, to explore all the raw data, furthering our understanding of transcriptional regulation using population genomics.

Outline of thesis

In this thesis, we aim to explore how genetic variation influences transcriptional regulation, a key process underlying many biological traits, so as to better understand how genetic variants ultimately influence complex and common phenotypes. More specifically, we try to investigate the local and distal influences genetic variants have on several molecular phenotypes, and the mechanism behind these. Furthermore, we aim to make causal claims about the relationships between molecular phenotypes, even in the face of confounding factors. We do this by investigating genetics, DNA methylation, and gene expression, through the development and deployment of several analysis strategies, including methylation QTL (meQTL) mapping, expression QTL mapping (eQTL), and Mendelian randomization-type approaches.

In **chapter 2**, we investigate meQTL mapping *in cis* from a methodological perspective. This shows that *cis*-effects are very local in nature, more so than previously appreciated. In addition, we show that a common approach to multiple testing often leads to an inflated number of CpGs identified as harboring a *cis*-meQTL.

In **chapter 3** we aim to directly relate genetic variants identified by many different GWA studies to the methylation levels of both nearby (*cis*) and distant (*trans*) autosomal CpG sites. This large undertaking resulted in the identification of *trans*meQTLs for one-third of tested genetic variants. We observe several variants each influencing the methylation levels of hundreds of CpG sites genome-wide, and propose these effects are likely due to *cis*-effects on transcription factor activity. This is supported by *cis*-eQTLs of these genetic variants on nearby transcription factor expression levels, as well as an enrichment of target CpG sites

overlapping with the corresponding transcription factor binding sites. In addition, we find one third of all tested CpGs to harbor a *cis*-meQTL, and are able to link variation in DNA methylation of CpG sites to the expression of different genes *in cis*.

In **chapter 4**, we further investigate the effects of SNPs on DNA methylation using X-chromosome inactivation. We describe an approach to identify autosomal loci influencing X-chromosomal methylation in a female-specific manner. Using this approach, we identify and replicate three loci that form a genetic basis of genes variably escaping X-chromosome inactivation (XCI) genes through hypomethylation of CpG islands (CGIs). These CGIs are located in regions known to variably escape XCI, and are associated to changes in the expression of nearby genes.

In **chapter 5** we go beyond molecular QTL mapping by utilizing and improving upon recent developments in data analysis to develop a resource of possible causal drivers of gene-gene interactions. We use genetics as a causal anchor, relieving the analysis from confounding. The identified drivers are often known transcription factors or chromatin remodelers, indirectly validating this approach. The results provide a resource from where novel biological insights into gene function and disease etiology can be distilled.

In conclusion, we aim to move towards generating causal hypotheses regarding transcriptional (dys)regulation using observational data. For example, we use data on transcription factors and chromatin remodelers to interpret molecular QTL mapping results (**chapter 3, 4, 5**), and use MR-type approaches to possibly directly establish hypotheses about causal relationships between molecular phenotypes (**chapter 5**). Together, these studies showcase how genetics can be utilized to systematically investigate transcriptional (dys)regulation, and better understand how this key process drives complex phenotypes, including common diseases.

References

Battle, A. et al. [2014]. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals, *Genome Res* **24**(1): 14–24.

Bell, J. T. et al. [2011]. Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines, *Genome Biol* **12**(1): R10.

Belshaw, R. et al. [2004]. Long-term reinfection of the human genome by endogenous retroviruses, *Proceedings of the National Academy of Sciences of the United States of America* **101**(14): 4894–4899.

Botstein, D. and Risch, N. [2003]. Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease, *Nature Genetics* **33**(3S): 228–237.

Boyle, E. A. et al. [2017]. An Expanded View of Complex Traits: From Polygenic to Omnigenic, *Cell* **169**(7): 1177–1186.

Buhule, O. D. et al. [2014]. Stratified randomization controls better for batch effects in 450K methylation analysis: A cautionary tale, *Frontiers in Genetics* **5**.

Burgess, S. et al. [2016]. Beyond Mendelian randomization: How to interpret evidence of shared genetic predictors, *Journal of Clinical Epidemiology* **69**: 208–216.

Davey Smith, G. and Hemani, G. [2014]. Mendelian randomization: genetic anchors for causal inference in epidemiological studies, *Hum Mol Genet* **23**(R1): R89–98.

de la Fuente, A. [2010]. From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases, *Trends in Genetics* **26**(7): 326–333.

Edwards, S. L., Beesley, J., French, J. D. and Dunning, A. M. [2013]. Beyond GWASs: illuminating the dark road from association to function, *Am J Hum Genet* **93**(5): 779–797.

Ferguson-Smith, A. C. [2011]. Genomic imprinting: The emergence of an epigenetic paradigm, *Nature Reviews Genetics* **12**(8): 565–575.

Garagnani, P. et al. [2012]. Methylation of ELOVL2 gene as a new epigenetic marker of age, *Aging Cell* **11**(6): 1132–1134.

Grundberg, E. et al. [2012]. Mapping cis- and trans-regulatory effects across multiple tissues in twins, *Nature genetics* **44**(10): 1084–1089.

GTEx Consortium [2017]. Genetic effects on gene expression across human tissues, *Nature* **550**: 204.

Gutierrez-Arcelus, M. et al. [2013]. Passive and active DNA methylation and the interplay with genetic variation in gene regulation, *Elife* **2**: e00523.

Hall, E. et al. [2014]. Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets, *Genome Biology* **15**(12): 522.

Heijmans, B. T. et al. [2007]. Heritable rather than age-related environmental and stochastic factors dominate variation in

DNA methylation of the human IGF2/H19 locus, *Hum Mol Genet* **16**(5): 547–554.

Heijmans, B. T. et al. [2008]. Persistent epigenetic differences associated with prenatal exposure to famine in humans, *Proc Natl Acad Sci* **105**(44): 17046–17049.

Hindorff, L. A. et al. [2009]. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *Proc Natl Acad Sci USA* **106**(23): 9362–9367.

Hu, S. et al. [2013]. DNA methylation presents distinct binding sites for human transcription factors, *eLife* **2**: e00726.

Illingworth, R. S. et al. [2010]. Orphan cpg islands identify numerous conserved promoters in the mammalian genome, *PLoS Genetics* **6**(9): e1001134.

Kass, S. U., Pruss, D. and Wolffe, A. P. [1997]. How does DNA methylation repress transcription?, *Trends in Genetics* **13**(11): 444–449.

Laird, P. W. [2005]. Cancer epigenetics, *Human Molecular Genetics* **14**(SPEC. ISS. 1).

Lango Allen, H. et al. [2010]. Hundreds of variants clustered in genomic loci and biological pathways affect human height, *Nature* **467**(7317): 832–838.

Lee, T. I. and Young, R. A. [2013]. Transcriptional regulation and its misregulation in disease, *Cell* **152**(6): 1237–1251.

Lemire, M. et al. [2015]. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci, *Nat Commun* **6**: 6326.

Lyon, M. F. [1961]. Gene Action in the X-chromosome of the Mouse (Mus musculus L.), *Nature* **190**(4773): 372–373.

Manolio, T. A. [2010]. Genomewide association studies and assessment of the risk of disease, *N Engl J Med* **363**(2): 166–176.

McCarthy, M. M. et al. [2009]. The epigenetics of sex differences in the brain, *The Journal of neuroscience : the official journal of the Society for Neuroscience* **29**(41): 12815–12823.

Miranda, T. B. and Jones, P. A. [2007]. DNA methylation: The nuts and bolts of repression, *Journal of Cellular Physiology* **213**(2): 384–390.

Morey, C. and Avner, P. [2010]. Genetics and epigenetics of the X chromosome, *Annals of the New York Academy of Sciences* **1214**.

Nica, A. C. and Dermitzakis, E. T. [2008]. Using gene expression to investigate the genetic basis of complex disorders, *Human molecular genetics* **17**(R2): R129–R134.

Pelikan, R. C. et al. [2018]. Enhancer histone-QTLs are enriched on autoimmune risk haplotypes and influence gene expression within chromatin networks, *Nature Communications* **9**(1).

Sharp, A. J., Stathaki, E., Migliavacca, E., Brahmachary, M., Montgomery, S. B., Dupre, Y. and Antonarakis, S. E. [2011]. DNA methylation profiles of human active and inactive X chromosomes, *Genome Res* **21**(10): 1592–1600.

Shi, J. et al. [2014]. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue, *Nat Commun* **5**: 3365.

Smith, A. K. et al. [2014]. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type, *BMC Genomics* **15**: 145.

Smith, Z. D. and Meissner, A. [2013]. DNA methylation: roles in mammalian development, *Nat Rev Genet* **14**(3): 204–220.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. [2003]. A gene-coexpression network for global discovery of conserved genetic modules, *Science* **302**(5643): 249–255.

Thomas, D. [2010]. Gene-environment-wide association studies: Emerging approaches, *Nature Reviews Genetics* **11**(4): 259–272.

van Iterson, M. et al. [2017]. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution, *Genome Biol* **18**(1): 19.

Visscher, P. M., Brown, M. A., McCarthy, M. I. and Yang, J. [2012]. Five years of GWAS discovery, *Am J Hum Genet* **90**(1): 7–24.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. and Yang, J. [2017]. 10 Years of GWAS Discovery: Biology, Function, and Translation, *American Journal of Human Genetics* **101**(1): 5–22.

Westra, H. J. et al. [2013]. Systematic identification of trans eQTLs as putative drivers of known disease associations, *Nature Genetics* **45**(10): 1238–1243.

Yang, F. et al. [2017]. Identifying cis -mediators for trans -eQTLs across many human tissues using genomic mediation analysis, *Genome research* pp. 1–13.

Yasukochi, Y. et al. [2010]. X chromosome-wide analyses of genomic DNA methylation states and gene expression in male and female neutrophils, *Proc Natl Acad Sci U S A* **107**(8): 3704–3709.

Yengo, L. et al. [2018]. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestryHuman Molecular Genetics, 2018, Vol. 00, No. 0Human Molecular Genetics, 2018, Vol. 00, No. 0, *Human Molecular Genetics* pp. ddy271–ddy271.

Zeilinger, S. et al. [2013]. Tobacco smoking leads to extensive genome-wide changes in DNA methylation, *PLoS ONE* **8**(5): e63812.

2 An alternative approach to multiple testing for methylation QTL mapping reduces the proportion of falsely identified CpGs

René Luijk, J.J. Goeman, E.P. Slagboom,
B.T. Heijmans, and E.W. van Zwet

Bioinformatics, **31(3):340-345 (2015)**

Abstract

An increasing number of studies investigates the influence of local genetic variation on DNA methylation levels, so called in *cis* methylation Quantitative Trait Loci (meQTL*s*). A common multiple testing approach in genome-wide *cis*-meQTL studies limits the false discovery rate (FDR) among all CpG-SNP pairs to 0.05 and reports on CpGs from the significant CpG-SNP pairs. However, a statistical test for each CpG is not performed, potentially increasing the proportion of CpGs falsely reported on. Here, we presented an alternative approach that does properly control for multiple testing at the CpG level.

We performed *cis*-meQTL mapping for varying window sizes using publicly available SNP and 450k data, extracting the CpGs from the significant CpG-SNP pairs (FDR < 0.05). Using a new bait-and-switch simulation approach, we show that up to 50% of the CpGs found in the simulated d ata may be false positives. We present an alternative, two-step multiple testing approach using the Simes and Benjamini-Hochberg procedures that does control the FDR among the CpGs, as confirmed by the bait-and-switch simulation. This approach indicates the use of window sizes in *cis*-meQTL mapping studies that are significantly smaller than commonly adopted.

Our approach to *cis* meQTL mapping properly controls the FDR at the CpG level, is computationally fast and can also be applied to *cis* eQTL studies.

Introduction

Genome-wide association studies (GWASs) are widely used to uncover the genetic basis of complex disease. Disease-associated genetic variants identified in GWASs are commonly located in non-coding regions, leaving the molecular mechanism underlying the associations unclear [Visscher et al., 2012]. The likely mechanism involves an effect on transcriptional activity of genes nearby (*cis*) or located distantly (*trans*), for example by influencing epigenetic regulation [Mill and Heijmans, 2013]. This can be studied by investigating the relationship between genetic variation, epigenetic marks including DNA methylation and gene expression. Already, many studies have reported on associations of specific genetic variants with variation in gene expression (expression QTL or eQTLs, Small et al. [2011]; Westra et al. [2013]) and DNA methylation, in particular the methylation of cytosines in CpG dinucleotides (*e.g.*, Heijmans et al. [2007]; Shi et al. [2014]; Wagner et al. [2014]) (DNA methylation quantitative trait loci or meQTLs). Creating catalogs of meQTLs and eQTLs will be instrumental in the discovery of genetic mechanisms determining DNA methylation and gene expression, the possible interplay between the two, and eventually the etiology of common diseases. To achieve this goal, further development of sound statistical methodology will be important.

Typically in meQTL and eQTL studies, a GWAS (*i.e.*, testing hundreds of thousands to millions of single nucleotide polymorphisms, SNPs) is performed for the level of methylation of every CpG measured or the level of transcription of every gene (more generally, for every transcript or exon), respectively, leading to a vast amount of possible combinations to investigate. While we will focus on *cis* meQTL studies, we note that the same principles and problems may also apply to *cis* eQTL studies. With the recent introduction of the Illumina 450k DNA methylation array [Bibikova et al., 2011], meQTL studies have become possible investigating over 400 thousand CpGs in large numbers of subjects. To test for associations of methylation at CpGs with genetic variants *in cis*, that is locally, studies have been considering SNPs anywhere between 5 kb [Gutierrez-Arcelus et al., 2013] to 1,000 kb [Gibbs et al., 2010] from measured CpGs. Particularly large window sizes will result in hundreds of millions statistical tests and thus brings about a huge multiple testing problem. A common strategy to account for multiple testing in meQTL studies is to control the false discovery rate (FDR; Benjamini and Hochberg [1995]) of all significantly associated CpG-SNP pairs at 0.05 (*e.g.*, Grundberg et al. [2013]; Drong et al. [2013]). This means that 5% of all significantly associated CpG-SNP pairs are expected to be false positives.

Due to the extensive linkage disequilibrium (LD) in the human genome, individual CpGs will frequently be associated with many SNPs. Hence, a particular CpG will often occur many times in the list of significant CpG-SNP pairs. In practice, this is redundant information because LD structure renders it impossible to pinpoint the causal SNP responsible for the variation in DNA methylation using statistical means (cf. GWAS; Pearson and Manolio [2008]; Feero et al. [2010]). Hence, the results reported on and further analyses generally focus on the CpGs in the list of significant CpG-SNP associations. That is, all CpGs that significantly

associate with at least one SNP (*e.g.*, Zhang et al. [2010]; Liu et al. [2013]; van Eijk et al. [2012]). We will refer to this approach as the CpG-SNP pair-based approach.

A large proportion of CpGs among the FDR significant CpG-SNP pairs may be false positives [Bell et al., 2011; Westra et al., 2013]. To obtain a list of CpGs influenced by genetic variation *in cis* that is properly controlled for multiple testing, we propose to formally test each CpG, obtaining a single valid P-value per CpG and control the FDR among those P-values, which we will refer to as the CpG-based approach. Using a new bait-and-switch simulation scheme we compare the proportion of falsely identified CpGs using the CpG-SNP pair-based approach and our proposed CpG-based approach in simulated data.

Methods

Data

We used Illumina 450k DNA methylation data [Heyn et al., 2013] and Illumina HumanHap 550k SNP data [Niu et al., 2010] on 96 unrelated healthy Caucasian-Americans. The DNA samples were obtained from lymphoblastoid cell lines included in the Human Variation panel (sample set HD100CAU; Coriell Cell Repositories). Both data sets are publicly available from the GEO data repository (accession numbers GSE36369 and GSE24260, respectively). The SNP array data were imputed to 30,038,302 SNPs based on the 1000 Genomes CEU reference panel and using IMPUTE v2 [Howie et al., 2009]. A dosage value ranging from 0 to 2 reflected the uncertainty in the imputation for the imputed SNPs. We selected SNPs with a minor allele frequency above 5%, a minimum call rate of 95%, and an imputation quality score of at least 0.4, leaving 6,596,758 SNPs for analysis.

The quality control of the 450k array was done based on the signal intensities and detection P-values. We set any beta values [Du et al., 2010], a measure of the DNA methylation fraction, with a corresponding detection P-value lower than 0.01 to missing. Next, we removed any samples with a log2 median intensity under 10.5 in either the methylated or the unmethylated signal. In addition, we removed any probes or samples with a call rate lower than 95%. Lastly, we removed probes mapping to the sex chromosomes, mapping ambiguously to the genome [Chen et al., 2013], or with a SNP in the interrogated CpG (MAF $> 1\%$ in 1000 Genomes). These filters resulted in 423,825 probes left for analysis out of the 482,421 probes on the array targeting CpG sites. The normalization of the 450k data consisted of a correction for background signal, followed by a dye-bias correction. Both procedures were performed using the methylumi package [Davis et al., 2013]. All further analyses were done using beta values. To verify that genotype and methylation data were linked to the correct sample identified, MixupMapper was used [Westra et al., 2011]. For 77 out of 93 samples remaining after quality control, SNP and methylation data could be linked (Supplementary Tables 1 and 2).

meQTL mapping

We tested all associations between genotypes and methylation of CpGs *in cis*, that is, locally, defined by window sizes from 1 kb to 500 kb around each CpG, calculating the Spearman rank correlation between the imputed dosage values and beta values. To this end, we use the Matrix eQTL package [Shabalin, 2012]. Because the Matrix eQTL package is only able to calculate the Pearson correlation, which is less robust to outliers than the Spearman rank correlation, we pre-calculated the ranks of the observed values for all CpGs and SNPs as input for Matrix eQTL to obtain a test on the basis of the Spearman correlation. The Matrix eQTL package provides a list of all CpG-SNP pairs tested across all windows evaluated and the P-values reflecting the statistical significance of the associations. Obtaining a list of statistically significant CpG-SNP pairs was achieved by limiting the false discovery rate (FDR) among the CpG-SNP pairs to 0.05.

Obtaining an FDR controlled list of CpGs influenced by genetic variation

While the FDR among the CpG-SNP pairs is controlled at 0.05, there is no guarantee that this is also true for the set of CpGs among these pairs. No formal statistical test is performed for each CpG individually, testing the global null hypothesis H_0 of no association between the variation in methylation and genetic variation *in cis*. In order to obtain a list of CpGs that is controlled at an FDR of 0.05, we proceed as follows. First, we perform a statistical test to assess the global null hypothesis $H_{0,i}$ of no association between a CpG i and the SNPs *in cis* to obtain one valid P-value for each CpG. Next, we apply the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] to these P-values to obtain an FDR controlled list of CpGs associated with genetic variation. Since commonly used software packages return P-values $p_{i,j}$ for all CpG-SNP pairs (i, j) tested, we propose to use the $p_{i,j}$ to test the global null hypothesis $H_{0,i}$. The Bonferroni correction multiplies the minimum of the observed P-values in a window by the number of such P-values

$$P_i = k_i \min(p_{1,i}, \ldots, p_{k_i,i}), \tag{2.1}$$

where k_i is the number of SNPs in that window. The Bonferroni correction is conservative in the case of dependent P-values (like in the case of LD between SNPs), since the effective number of tests done may be smaller than the number of tests corrected for by Bonferroni. Hence, we propose to use the Simes procedure [Simes, 1986] (see Supplemental Materials), a method developed specifically to test a global null hypothesis H_0. This method makes the extra assumption of positive dependence among the P-values, similar to the Benjamini-Hochberg procedure [Goeman and Solari, 2014]. The Simes procedure implicitly takes these dependencies into account, yielding a less conservative P-value than the Bonferroni correction. The Simes procedure orders the P-values belonging to CpG i in ascending order, such that $p_{(1),i} \leq \cdots \leq p_{(k_i),i}$. Next, a P-value P_i for CpG

i is calculated by multiplying each $p_{(j),i}$ by a smaller factor k_i/j and taking the minimum of these corrected P-values:

$$P_i = \min \left\{ j : \frac{k_i}{j} p_{(j),i} \right\} \tag{2.2}$$

Both the Bonferroni procedure and the Simes procedure multiply the smallest P-value $p_{(1),i}$ by k_i. However, the Simes procedure multiplies the larger $p_{i,j}$ by a smaller factor, making the Simes procedure a more liberal procedure in the case of positively correlated P-values.

Estimating the CpG level false discovery proportion in a simulated setting using the bait-and-switch simulation procedure

We have discussed two approaches to compiling a list of CpGs influenced by genetic variation: the CpG-SNP pair-based approach and our CpG-based approach. We will now discuss a novel data-based simulation scheme called the bait-and-switch simulation to provide an assessment of the performance of these approaches in terms of the proportion of CpGs falsely identified as being significantly associated with genetic variation in a realistic simulation setting. Because simulation of realistic genome-wide genotype and methylation data is hard to do from scratch, we choose to modify the current data set in such a way that we have knowledge of what null hypotheses are true, i.e. which CpGs should not associate with any genetic variation. This simulation consists of several steps and is depicted in Figure 2.2A:

1. Within-window correction: perform the Simes correction within each CpG's window separately. Take the minimum adjusted P-value as the P-value for this CpG.

2. Between-windows correction: control the FDR among the newly calculated P-values to obtain a list of FDR significant CpGs.

3. The data consisting of FDR significant CpGs will be called the bait set. The rest of the data, the non-significant CpGs, are called the switch set.

4. Permute the methylation values for the switch set, leaving the data in the bait set and the genotype data intact.

5. Perform the CpG-SNP pair-based approach and the CpG-based approach on the simulated data, obtaining a list of significant CpGs for each approach.

To get an estimate of the CpG level FDR, we calculate the proportion of the CpGs obtained in step 3 coming from the switch set. Although we do not know which of the CpGs in the bait set are truly associated with genetic variation, we do know that none of the CpGs in switch set have any such association. As a result, the calculated FDP is a lower bound. The CpG level FDR is the average of the different realizations of the FDP coming from many repetitions of the same simulation experiment.

Results

CpG-SNP pair-based meQTL mapping approach

We performed *cis* meQTL mapping, varying the window size from 1 kb to 500 kb. For each window size, we applied the the CpG-SNP pair-based approach, obtaining a list of statistically significant CpG-SNP pairs and a list of the CpGs among these CpG-SNP pairs, i.e. the CpGs that are associated with at least one SNP. Despite the relatively small sample size, Figure 2.1 shows that the Benjamini-Hochberg method finds an increasing number of CpG-SNP pairs with increasing window size, with a maximum of 223,428 CpG-SNP pairs at a 200 kb window and a maximum of 10,034 CpGs at the 100 kb window size. If we keep expanding the search window around each CpG the multiple testing burden becomes too great, leading to a slight decrease in the number of CpG-SNP pairs and CpGs in that list. The increase in the number of CpG-SNP pairs can be mainly attributed to linkage disequilibrium (LD). When observing a statistically significant CpG-SNP pair, LD may virtually guarantee finding more significant CpG-SNP pairs if that SNP is strongly correlated to other nearby SNPs and we expand the window around each CpG. This is illustrated by the LocusZoom plot [Pruim et al., 2010] for a CpG (cg12247378) associated with several SNPs on 22q13.1 in Figure 2.1B. Many of the SNPs associated with this CpG are in LD and will be included with an increasing window size.

Evaluating the CpG level false discovery proportion in a simulated setting using the bait-and-switch simulation

LD causes identification of the causal SNP responsible for the variation in methylation to be impossible by statistical means. Therefore, it would be more insightful to consider individual CpGs only, instead of focusing on all CpG-SNP pairs. Following the CpG-SNP pair-based approach, we report on the CpGs from the FDR significant CpG-SNP pairs found (see Figure 2.1), i.e. the CpGs that associate with at least one SNP. However, this set of CpGs has no guarantee of FDR control and likely includes many false positive CpGs.

To evaluate the CpG level FDP among the in a controlled setting, we use the bait-and-switch simulation scheme. We construct a new, simulated data set that is very similar to the original data, but allows us to compute a lower bound on the FDP among the CpGs. Performing the CpG-SNP pair-based approach for varying

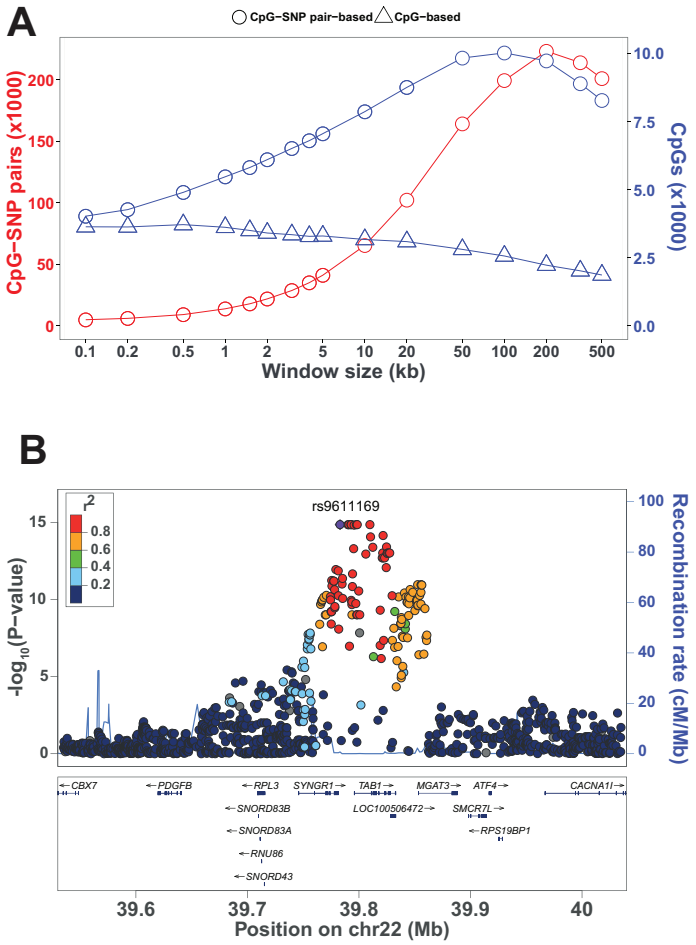

Figure 2.1: (**A**) The number of CpG-SNP pairs and the number of CpGs among them for different window sizes in the real data. The grey line shows the number of CpG-SNP pairs (FDR < 0.05). The black lines show the number of CpGs found. The two different symbols denote the CpG-SNP pair-based approach (circles) and our proposed CpG-based approach (triangles). Both the number of CpG-SNP pairs and the CpGs among them increase with window size when using the CpG-SNP pair-based approach. The CpG-based approach finds less CpGs, and reaches an optimum at at a 500 base pair window size. (**B**) CpGs associated with genetic variation are often associated with many SNPs due to LD. The LocusZoom plot shows the associations between CpG cg12247378 (22q13.1) and the SNPs in its window. The left y-axis shows the P-value corresponding to the association with the methylation levels on a -$\log_{10}$-scale, the right axis shows the recombination rate. The color coding indicates the r^2 between the SNPs, based on 1000 Genomes, build hg19. Many of the associated SNPs are in strong LD with one another.

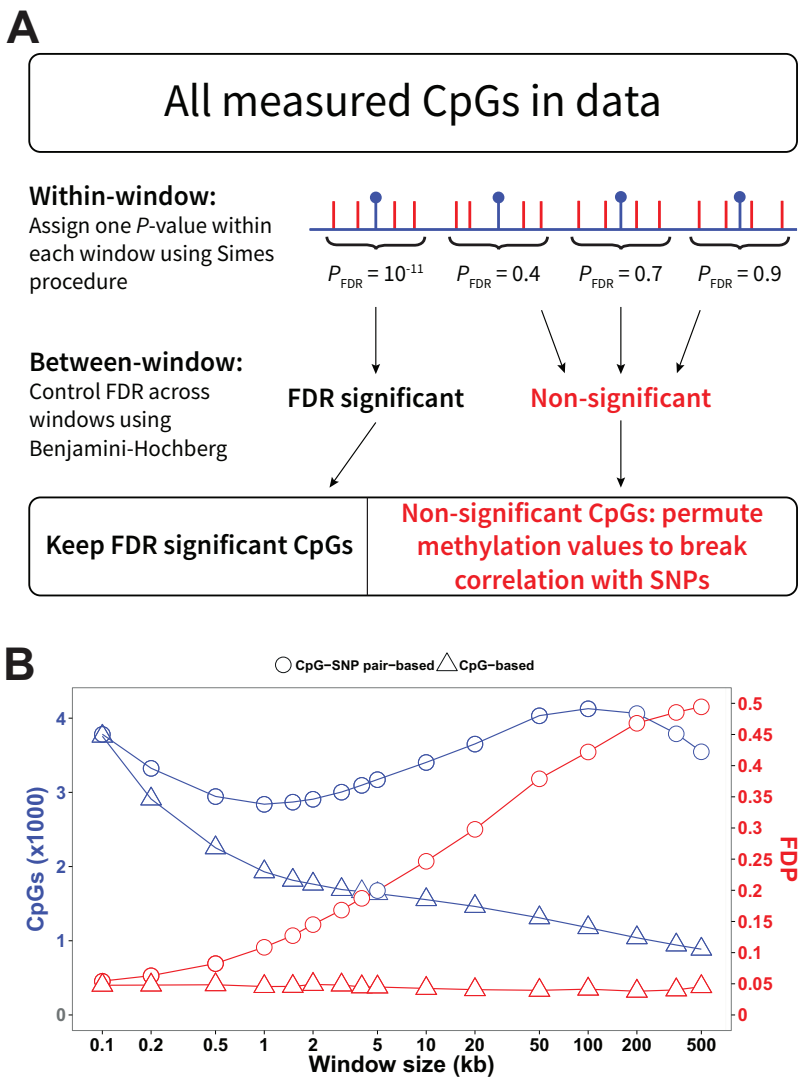

Figure 2.2: The number of CpGs found using the CpG-SNP pair-based approach, our proposed CpG-based approach and the corresponding CpG level FDP for different window sizes in the bait-and-switch simulated data. (**A**) An overview of the bait-and-switch simulation. (**B**) The grey line shows the number of CpGs. The black lines show the corresponding CpG level FDP. The two different symbols denote the CpG-SNP pair-based approach (circles) and our proposed CpG-based approach (triangles).

window sizes on the simulated data set yields a list of CpGs associated with at least one SNP and a list of all CpG-SNP pairs at an FDR of 0.05, similar to the results in Figure 2.1A. In the simulated data set we know for which CpGs we permuted the methylation values and thus are false positives (see Figure 2.2B). Strikingly, a large portion of the identified CpGs using the CpG-SNP pair-based approach seem to be false positives, especially for larger window sizes (Figure 2.2). Even when using a very small 0.5 kb window size, we find an estimated FDP of 0.1 (SE = 0.0006, based on 5 permutations), meaning at least 10% of the CpGs found among the CpG-SNP pairs in the simulated data are coming from the permuted switch set, i.e. are not truly associated with a SNP. This number greatly increases to 49.1% (FDP = 0.49, SE = 0.002, based on 5 permutations) for the 500 kb window size. While we can only claim that up to 50% of the CpGs found in the simulated data are false positives, this approach will probably yield an inflated proportion of falsely identified CpGs in the original data too. Our proposed CpG-based approach, however, controls the FDP at 0.05 (SE 0.001-0.005, based on 5 permutations) for all window sizes.

A FDR controlled list of CpGs influenced by genetic variation

To obtain a valid list of CpGs that are significantly associated with genetic variation *in cis* in the original data, we calculated one P-value per CpG, testing the global hypothesis of no association between variation in methylation any of the SNPs in its window. We calculated these P-values by means of the the Simes procedure. The Simes procedure implicitly takes into account the correlation structure among the SNPs, making it a more powerful method than, *e.g.*, the conservative Bonferroni method. After this within-window correction, we applied the Benjamini-Hochberg procedure to the resulting P-values, controlling the FDR among the CpGs to 0.05. Figure 2.1A shows that this approach identifies a maximum of 3,721 CpGs at a 500 base pair window size (black line, triangles). This suggests that strongly associated SNPs are often in close proximity to the CpG, as reported earlier [Bell et al., 2011; Gutierrez-Arcelus et al., 2013]. To show that this approach does control the FDP among the CpGs at the desired level, we again conducted the same bait-and-switch simulation experiment, applying our proposed CpG-based approach on the simulated data set. While our approach seemingly discovers fewer CpGs than the CpG-SNP pair-based approach to meQTL mapping when applied to the original data, the FDR among the CpGs identified in the simulated data is controlled at 0.05 (Figure 2.2B).

Discussion

We report on a CpG-based multiple testing approach in meQTL mapping to identify individual CpGs whose methylation level is influenced by genetic variation *in cis*. Our approach is based on the application of the Simes procedure within a window around each CpG to obtain a single P-value per CpG, followed by the Benjamini-Hochberg procedure to control the FDR across CpGs. Strikingly, this approach

suggests that optimal window sizes for the identification of *cis* meQTLs are much smaller than frequently used in the literature (up to 10s of kb instead of 100s of kb). These smaller window sizes are in line with reports that SNPs strongly associated with a CpG are often in close proximity to the CpG [Bell et al., 2011; Gutierrez-Arcelus et al., 2013]. The large window sizes used in literature may stem from the CpG-SNP pair-based approach reporting on the CpGs from a list of all FDR significant CpG-SNP pairs. Using the bait-and-switch simulation we show that the latter approach yields up to 50% falsely identified CpGs in simulated data. Our proposed approach controls the CpG level FDR at the desired level and still identifies a substantial number of CpGs associated with genetic variation *in cis*.

Our method can be directly applied to the output of commonly used QTL mapping software, *e.g.*, Matrix EQTL, which returns P-values corresponding to every CpG-SNP pair tested. In addition, the current method does not require the use of permutations to control the FDR, making it a fast and easy-to-use approach. While permutations are still feasible for small 450k array data sets, this becomes burdensome for large data sets, particularly when using bisulphite-sequencing data measuring millions of CpG sites.

When calculating one P-value for the window around each CpG site it is important to account for LD between SNPs in the window. Not doing so will substantially reduce statistical power. Therefore, some methods, like the Bonferroni correction, may be too conservative. The Simes procedure implicitly takes LD into account by multiplying larger P-values with smaller factors. Although the Simes procedure seems to perform well in terms of CpGs found, it still does not fully capture the correlation structure. A possible solution would be to estimate the number of independent tests for each window, *e.g.*, using GATES [Li et al., 2011] or TATES [van der Sluis et al., 2013], accounting for the number of independent tests done. However, this may be computationally expensive. Our proposed approach is unable to distinguish between two independent SNP effects on the methylation levels of a CpG. It only allows for making claims about the global null hypothesis of no association with any genetic variant *in cis*. This approach takes into account that the causal variant cannot be identified with statistical means only. Another limitation is that there currently is no valid method to determine the optimal window size for a study prior to QTL-mapping. In general, the optimal window size will be greater for studies with higher statistical power. Our study suggests that the optimal window size will be 10-50 kb instead of the commonly used 100s kb , which will reduce statistical power by dramatically increasing the number of tests.

In this paper we introduced the bait-and-switch simulation method to estimate the true false discovery proportion among CpGs with a meQTL *in cis* in simulated data. This approach indicated up to 50% of identified CpGs in our simulated data may be false positive. While we know this is true in the simulated data, we cannot extrapolate this to the original data. It is likely that the common approach to multiple testing also brings about an increased CpG level FDR in the real data. This finding may also be an issue for *cis* expression QTL studies and possibly *trans* QTL studies. Interpretation of results based on the common approach evaluated here should be interpreted with caution.

Development of statistical methodology will aid in getting a complete catalogue of meQTLs and eQTLs that is key in understanding the mechanisms underlying the association of non-coding genetic variants with disease phenotypes.

Acknowledgements

This work was done within the framework of the Biobank-Based Integrative Omics Studies (BIOS) Consortium funded by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO 184.021.007).

References

Bell, J. T. et al. [2011]. Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines, *Genome Biol* **12**(1): R10.

Benjamini, Y. and Hochberg, Y. [1995]. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J Roy Stat Soc B Met.* pp. 289–300.

Bibikova, M. et al. [2011]. High density dna methylation array with single cpg site resolution, *Genomics* **98**(4): 288–295.

Chen, Y. et al. [2013]. Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray, *Epigenetics* **8**(2): 203.

Davis, S. et al. [2013]. Methylumi: Handle illumina methylation data 2012, *R package* **2**(1).

Drong, A. W. et al. [2013]. The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of dna methylation in adipose tissue, *PLoS One* **8**(2): e55923.

Du, P. et al. [2010]. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis, *BMC Bioinformatics* **11**(1): 587.

Feero, W. G. et al. [2010]. Genomewide association studies and assessment of the risk of disease, *N Engl J Med* **363**(2): 166–176.

Gibbs, J. R. et al. [2010]. Abundant quantitative trait loci exist for dna methylation and gene expression in human brain, *PLoS Genet* **6**(5): e1000952.

Goeman, J. J. and Solari, A. [2014]. Multiple hypothesis testing in genomics, *Stat Med* **33**(11): 1946–1978.

Grundberg, E. et al. [2013]. Global analysis of dna methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements, *Am J Hum Genet* **93**(5): 876–890.

Gutierrez-Arcelus, M. et al. [2013]. Passive and active dna methylation and the interplay with genetic variation in gene regulation, *eLife* **2**.

Heijmans, B. T. et al. [2007]. Heritable rather than age-related environmental and stochastic factors dominate variation in dna methylation of the human igf2/h19 locus, *Hum Mol Genet* **16**(5): 547–554.

Heyn, H. et al. [2013]. Dna methylation contributes to natural human variation, *Genome Res* **23**(9): 1363–1372.

Howie, B. N. et al. [2009]. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, *PLoS Genet* **5**(6): e1000529.

Li, M.-X. et al. [2011]. Gates: a rapid and powerful gene-based association test using extended simes procedure, *Am J Hum Genet* **88**(3): 283–293.

Liu, Y. et al. [2013]. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis, *Nat Biotechnol* **31**(2): 142–147.

Mill, J. and Heijmans, B. T. [2013]. From promises to practical strategies in epigenetic epidemiology, *Nat Rev Genet* **14**(8): 585–594.

Niu, N. et al. [2010]. Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines, *Genome Res* **20**(11): 1482–1492.

Pearson, T. A. and Manolio, T. A. [2008]. How to interpret a genome-wide association study, *JAMA* **299**(11): 1335–1344.

Pruim, R. J. et al. [2010]. Locuszoom: regional visualization of genome-wide association scan results, *Bioinformatics* **26**(18): 2336–2337.

Shabalin, A. A. [2012]. Matrix eqtl: ultra fast eqtl analysis via large matrix operations, *Bioinformatics* **28**(10): 1353–1358.

Shi, J. et al. [2014]. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue, *Nat Commun* **5**.

Simes, R. J. [1986]. An improved bonferroni procedure for multiple tests of significance, *Biometrika* **73**(3): 751–754.

Small, K. S. et al. [2011]. Identification of an imprinted master trans regulator at the klf14 locus related to multiple metabolic phenotypes, *Nat Genet* **43**(6): 561–564.

van der Sluis, S. et al. [2013]. Tates: efficient multivariate genotype-phenotype analysis for genome-wide association studies, *PLoS Genet* **9**(1): e1003235.

van Eijk, K. R. et al. [2012]. Genetic analysis of dna methylation and gene expression levels in whole blood of healthy human subjects, *BMC Genomics* **13**(1): 636.

Visscher, P. M. et al. [2012]. Five years of gwas discovery, *Am. J. Hum. Genet.* **90**: 7–24.

Wagner, J. R. et al. [2014]. The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts, *Genome Biol* **15**(2): R37.

Westra, H. et al. [2011]. Mixupmapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects, *Bioinformatics* **27**(15): 2104–2111.

Westra, H. et al. [2013]. Systematic identification of trans eqtls as putative drivers of known disease associations, *Nat Genet* **45**(10): 1238–1243.

Zhang, D. et al. [2010]. Genetic control of individual differences in gene-specific methylation in human brain, *Am J Hum Genet* **86**(3): 411–419.

3 DISEASE VARIANTS ALTER TRANSCRIPTION FACTOR LEVELS AND METHYLATION LEVELS OF THEIR BINDING SITES

M.J. Bonder*, **René 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. 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. Stehouwer, C.J. van der Kallen, C.G. Schalkwijk, L.H. van den Berg, E.W. van Zwet, H. Mei, Y. Li, M. Lemire, T.J. Hudson, BIOS Consortium, P.E. Slagboom, C. Wijmenga, J.H. Veldink, M.M. van Greevenbroek, C.M. van Duijn, D.I. Boomsma, A. Isaacs, R. Jansen, J.B. van Meurs, P.A.C. 't Hoen, L. Franke, B.T. Heijmans

* *Contributed equally*

Nature Genetics, **49**(1):131-138 (2017)

Main

Most disease-associated genetic variants are noncoding, making it challenging to design experiments to understand their functional consequences [Manolio, 2010; Visscher et al., 2012]. Identification of expression quantitative trait loci (eQTLs) has been a powerful approach to infer the downstream effects of disease-associated variants, but most of these variants remain unexplained [Westra et al., 2013; Wright et al., 2014]. The analysis of DNA methylation, a key component of the epigenome [Bernstein et al., 2007; Mill and Heijmans, 2013], offers highly complementary data on the regulatory potential of genomic regions [Gutierrez-Arcelus et al., 2013; Tsankov et al., 2015]. Here we show that disease-associated variants have widespread effects on DNA methylation *in trans* that likely reflect differential occupancy of *trans* binding sites by *cis*-regulated transcription factors. Using multiple omics data sets from 3,841 Dutch individuals, we identified 1,907 established trait-associated SNPs that affect the methylation levels of 10,141 different CpG sites *in trans* (false discovery rate (FDR) < 0.05). These included SNPs that affect both the expression of a nearby transcription factor (such as *NFKB1*, *CTCF* and *NKX2-3*) and methylation of its respective binding site across the genome. *Trans* methylation QTLs effectively expose the downstream effects of disease-associated variants.

To systematically study the role of DNA methylation in explaining the downstream effects of genetic variation, we analyzed genome-wide genotype and DNA methylation in whole blood from 3,841 samples from five Dutch biobanks [Tigchelaar et al., 2015; van Greevenbroek et al., 2011; Schoenmaker et al., 2006; Willemsen et al., 2013; Hofman et al., 2013] (Figure 3.1, Supplementary Table 1 and Supplementary Note). We found *cis* methylation quantitative trait locus (meQTL) effects for 34.4% of all 405,709 CpGs tested (n $= 139,566$ at a CpG-level FDR of 5%, $P < 1.38 \times 10^{-4}$), typically with a short physical distance between the SNP and CpG (median distance = 10 kb; Supplementary Figure 3.1). By regressing out the effect of the primary meQTL for each of these CpGs and repeating the *cis*-meQTL mapping, we observed up to 16 independent *cis*-meQTLs for each CpG site (Supplementary Table 2), totaling 272,037 independent *cis*-meQTL effects. We found that few factors determine whether a CpG site shows a *cis*-meQTL effect other than variance in the methylation levels of the CpG site involved (Supplementary Figures 2 and 3). The proportion of variance in methylation explained by SNPs, however, is typically small (Supplementary Figure 3.3b). When accounting for this strong effect of CpG variation, we found only modest enrichments and depletions of *cis*-meQTL CpG sites in CpG island and genic annotations (Supplementary Figure 3e) or when using annotations for biological function based on chromatin segmentations of 27 blood cell types (Figure 3.2a).

We contrasted these modest functional enrichments to those of CpGs whose methylation levels correlated with gene expression in *cis* (that is, expression quantitative trait methylation (eQTM)) by generating RNA-seq data for 2,101 of 3,841 individuals in our study. Using a conservative approach that maximally accounts for potential biases (Online Methods), we identified 12,809 unique

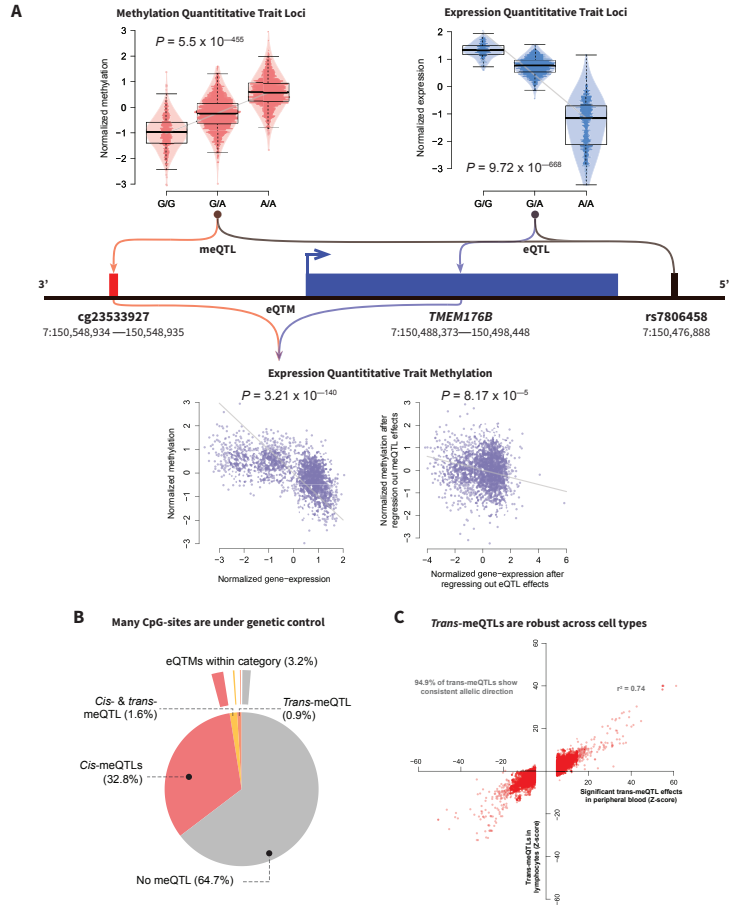

Figure 3.1: (**a**) In the illustration, the relationships between a SNP, DNA methylation at nearby CpGs and associations with the gene itself are shown. Boxes represent the median and interquartile range (IQR); whiskers extend to the outer quartile plus 1.5 times the IQR. The top left plot shows the observed meQTL between cg23533927 and rs7806458. The top right plot shows the observed eQTL between *TMEM176B* and rs7806458. The observed methylation-expression association (eQTM) between *TMEM176B* and cg23533927 is shown below the gene. The bottom left plot shows the data before correction for the *cis*-eQTL and *cis*-meQTL; the eQTM effect after correction for *cis*-eQTLs and *cis*-meQTLs is shown in the bottom right plot. (**b**) Two overlaid pie charts. The inner chart indicates the proportion of tested CpGs harboring meQTLs. Over 35% of all tested CpGs show evidence of harboring a meQTL, either *in cis* or *trans*. The outer chart indicates what CpGs are associated with gene expression *in cis* (in total, 3.2%). (**c**) Replication of peripheral blood *trans*-meQTLs in lymphocytes.

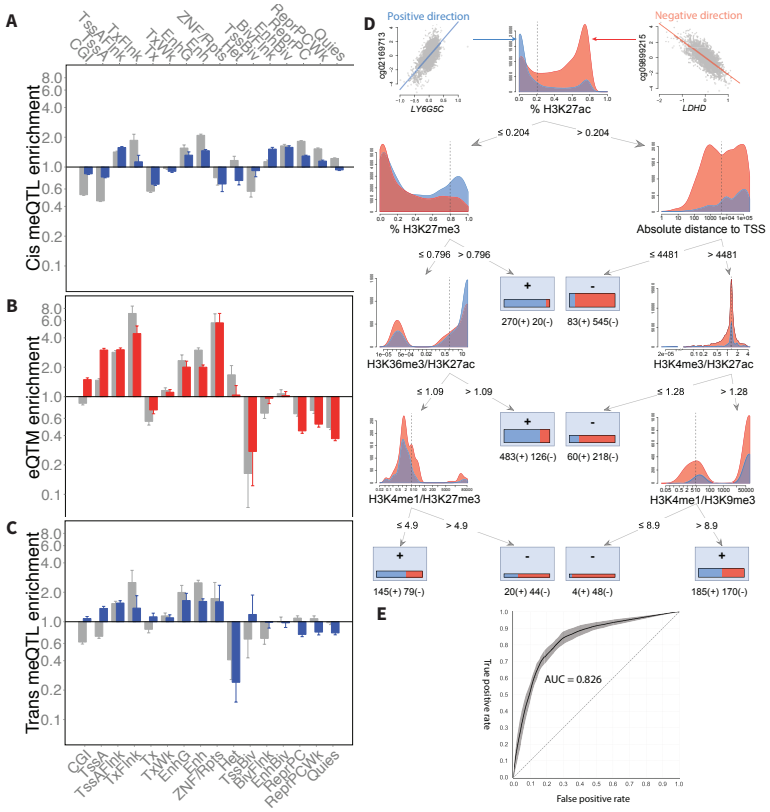

Figure 3.2: (**a-c**) Over- or under-representation of CpGs in predicted chromatin states for *cis*-meQTLs (**a**), eQTMs (**b**) and *trans*-meQTLs (**c**). Gray bars reflect uncorrected enrichment; colored bars reflect enrichment after correction for factors influencing the likelihood of harboring a meQTL or eQTM, including methylation variability. Bar graphs show odds ratios; error bars, 95% confidence intervals. CGI, CpG island; TssA, active TSS; TssAFlnk, flanking active TSS; TxFlnk, transcribed at a gene's 5′ or 3′ end; Tx, strong transcription; TxWk, weak transcription; EnhG, genic enhancer; Enh, enhancer; ZNF/Rpts, ZNF genes and repeats; Het, heterochromatin; TssBiv, bivalent/poised TSS; BivFlnk, flanking bivalent TSS/enhancer; EnhBiv, bivalent enhancer. (**d**) Decision tree for predicting the direction of the effects of eQTMs. Each subplot shows the distributions for positive (blue) and negative (red) associations for that subset of the data. Dashed vertical lines correspond to the optimal split used by the algorithm. The boxes in the leaves show the number of positive and negative effects in each of the leaves. (**e**) Receiver operator characteristic curve showing the performance of the decision tree. AUC, area under the receiver operating curve.

CpGs that correlated with 3,842 unique genes in *cis* (CpG-level FDR < 0.05). eQTMs were enriched for mapping to active regions, for example, in and around active transcription start sites (TSSs) (3-fold enrichment, $P = 1.8 \times 10^{-91}$) and enhancers (2-fold enrichment, $P = 1.1 \times 10^{-139}$; Figure 3.2b). The majority of eQTMs showed the canonical negative correlation with transcriptional activity (69.2%), but a substantial minority of correlations were positive (30.8%), in line with recent evidence that DNA methylation does not always negatively correlate with gene expression [Hu et al., 2013]. As expected, negatively correlated eQTMs were enriched in active regions such as active TSSs (3.7-fold enrichment, $P = 9.5 \times 10^{-202}$). Positive correlations primarily occurred in repressed regions (for example, Polycomb-repressed regions, 3.4-fold enrichment, $P = 5.8 \times 10^{-103}$) (Supplementary Figure 4). The sharp contrast between positively and negatively associated eQTMs enabled us to predict the direction of the correlation. A decision tree trained on the strongest eQTMs (those with FDR $< 9.7 \times 10^{-6}$, n $= 5,137$), using data on histone marks and distance relative to genes, could predict the direction with an area under the curve of 0.83 (95% confidence interval, $0.78 - 0.87$) (Figure 3.2d,e).

We next ascertained whether *trans*-meQTLs are biologically informative, as previous *trans*-eQTL mapping studies demonstrated that identifying *trans* expression effects provides a powerful tool to uncover and understand the downstream biological effects of disease-associated SNPs [Westra et al., 2013; Yao et al., 2015; Huan et al., 2015]. We focused on 6,111 SNPs that were previously associated with complex traits and diseases ('trait-associated SNPs'; Online Methods and Supplementary Table 3). We observed that one-third of these trait-associated SNPs (1,907 SNPs; 31.2%) affected methylation *in trans* at 10,141 CpG sites, totaling 27,816 SNP-CpG combinations (FDR < 0.05, $P < 2.6 \times 10^{-7}$; Figure 3.3a). This represents a fivefold increase in the number of CpG sites affected as compared with a previous *trans*-meQTL mapping study [Lemire et al., 2015]. We evaluated whether the trait-associated SNPs themselves were likely to underlie the *trans* effects or whether the associations could be attributed to other SNPs in moderate linkage disequilibrium (LD). Of the 1,907 trait-associated SNPs with *trans* effects, 1,538 (87.2%) were in strong LD with the top SNP ($r^2 > 0.8$), indicating that the GWAS SNPs are indeed the driving force behind many of the *trans*-meQTLs. Of note, because of the sparse coverage of the Illumina HumanMethylation450 BeadChip, the true number of CpGs in the genome that are altered by these trait-associated SNPs will be substantially higher.

To validate our *trans*-meQTLs, we performed a replication analysis in a set of 1,748 lymphocyte samples [Lemire et al., 2015]. Of the 18,764 overlapping *trans*-meQTLs, 94.9% had a consistent allelic direction in the replication data (Figure 3.1e and Supplementary Table 4). This indicates that the identified *trans*-meQTLs are robust and are not caused by differences in cell type composition. Further analysis of SNPs known to influence blood cell composition [Orru et al., 2013; Roederer et al., 2015] showed no or only few effects in *trans* and alternative adjustments of the methylation data corroborated the stability of the *trans* effects, with both approaches indicating a limited influence of cell type composition (Supplementary Tables 5, 6, 7 and Supplementary Note).

After identifying *trans*-meQTLs, we assessed whether their respective SNPs also affected the expression of the genes associated with the CpGs *in trans*. By overlaying the *trans*-meQTLs and *cis*-eQTMs, we could link 436 SNPs to 850 genes, totaling 2,889 SNP-gene pairs. We found significant associations (*trans*-eQTLs; FDR < 0.05) for 8.4% of these effects, and 91% of these effects showed the expected direction of effect given the directions of effect for the *trans*-meQTL and *cis*-eQTM (Supplementary Table 8).

In contrast to *cis*-meQTL CpGs, *trans*-meQTL CpGs showed substantial functional enrichment: they were enriched around TSSs and depleted in heterochromatin (Figure 3.2c) and were strongly enriched for being an eQTM (1,913 CpGs (18.9%), 5.2-fold enrichment, $P = 2.3 \times 10^{-101}$). Among the 1,907 trait-associated SNPs that made up the *trans*-meQTLs, there was an over-representation of GWAS-identified SNPs associated with immune- and cancer-related traits (Figure 3.3a). The large majority of *trans*-meQTLs were interchromosomal (93%; 9,429 CpG-SNP pairs) and included 12 *trans*-meQTL SNPs (yielding 3,616 unique CpG-SNP pairs) that each showed downstream *trans*-meQTL effects across all 22 autosomal chromosomes (*trans* bands; Figure 3.3b).

We subsequently studied the nature of these *trans*-meQTLs. Using high-resolution Hi-C data [Rao et al., 2014], we identified 720 SNP-CpG pairs (including 402 CpG sites and 172 SNPs) among the *trans*-meQTLs that overlapped with an interchromosomal contact, which is 2.9-fold more than expected by chance ($P = 3.7 \times 10^{-126}$; Figure 3.3a,b). The enrichment for Hi-C interchromosomal contacts remained after removing SNPs that were responsible for *trans* bands ($P = 1.7 \times 10^{-61}$). Hence, interchromosomal contacts may produce associations between SNPs and CpGs *in trans*. To characterize the 720 SNP-CpG pairs overlapping with interchromosomal contacts, we examined motif enrichment using three motif enrichment analysis tools (HOMER, PWMEnrich and DEEPbind; Heinz et al. [2010]; Alipanahi et al. [2015]). These analyses showed that the 402 CpG sites involved frequently overlapped with binding sites for CTCF, RAD21 and SMC3 ($P = 2.3 \times 10^{-5}$, $P = 3.5 \times 10^{-5}$ and $P = 5.1 \times 10^{-5}$, respectively), factors known to regulate chromatin architecture [Zuin et al., 2014; Splinter et al., 2006]. An analysis of ChIP-seq data on CTCF binding confirmed this finding (1.8-fold enrichment, $P = 5.2 \times 10^{-7}$).

We next tested whether the *trans*-meQTLs reflected the effect of differential transcription factor binding for transcription factors that mapped close to the SNPs. The rationale for this hypothesis is that binding of transcription factors has been linked to changes in local DNA methylation, primarily loss of methylation upon transcription factor binding and gain of methylation after loss of transcription factor occupancy [Gutierrez-Arcelus et al., 2013; Tsankov et al., 2015]. This model suggests that *trans*-meQTLs may be attributed to SNPs affecting the expression of a transcription factor *in cis* and that the SNP allele preferentially has a unidirectional effect on DNA methylation. In line with this prediction, we observed that, if a SNP was associated with multiple CpG sites in *trans* (at least 10, $n = 305$), the direction of the association of the SNP was consistently skewed toward either increased or decreased DNA methylation. On average, 76% of the CpGs for each *trans*-meQTL SNP displayed the same direction of effect (50%

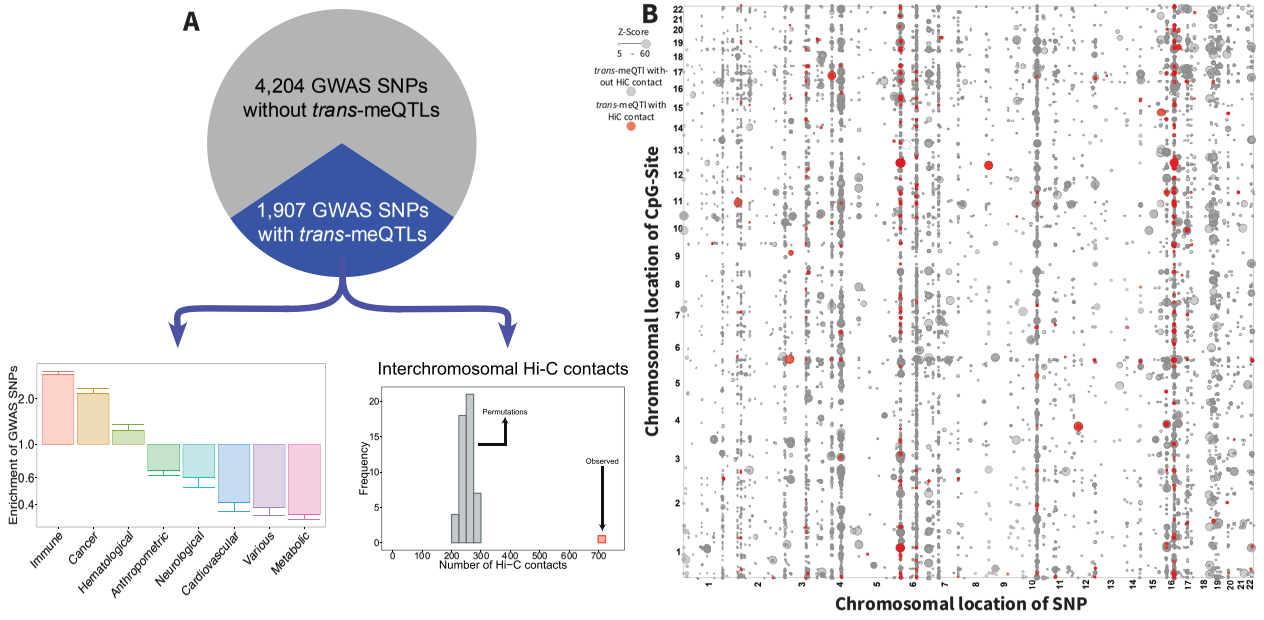

Figure 3.3: (**a**) Distribution of the tested trait-associated SNPs influencing DNA methylation *in trans*. Over 1,900 (31.2%) of all tested SNPs have downstream effects on DNA methylation. Bottom left, for the associated GWAS SNPs, we show the over-representation of SNPs with *trans*-meQTLs in different GWAS trait categories, where the y-axis shows the odds ratio and the bars depict the error margin. Bottom right, Hi-C contacts are over-represented among *trans*-meQTLs. Gray bars show the number of Hi-C contacts using permutated data, and the red bar corresponds to the actually observed number in our data. (**b**) Dot plot depicting the *trans*-meQTLs. Effect strength is reflected by the size of each dot. Red dots correspond to *trans*-meQTLs that overlap with a Hi-C contact site. Several SNPs with widespread *trans*-meQTLs show interchromosomal contacts across the genome, further implicating an important role for those SNPs in development of the associated trait.

expected, $P = 10^{-111}$; Figure 3.4a). A significant skew in the direction of the allelic effect was present for 59.7% of the 305 individual SNPs with at least 10 *trans*-meQTL effects, and this proportion increased to 95.2% for the 104 SNPs with at least 50 *trans*-meQTL effects (binomial $P < 0.05$), suggesting that differential transcription factor binding might explain a substantial fraction of *trans*-meQTLs.

To explore this mechanism further, we combined ChIP-seq data on transcription factor binding at CpGs with the expression effects *in cis* of SNPs to directly examine the involvement of transcription factors in mediating *trans*-meQTLs. Among the trait-associated SNPs influencing at least 10 CpGs *in trans* (n = 305), we identified 13 *trans*-meQTL SNPs with strong support for a role of transcription factors (Figure 3.4a).

The most striking example was a locus on chromosome 4 (Figure 3.4b), where two SNPs (rs3774937 and rs3774959; in strong LD) were associated with ulcerative colitis [Jostins et al., 2012]. The top SNP, rs3774937, was associated with differential DNA methylation at 413 CpG sites across the genome, 92% of which showed the same direction of effect—that is, lower methylation—associated with the minor allele (binomial $P = 2.72 \times 10^{-69}$). Of the 380 CpG sites with lower methylation, 147 (38.7%) overlapped with a nuclear factor (NF)-κB transcription factor binding site (2.75-fold enrichment, $P = 5.3 \times 10^{-32}$), as derived from Encyclopedia of DNA Elements (ENCODE) *NF-κB* ChIP-seq data in blood cell types (Figure 3.4c). Three motif enrichment analysis tools (HOMER, PWMEnrich and DEEPbind) [Heinz et al., 2010; Alipanahi et al., 2015] corroborated the enrichment of *NF-κB*-binding motifs for the 413 CpG sites (Figure 3.4c). Notably, SNP rs3774937 is located in the first intron of *NFKB1*, and we found that the minor allele was associated with higher *NFKB1* expression (Figure 3.4a). Of the 413 CpGs *in trans*, 64 were eQTMs and showed a coherent gene network (Figure 3.4d) that was enriched for immunological processes related to *NFKB1* function [Pers et al., 2015] (Figure 3.4e). Taken together, these results support the idea that the minor allele of rs3774937, which is associated with increased risk of ulcerative colitis, decreases DNA methylation *in trans* by increasing *NFKB1* expression *in cis*.

The same analysis approach indicated that the 779 methylation effects of rs8060686 *in trans* (associated with various phenotypes, including metabolic syndrome [Kristiansson et al., 2012] and coronary heart disease [Lettre et al., 2011]) were mediated by altered CTCF binding, which mapped 315 kb from the *trans*-meQTL SNP. We observed strong CTCF ChIP-seq enrichment (603 of the 779 CpGs *in trans* overlapping with CTCF binding; $P = 1.6 \times 10^{-232}$) and enrichment for CTCF motifs (Figure 3.5). Of these *trans* CpGs, only 13 were observed previously in lymphocytes [Lemire et al., 2015]. Hence, the minor allele of rs8060686 increased DNA methylation *in trans*, which could be attributed to lower *CTCF* gene expression *in cis*.

We found another example of this phenomenon: 228 *trans*-meQTL effects of four SNPs on chromosome 10, mapping near *NKX2-3* and implicated in inflammatory bowel disease [Jostins et al., 2012], were strongly enriched for NKX2 transcription factor motifs and associated with *NKX2-3* expression. Again, a negative correlation was observed, in which the minor allele of rs11190140 decreased DNA methylation *in trans* at NKX2-3-binding sites and increased *NKX2-*

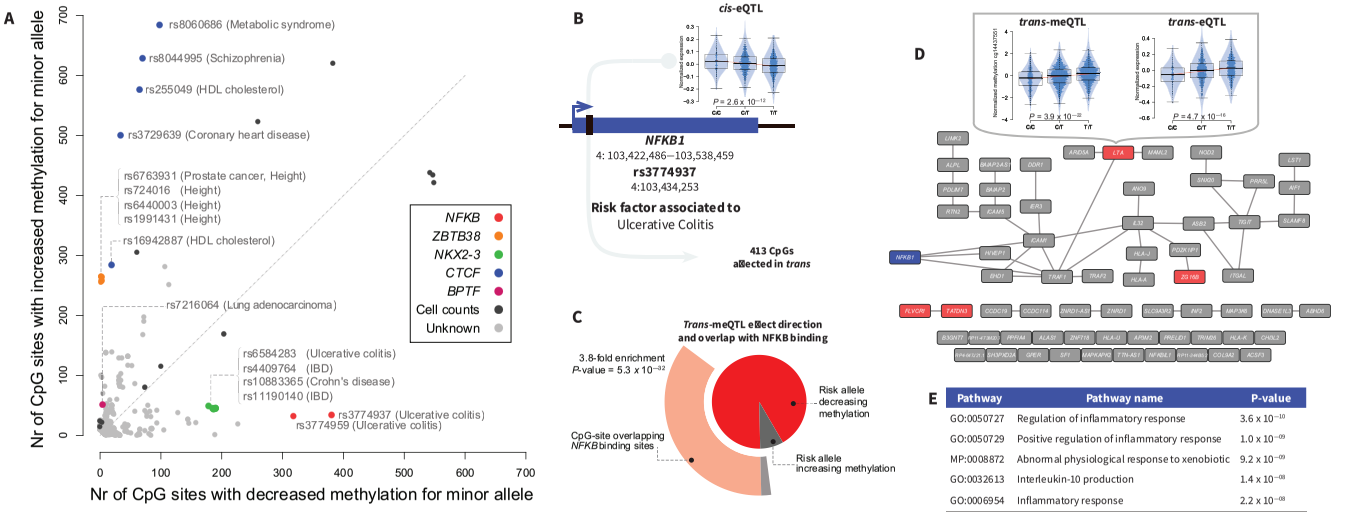

Figure 3.4: (**a**) Each dot represents a SNP with at least ten *trans*-meQTL effects. The x axis shows the number of *trans* effects where the minor allele decreases methylation, and the y axis shows the number of *trans* effects where the minor allele increases methylation. SNPs with a multitude of effects of which many have the same allelic direction often exhibit evidence of a *cis*-eQTL on a transcription factor (colored dots) and an over-representation of *trans*-CpGs overlapping binding sites for that transcription factor. (**b**) Depiction of the *NFKB1* gene and rs3774937, for which the risk and minor allele C is associated with ulcerative colitis and increased expression of *NFKB1*. Boxes show the median and IQR; whiskers extend to the outer quartile plus 1.5 times the IQR. (**c**) In addition to influencing *NFKB1* expression, rs3774937 also relates to DNA methylation at 413 CpGs *in trans*, decreasing methylation levels at 93% of the affected CpG sites (dark gray). Outer chart, many of the CpG sites (37.3%) overlap with NF-κB-binding sites (3.8-fold enrichment, $P = 5.3 \times 10^{-32}$). (**d**) Gene network of the eQTM genes associated with 72 of the 413 CpGs (17.4%) that show a *trans*-meQTL and a *trans*-eQTL (in red). *NFKB1* is depicted in blue. The illustrations above show theobserved *trans*-meQTL (left plot) and *trans*-eQTL (right plot) effects of rs3774937. (**e**) Top pathways as identified by DEPICT for which the genes in **d** were over-represented. Many of the identified pathways are related to inflammation, in line with the inflammatory nature of ulcerative colitis.

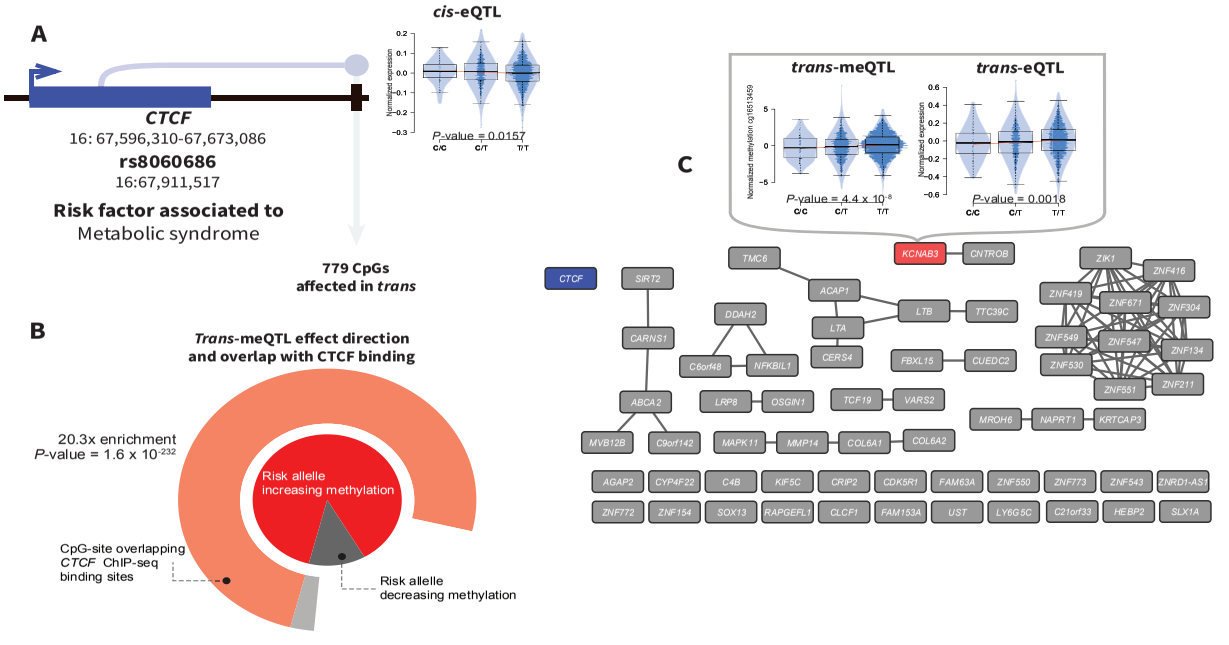

Figure 3.5: (**a**) Depiction of the *CTCF* gene and rs8060686, associated with metabolic syndrome. The plot shows increased expression of *CTCF* for the risk allele C. (**b**) In addition to influencing *CTCF* expression, rs8060686 also influences DNA methylation at 779 CpGs *in trans*, increasing methylation levels at 87.7% of the affected CpG sites (dark gray). Outer chart, many of the CpG sites (77.4%) overlap with CTCF-binding sites (20.3-fold enrichment, $P = 1.6x10^{-232}$). (**c**) Gene network of the genes associated with 60 of the 779 CpGs (7.7%) with a *trans*-meQTL. In the top part of the figure, there is an illustration of overlapping *trans*-meQTL (left) and *trans*-eQTL effects (right) for rs8060686.

3 gene expression *in cis* (Supplementary Figure 5).

A height-associated locus [Soranzo et al., 2009] harboring four SNPs and associated with 267 *trans* CpGs implicated a role for *ZBTB38* in mediating *trans*-meQTL effects (Supplementary Figure 6). In contrast to the aforementioned transcription factors, which are all transcriptional activators, *ZBTB38* is a transcriptional repressor [Filion et al., 2006; Sasai and Defossez, 2009] and its expression was positively correlated with methylation *in trans*, in line with our observation that eQTMs in repressed regions are enriched for positive correlations. Finally, the methylation effects *in trans* of rs7216064 (64 *trans* CpGs), associated with lung carcinoma [Shiraishi et al., 2012], preferentially occurred at regions binding CTCF, while the SNP was located in the *BPTF* gene, which encodes a protein known to occupy CTCF-binding sites [Qiu et al., 2015] (Supplementary Figure 7).

The possibility of linking *trans*-meQTL effects to an association with transcription factor expression *in cis* and concomitant differential methylation *in trans* at the respective binding site for the transcription factor is limited to transcription factors for which ChIP-seq data or motif information is available. To make inferences on transcription factors for which such data are not yet available, we ascertained whether *trans*-meQTL SNPs were more often associated with transcription factor gene expression *in cis* as compared with SNPs without a *trans*-meQTL effect. We observed that 13.1% of the trait-associated SNPs that produced *trans*-meQTLs also affected transcription factor gene expression *in cis*, whereas only 4.5% of the trait-associated SNPs without a *trans*-meQTL affected transcription factor gene expression *in cis* (Fisher's exact $P = 6.6 \times 10^{-13}$).

Here we report that one-third of known disease- and trait-associated SNPs have downstream effects on methylation *in trans* and often are associated with multiple regions across the genome. Our data suggest that the biological mechanism underlying *trans*-meQTLs commonly involves a local effect on the expression of a nearby transcription factor that influences DNA methylation at the distal binding sites of that particular transcription factor. The direction of downstream methylation effects is remarkably consistent for each SNP and indicates that decreased DNA methylation is a signature of increased binding of transcriptional activators. As such, our study identifies the previously unrecognized functional consequences of disease-associated variants in noncoding regions. These can be viewed online (see URLs) and will provide leads for experimental follow-up.

Methods

Cohort descriptions

The five cohorts used in our study are described briefly below. The number of samples per cohort and references to full cohort descriptions can be found in Supplementary Table 1.

CODAM

The Cohort on Diabetes and Atherosclerosis Maastricht (CODAM) [van Greevenbroek et al., 2011] consists of a selection of 547 subjects from a larger population-based cohort [van Dam et al., 2001]. Inclusion of subjects into CODAM was based on a moderately increased risk of developing cardiometabolic diseases, such as type 2 diabetes and/or cardiovascular disease. Subjects were included if they were of European ancestry and over 40 years of age and additionally met at least one of the following criteria: increased body mass index (BMI; > 25), a positive family history for type 2 diabetes, a history of gestational diabetes and/or glycosuria, or use of antihypertensive medication.

LifeLines-DEEP

The LifeLines-DEEP (LLD) cohort [Tigchelaar et al., 2015] is a subcohort of the LifeLines cohort [Scholtens et al., 2015]. LifeLines is a multidisciplinary prospective population-based cohort study examining the health and health-related behaviors of 167,729 individuals living in the northern parts of the Netherlands using a unique three-generation design. It employs a broad range of investigative procedures assessing biomedical, sociodemographic, behavioral, physical and psychological factors contributing to health and disease in the general population. A subset of 1,500 LifeLines participants also take part in LLD [Tigchelaar et al., 2015]. For these participants, additional molecular data are generated, allowing for a more thorough investigation of the association between genetic and phenotypic variation.

LLS

The aim of the Leiden Longevity Study (LLS) [Schoenmaker et al., 2006] is to identify genetic factors influencing longevity and examine their interaction with the environment as a means to develop interventions to increase health at older ages. To this end, long-lived siblings of European descent were recruited together with their offspring and their offspring's partners, on the condition that at least two long-lived siblings were alive at the time of ascertainment. For men, the age criterion was 89 years or older; for women, the age criterion was 91 years or older. These criteria led to the ascertainment of 944 long-lived siblings from 421 families, together with 1,671 of their offspring and 744 partners.

NTR

The Netherlands Twin Register (NTR) [Willemsen et al., 2013; Boomsma et al., 2002, 2008] was established in 1987 to study the extent to which genetic and environmental influences cause phenotypic differences between individuals. To this end, data from twins and their families (nearly 200,000 participants) from all over the Netherlands are collected, with a focus on health, lifestyle, personality, brain development, cognition, mental health and aging.

RS

The Rotterdam Study [Hofman et al., 2013] is a single-center, prospective population-based cohort study conducted in Rotterdam, the Netherlands. Subjects were included in different phases, with a total of 14,926 men and women aged 45 years and over included as of late 2008. The main objective of the Rotterdam Study is to investigate the prevalence and incidence of and risk factors for chronic diseases to contribute to better prevention and treatment of such diseases in the elderly.

Genotype data

Data generation

Genotype data was generated for each cohort individually. Details on the methods used can be found in the individual papers (CODAM [van Dam et al., 2001]; LLD [Tigchelaar et al., 2015]; LLS [Deelen et al., 2014a]; NTR [Willemsen et al., 2013]; RS [Hofman et al., 2013]).

Imputation and QC

For each cohort separately, the genotype data were harmonized toward the Genome of the Netherlands (GoNL) using Genotype Hamonizer [Deelen et al., 2014b] and subsequently imputed per cohort using Impute2 [Howie et al., 2009] using GoNL [Deelen et al., 2014c] reference panel (v5). Quality control was also performed per cohort. We removed SNPs based on imputation info-score (< 0.5), HWE ($P < 10^{-4}$), call rate ($< 95\%$) and minor allele frequency (> 0.05), resulting in 5,206,562 SNPs that passed quality control in each of the data sets.

Methylation data

Data generation

For the generation of genome-wide DNA methylation data, 500 ng of genomic DNA was bisulfite modified using the EZ DNA Methylation kit (Zymo Research) and hybridized on Illumina 450K arrays according to the manufacturer's protocols. The original IDAT files were generated by the Illumina iScan BeadChip scanner. We collected methylation data for a total of 3,841 samples. Data was generated by the Human Genotyping facility (HugeF) of ErasmusMC, The Netherlands (see URLs).

Probe remapping and selection

We remapped the 450K probes to the human genome reference (hg19) to correct for inaccurate mappings of probes and identify probes that mapped to multiple locations on the genome. Details on this procedure can be found in Bonder et al. [2014]. Next, we removed probes with a known SNP (GoNL, MAF > 0.01) at the single base extension (SBE) site or CpG site. Lastly, we removed all probes

on the sex chromosomes, leaving 405,709 high quality methylation probes for the analyses.

Normalization and QC

Methylation data was processed using a custom pipeline based on the pipeline developed by Touleimat and Tost [2012]. First, we used methylumi to extract the data from the raw IDAT files. Next, we removed incorrectly mapped probes and checked for outlying samples using the first two principal components (PCs) obtained using principal component analysis (PCA). None of the samples failed our quality control checks, indicating high quality data. Following quality control, we performed background correction and probe type normalization as implemented in DASEN [Pidsley et al., 2013]. Normalization was performed per cohort, followed by quantile normalization on the combined data to normalize the differences per cohort. We used mix-up mapper [Westra et al., 2011] to identify sample mix-ups between genotype and DNA methylation data, detecting and correcting 193 mix-ups. Lastly, in order to correct for known and unknown confounding sources of variation in the methylation data and increase statistical power, we removed the first components which were not affected by genetic information (22 PCs) from the methylation data using methodology we have successfully used in *trans*-eQTL [Westra et al., 2013; Fehrmann et al., 2011] and meQTL analyses [Touleimat and Tost, 2012].

RNA sequencing

Total RNA from whole blood was depleted of globin transcripts using the Ambion GLOBIN clear kit and subsequently processed for sequencing using the Illumina TruSeq version 2 library preparation kit. Paired-end sequencing of 2×50-bp reads was performed using the Illumina HiSeq 2000 platform, pooling ten samples per lane. Finally, read sets were generated for each sample using CASAVA, retaining only reads passing the Illumina Chastity Filter for further processing. Data were generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (see URLs).

Initial quality control was performed using FastQC v0.10.1 (see URLs), removal of adaptors was performed using cutadapt [Martin, 2011] (v1.1) and Sickle v1.2 (see URLs) was used to trim low-quality ends from the reads (min length 25, min quality 20). Sequencing reads were mapped to the human genome (hg19) using STAR [Dobin et al., 2013] v2.3.125. Gene expression quantification was performed by HTseq-count. The gene definitions used for quantification were based on Ensembl version 71, with the extension that regions with overlapping exons were treated as separate genes and reads mapping within these overlapping parts did not count toward expression of the normal genes.

Expression data on the gene level were first normalized using trimmed mean of M values [Robinson and Oshlack, 2010]. Then, expression values were log2 transformed, and gene and sample means were centered to zero. To correct for batch effects, principal-component analysis (PCA) was run on the sample

correlation matrix and the first 25 principal components were removed using methodology that we have used before [Westra et al., 2013; Fehrmann et al., 2011]; details are provided in Zhernakova et al. [2016].

Cis-meQTL mapping

To determine the effect of nearby genetic variation on methylation levels (*cis*-meQTL, here defined as the relationship between a CpG and a SNP no further than 250 kb apart), we performed *cis*-meQTL mapping using 3,841 samples for which both genotype data and methylation data were available. To this end, we calculated the Spearman rank correlation for each cohort, followed by meta-analysis using a weighted Z-method described previously [Westra et al., 2013]. To detect all possible independent SNPs regulating methylation at a single CpG site, we regressed out all primary *cis*-meQTL effects and then performed *cis*-meQTL mapping for the same CpG site to find secondary *cis*-meQTLs. We repeated this in a stepwise fashion until no more independent *cis*-meQTLs were found.

To filter out potential false positive *cis*-meQTLs caused by SNPs affecting the binding of a probe on the array, we filtered the *cis*-meQTL effects by removing any CpG-SNP pairs for which the SNP was located in the probe. In addition, all other CpG-SNP pairs for which the SNP was outside the probe but in LD ($r^2 > 0.2$ or D$\prime > 0.2$) with a SNP inside the probe were also removed. We tested for LD between SNPs in probes and in surrounding *cis* areas in the individual genotype data sets, as well as in GoNL v5, to be as strict as possible in marking a QTL as a true positive.

To correct for multiple testing, we empirically controlled the FDR at 5%. For this, we compared the distribution of observed P values to the distribution obtained from performing the analysis on permuted data. Permutation was performed by shuffling the sample identifiers of one data set, thereby breaking the link between, for example, the genotype data and the methylation or expression data. We repeated this procedure ten times to obtain a stable distribution of P-values under the null distribution. The FDR was determined by only selecting the strongest effect for each CpG [Westra et al., 2013] in both the real analysis and the permutations (probe-level FDR $< 5\%$).

Cis-eQTL mapping

For a set of 2,116 BIOS samples we had also generated RNA-seq data. We used this data to identify *cis*-eQTLs. *Cis*-eQTL mapping was performed using the same method as *cis*-meQTL mapping. Details on these eQTLs are described in a separate paper [Zhernakova et al., 2016].

Expression quantitative trait methylation analysis

To identify associations between methylation levels and the expression levels of nearby genes (*cis*-eQTMs), we first corrected our expression and methylation data for batch effects and covariates by regressing out the principal components

and regressing out the identified *cis*-meQTLs and *cis*-eQTLs, to ensure that the associations identified between CpG sites and gene expression levels were not due to shared genetic effects. We mapped the eQTMs in a window of 250 kb around the TSS of a transcript. Further statistical analysis was identical to that for *cis*-meQTL mapping. For this analysis, we were able to use a total of 2,101 samples for which both genetic, methylation and gene expression data were available. To correct for multiple testing, we controlled the FDR at 5%; the FDR was determined by only selecting the strongest effect for each CpG [Westra et al., 2013] in both the real analysis and the permutations.

Trans-meQTL mapping

To identify the effects of distal genetic variation on methylation (*trans*-meQTLs), we used the same 3,841 samples that we had used for cis-meQTL mapping. To focus our analysis and limit the multiple-testing burden, we restricted our analysis to SNPs that have previously been found to be significantly correlated with traits and diseases. We extracted these SNPs from the NHGRI GWAS catalog and also used recent GWAS not yet in the NHGRI GWAS catalog and studies on the Immunochip and Metabochip platforms that are not included in the NHGRI GWAS catalog. We compiled this list of SNPs in December 2014. For each SNP, we only investigated CpG sites that mapped at least 5 Mb from the SNP or on other chromosomes. Before mapping *trans*-meQTLs, we regressed out the identified cis-meQTLs to increase the statistical power of *trans*-meQTL detection (as done previously for *trans*-eQTLs [Westra et al., 2013]) and to avoid designating an association as *trans* that might be due to long-range LD (for example, within the human leukocyte antigen (HLA) region). To ascertain the stability of the *trans*-meQTLs, we also performed *trans* mapping using uncorrected methylation data and data corrected for cell type proportions. In addition, we performed meQTL mapping on SNPs known to influence cell type proportions in blood [Orru et al., 2013; Roederer et al., 2015].

To filter out potential false positive *trans*-meQTLs due to cross-hybridization of the probe, we remapped the methylation probes with very relaxed settings identical to those used in Westra et al. [2013], with the difference that we only accepted mappings if the last bases of the probe including the SBE site were accurately mapped to the alternative location. If the probe mapped within our minimal *trans* window, 5 Mb from the SNP, we removed the effect as being a false positive *trans*-meQTL.

We controlled the FDR at 5%, identical to in the aforementioned *cis*-meQTL analysis.

Trans-eQTL mapping

To check whether *trans*-meQTL effects also showed in gene expression levels, we annotated the CpGs with a *trans*-meQTL to genes using our eQTMs. Using the 2,101 samples for which both genotype and gene expression data were available,

we performed *trans*-eQTL mapping, associating SNPs known to be associated with DNA methylation in *trans* with their corresponding eQTM genes.

Annotation and enrichment tests

Annotation of CpG sites was performed using Ensembl [Flicek et al., 2013] (v70), the UCSC Genome Browser [Kent et al., 2002] and data from the Epigenomics Roadmap project [Kundaje et al., 2015]. We used Epigenomics Roadmap annotation for the SBE site of the methylation site using 27 blood cell types. We used both the histone mark information and the chromatin marks in blood-related cell types only, as generated by the Epigenomics Roadmap project. Summarizing the information over the 27 blood cell types was carried out by counting the presence of histone marks in all the cell types and scaling the abundance: that is, the score would be 1 if a mark is bound in all cell types, whereas the score would be 0 if it is present in none of the blood cell types.

To calculate enrichment of meQTLs or eQTMs for any particular genomic context, we used logistic regression because this allowed us to account for covariates such as CpG methylation variation. For *cis*-meQTLs, we used the variability in DNA methylation, the number of SNPs tested and the distance to the nearest SNP for each CpG as covariates. For all other analyses, we used only the variability in DNA methylation as a covariate.

We used transcription factor ChIP-seq data from the ENCODE project for blood-related cell lines (narrow-peak data). We overlapped CpG locations with ChIP-seq signals and performed a Fisher's exact test to determine whether the *trans*-meQTL probes associated with a SNP overlapped a ChIP-seq region more often than other *trans*-meQTL probes.

Enrichment of known sequence motifs among *trans*-CpGs was assessed using the PWMEnrich package in R, HOMER [Heinz et al., 2013] and DEEPbind [Alipanahi et al., 2015]. For PWMEnrich, the 100-bp sequence around each interrogated CpG site was used, and as a background set we used the top CpGs from the 50 permutations used to determine the FDR threshold of the *trans*-meQTLs. For HOMER, the default settings for the identification of motif enrichment were used, and the same CpG sites derived from the permutations were used as background. For DEEPbind, we used both the permutation background as described for HOMER and the permutation background as described for PWMEnrich.

Using data published by Rao et al. [2014], we were able to intersect the *trans*-meQTLs with information about the 3D structure of the human genome using combined Hi-C data for both inter- and intrachromosomal data at 1 kb and the quality threshold of E30 in the GM12878 LCL. Both the *trans*-meQTL SNPs and *trans*-meQTL probes were put in the relevant 1-kb blocks, and for these blocks we looked up the chromosomal contact value in the measurements by Rao *et al*. Surrounding the *trans*-meQTL SNPs, we used an LD window that spanned maximally 250 kb from the *trans*-meQTL SNP and had a minimal r^2 value of 0.8. If a Hi-C contact was indicated between a SNP block and a CpG site, we flagged the region as positive for Hi-C contacts. As background, we used the combinations

found in our 50 permutated *trans*-meQTL analyses, taking for each permutation the top *trans*-meQTLs that were similar in size to those from the real analysis.

Prediction of eQTM direction

We predicted the direction of eQTM effects using both a decision tree and a naive Bayes model (as implemented by Rapid-miner v6.3 [Hofmann and Klinkenberg, 2013]). We built the models on the strongest eQTMs (FDR $< 9.73 \times 10^{-6}$). For the decision tree, we used a standard cross-validation setup with 20 folds. For the naive Bayesian model, we used double-loop cross-validation: performance was evaluated in the outer loop using 20-fold cross-validation, while feature selection (using both backward elimination and forward selection) took place in the inner loop using tenfold cross-validation. Details about double-loop cross-validation can be found in de Ronde et al. [2014]. During the training of the model, we balanced the two classes, making sure we had an equal number of positively correlating and negatively correlating CpG-gene combinations, by randomly sampling a subset of the over-represented negatively correlating CpG-gene combination group. We chose to do so to circumvent labeling al eQTMs as negative, as this is the class to which the majority of the eQTMs belonged.

In the models, we used CpG-centric annotations: overlap with Epigenomics Roadmap chromatin states, histone marks and relationships between the histone marks, GC content surrounding the CpG site and relative locations from the CpG site to the transcript.

DEPICT

To investigate whether there was biological coherence in the *trans*-meQTLs identified for the *NFKB1* locus, we performed gene set enrichment analysis for the genes near the *trans*-CpG sites of the ulcerative colitis genetic risk factor (which maps in the *NFKB1* locus). To do so, we adapted DEPICT [Pers et al., 2015], a pathway enrichment analysis method that we originally developed for GWAS. Instead of defining loci with genes by using the top associated SNPs (as is done when analyzing GWAS data), we used the eQTM information to empirically link *trans*-CpGs to genes (that map close to the CpGs). Within DEPICT gene set enrichment, significance is determined by using a background set of genes. As background in the adapted DEPICT enrichment analyses, we matched our background to the results from the actual *trans*-meQTL and eQTM analyses: matching was performed by generating a set of background CpGs (and corresponding correlating eQTM genes), by selecting an equal number of CpGs for which we had found *trans*-meQTL effects with SNPs that map outside the *NFKB1* locus. By doing so, we ensured that the characteristics of these background CpGs were the same as those for the real *NFKB1 trans*-meQTL CpGs, both in terms of CpG variance and the requirement that they also show a significant correlation with expression levels of genes close to the CpG (that is, a *cis*-eQTM), ensuring that the corresponding input genes for DEPICT had the same expression variation distribution in the actual *NFKB1* analysis and in the background. Subsequent

pathway enrichment analysis was conducted as described before [Pers et al., 2015], and significance was determined by controlling the FDR at 5%.

URLs

All results can be queried using our dedicated QTL browser at http://www.genenetwork.nl/biosqtlbrowser. Data were generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (http://www.glimDNA.org/). Cohort webpages are as follows: LifeLines, http://lifelines.nl/lifelines-research/general; Leiden Longevity Study, http://www.healthy-ageing.nl/ and http://www.leidenlangleven.nl/; Netherlands Twin Registry, http://www.tweelingenregister.org/; Rotterdam Studies, http://www.erasmusmc.nl/epi/research/The-Rotterdam-Study/; Genetic Research in Isolated Populations program, http://www.epib.nl/research/geneticepi/research.html#gip; CODAM study, http://www.carimmaastricht.nl/; PAN study, http://www.alsonderzoek.nl/. Software used included the following: FastQC, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; Sickle, https://github.com/najoshi/sickle; PWMEnrich: PWM enrichment analysis v.4.6.0, https://bioconductor.riken.jp/packages/3.2/bioc/html/PWMEnrich.html.

Accession codes

All results can be queried using our dedicated QTL browser (see URLs). Raw data were submitted to the European Genome-phenome Archive (EGA) under accession EGAS00001001077.

References

Alipanahi, B. et al. [2015]. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning, *Nat. Biotechnol.* **33**: 831–838.

Bernstein, B. E., Meissner, A. and Lander, E. S. [2007]. The mammalian epigenome, *Cell* **128**: 669–681.

Bonder, M. J. et al. [2014]. Genetic and epigenetic regulation of gene expression in fetal and adult human livers, *BMC Genomics* **15**: 860.

Boomsma, D. I. et al. [2002]. Netherlands twin register: a focus on longitudinal research, *Twin Res.* **5**: 401–406.

Boomsma, D. I. et al. [2008]. Genome-wide association of major depression: description of samples for the gain major depressive disorder study: Ntr and nesda biobank projects, *Eur. J. Hum. Genet.* **16**: 335–342.

de Ronde, J. J. et al. [2014]. Breast cancer subtype specific classifiers of response to neoadjuvant chemotherapy do not outperform classifiers trained on all subtypes, *PLoS One* **9**: e88551.

Deelen, J. et al. [2014a]. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age, *Hum. Mol. Genet.* **23**: 4420–4432.

Deelen, P. et al. [2014b]. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration, *BMC Res. Notes* **7**: 901.

Deelen, P. et al. [2014c]. Improved imputation quality of low-frequency and rare variants in european samples using the 'genome of the netherlands', *Eur. J. Hum. Genet.* **22**: 1321–1326.

Dobin, A. et al. [2013]. Star: ultrafast universal rna-seq aligner, *Bioinformatics* **29**: 15–21.

Fehrmann, R. S. N. et al. [2011]. Trans-eqtls reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the hla, *PLoS Genet.* **7**: e1002197.

Filion, G. J. P. et al. [2006]. A family of human zinc finger proteins that bind methylated dna and repress transcription, *Mol. Cell. Biol.* **26**: 169–181.

Flicek, P. et al. [2013]. Ensembl 2013, *Nucleic Acids Res.* **41**: D48–D55.

Gutierrez-Arcelus, M. et al. [2013]. Passive and active dna methylation and the interplay with genetic variation in gene regulation, *eLife* **2**.

Heinz, S. et al. [2010]. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities, *Mol. Cell* **38**: 576–589.

Heinz, S. et al. [2013]. Effect of natural genetic variation on enhancer selection and function, *Nature* **503**: 487–492.

Hofman, A. et al. [2013]. The rotterdam study: 2014 objectives and design update, *Eur. J. Epidemiol.* **28**: 889–926.

Hofmann, M. and Klinkenberg, R. [2013]. *Rapid Miner Data Mining Use Cases and Business Analytics Applications* .

Howie, B. N. et al. [2009]. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, *PLoS Genet* **5**(6): e1000529.

Hu, S. et al. [2013]. Dna methylation presents distinct binding sites for human transcription factors, *eLife* **2**: e00726.

Huan, T. et al. [2015]. A meta-analysis of gene expression signatures of blood pressure and hypertension, *PLoS Genet.* **11**: e1005035.

Jostins, L. et al. [2012]. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease, *Nature* **491**: 119–124.

Kent, W. J. et al. [2002]. The human genome browser at ucsc, *Genome Res.* **12**: 996–1006.

Kristiansson, K. et al. [2012]. Genome-wide screen for metabolic syndrome susceptibility loci reveals strong lipid gene contribution but no evidence for common genetic basis for clustering of metabolic syndrome traits, *Circ Cardiovasc Genet* **5**: 242–249.

Kundaje, A. et al. [2015]. Integrative analysis of 111 reference human epigenomes, *Nature* **518**: 317–330.

Lemire, M. et al. [2015]. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci, *Nat. Commun.* **6**: 6326.

Lettre, G. et al. [2011]. Genome-wide association study of coronary heart disease and its risk factors in 8,090 african americans: the nhlbi care project, *PLoS Genet.* **7**: e1001300.

Manolio, T. A. [2010]. Genomewide association studies and assessment of the risk of disease, *N. Engl. J. Med.* **363**: 166–176.

Martin, M. [2011]. Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal* **17**: 10–12.

Mill, J. and Heijmans, B. T. [2013]. From promises to practical strategies in epigenetic epidemiology, *Nat Rev Genet* **14**(8): 585–594.

Orru, V. et al. [2013]. Genetic variants regulating immune cell levels in health and disease, *Cell* **155**: 242–256.

Pers, T. H. et al. [2015]. Biological interpretation of genome-wide association studies using predicted gene functions, *Nat. Commun.* **6**: 5890.

Pidsley, R. et al. [2013]. A data-driven approach to preprocessing illumina 450k methylation array data, *BMC Genomics* **14**: 293.

Qiu, Z. et al. [2015]. Functional interactions between nurf and ctcf regulate gene expression, *Mol. Cell. Biol.* **35**: 224–237.

Rao, S. S. P. et al. [2014]. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping, *Cell* **159**: 1665–1680.

Robinson, M. D. and Oshlack, A. [2010]. A scaling normalization method for differential expression analysis of rna-seq data, *Genome Biol.* **11**: R25.

Roederer, M. et al. [2015]. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis, *Cell* **161**: 387–403.

Sasai, N. and Defossez, P. A. [2009]. Many paths to one goal?: The proteins that recognize methylated dna in eukaryotes, *Int. J. Dev. Biol.* **53**: 323–334.

Schoenmaker, M. et al. [2006]. Evidence of genetic enrichment for exceptional survival using a family approach: the leiden longevity study, *Eur. J. Hum. Genet.* **14**: 79–84.

Scholtens, S. et al. [2015]. Cohort profile: Lifelines, a three-generation cohort study and biobank, *Int. J. Epidemiol.* **44**: 1172–1180.

Shiraishi, K. et al. [2012]. A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the japanese population, *Nat. Genet.* **44**: 900–903.

Soranzo, N. et al. [2009]. Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size, *PLoS Genet.* **5**: e1000445.

Splinter, E. et al. [2006]. Ctcf mediates long-range chromatin looping and local histone modification in the [beta]-globin locus, *Genes Dev.* **20**: 2349–2354.

Tigchelaar, E. F. et al. [2015]. Cohort profile: Lifelines deep, a prospective, general population cohort study in the northern netherlands: study design and baseline characteristics, *BMJ Open* **5**: e006772.

Touleimat, N. and Tost, J. [2012]. Complete pipeline for infinium([reg]) human methylation 450k beadchip data processing using subset quantile normalization for accurate dna methylation estimation, *Epigenomics* **4**: 325–341.

Tsankov, A. M. et al. [2015]. Transcription factor binding dynamics during human es cell differentiation, *Nature* **518**: 344–349.

van Dam, R. M. et al. [2001]. Parental history of diabetes modifies the association between abdominal adiposity and hyperglycemia, *Diabetes Care* **24**: 1454–1459.

van Greevenbroek, M. M. J. et al. [2011]. The cross-sectional association between insulin resistance and circulating complement c3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the codam study), *Eur. J. Clin. Invest.* **41**: 372–379.

Visscher, P. M. et al. [2012]. Five years of gwas discovery, *Am. J. Hum. Genet.* **90**: 7–24.

Westra, H. et al. [2011]. Mixupmapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects, *Bioinformatics* **27**(15): 2104–2111.

Westra, H. et al. [2013]. Systematic identification of trans eqtls as putative drivers of known

disease associations, *Nat Genet* **45**(10): 1238–1243.

Willemsen, G. et al. [2013]. The adult netherlands twin register: twenty-five years of survey and biological data collection, *Twin Res. Hum. Genet.* **16**: 271–281.

Wright, F. A. et al. [2014]. Heritability and genomics of gene expression in peripheral blood, *Nat. Genet.* **46**: 430–437.

Yao, C. et al. [2015]. Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes, *Circulation* **131**: 536–549.

Zhernakova, D. V. et al. [2016]. Identification of context-dependent expression quantitative trait loci in whole blood, *Nat. Genet.* .

Zuin, J. et al. [2014]. Cohesin and ctcf differentially affect chromatin architecture and gene expression in human cells, *Proc. Natl. Acad. Sci. USA* **111**: 996–1001.

4 AUTOSOMAL GENETIC VARIATION IS ASSOCIATED WITH DNA METHYLATION IN REGIONS VARIABLY ESCAPING X-CHROMOSOME INACTIVATION

René Luijk, H. Wu, C.K. Ward-Caviness, E. Hannon, E. Carnero-Montoro, J.L. Min, P. Mandaviya, M. Müller-Nurasyid, H. Mei, S.M. van der Maare, BIOS Consortium, C. Relton, J. Mill, M. Waldenberger, J.T. Bell, R. Jansen, A. Zhernakova, L. Franke, P.A.C. 't Hoen, D.I. Boomsma, C.M. van Duijn, M.M.J. van Greevenbroek, J.H. Veldink, C. Wijmenga, J. van Meurs, L. Daxinger, P.E. Slagboom, E.W. van Zwet, B.T. Heijmans

Nature Communications, **9**(1) (2018)

Abstract

The inactivation of one of the female X chromosomes restores equal expression of X-chromosomal genes between females and males. While most of the X-chromosomal genes are silenced by X-chromosome inactivation (XCI), 15% of genes remain bi-allelically expressed, and 10% show variable degrees of escape from XCI between females. However, little is known about genes involved in human XCI, or the causes of variable XCI. Using a discovery data-set of 1,867 females and 1,398 males and an independent replication sample of 3,351 females, we show that genetic variation at three autosomal loci is associated with female-specific changes in X-chromosome methylation. Through cis-eQTL expression analysis in the same 1,867 females, we mapped the loci to the genes *SMCHD1/METTL4*, *TRIM6/HBG2*, and *ZSCAN9*. Low-expression alleles of the loci were predominantly associated with mild hypomethylation of CpG islands near genes known to variably escape XCI, implicating the autosomal genes in variable XCI. Together, these results suggest a genetic basis for variable escape from XCI and highlight the potential of a population genomics approach to identify genes involved in XCI.

Introduction

To achieve dosage equivalency between male and female mammals, one of two X-chromosomes is silenced early in female embryonic development resulting in one inactive (Xi) and one active (Xa) copy of the X-chromosome [Lyon, 1961]. While the Xi-linked gene *XIST* is crucial for the initiation of X-chromosome inactivation (XCI), autosomal genes appear to be critically involved in XCI establishment and maintenance [Galupa and Heard, 2015]. An abundance of repressive histone marks [Brinkman et al., 2006; Heard et al., 2001; Plath et al., 2003] and DNA methylation [Sharp et al., 2011; Yasukochi et al., 2010] throughout XCI on Xi is in line with a prominent role of epigenetic regulation in both phases. However, the Xi is not completely inactivated. With an estimated 15% of X-chromosomal genes consistently escaping XCI, and an additional 10% escaping XCI to varying degrees [Carrel and Willard, 2005; Cotton et al., 2013], escape from XCI is fairly common in humans [Carrel and Willard, 1999; Cotton et al., 2014; Zhang et al., 2013], much more so than in mice [Yang et al., 2010]. Genes escaping XCI are characterized by distinct epigenetic states [Peeters et al., 2014] and are thought to be associated with adverse outcomes, including mental impairment [Peeters et al., 2014; Yang et al., 2010; Zhang et al., 2013].

In the mouse, an example of an autosomal gene involved in XCI is *Smchd1*. *Smchd1* is an epigenetic repressor and plays a critical role in the DNA methylation maintenance of XCI in mice [Blewitt et al., 2008; Nozawa et al., 2013]. However, in humans, in-depth knowledge on the role of autosomal genes in XCI maintenance is lacking, despite earlier *in vitro* efforts [Massah et al., 2014]. Furthermore, the mechanisms underlying variable XCI, a common feature of human XCI [Carrel and Willard, 2005; Cotton et al., 2013], are unknown.

Here, we report on the identification of four autosomal loci associated with female-specific changes in X-chromosome DNA methylation using a discovery set of 1,867 females and 1,398 males, and replication of three of these loci in an independent replication set consisting of 3,351 female samples. The replicated loci map to the genes *SMCHD1/METTL4*, *TRIM6/HBG2* and *ZSCAN9* through eQTL analysis. All three preferentially influenced the methylation of CpGs located in CpG islands near genes known to variably escape XCI between individuals, providing evidence for a genetic basis of this phenomenon.

Results

Identification of female-specific genetic effects on X-chromosome methylation

To identify genetic variants involved in XCI, we employed a global test approach [Luijk et al., 2015] to evaluate the association of 7,545,443 autosomal genetic variants with DNA methylation at any of 10,286 X-chromosomal CpGs measured in whole blood of 1,867 females (Supplementary Data 1) using the Illumina 450k array (see Methods). The analysis was corrected for covariates, including

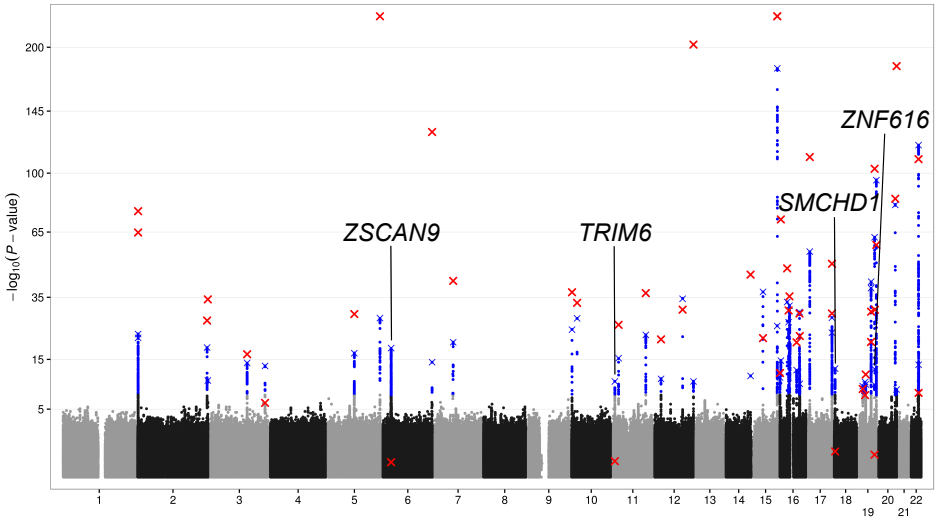

Figure 4.1: Manhattan plot showing all tested autosomal SNPs for an overall effect on X-chromosomal methylation in females. Significant associations are depicted in blue (Wald $P < 5 \times 10^{-8}$). The sentinel variant per independent locus is indicated with a blue cross. Testing the effects of these 48 sentinel variants in males, we found 44 replicated in males (Wald $P < 1.1 \times 10^{-6}$, red cross), whereas the other 4 loci were female-specific, as they clearly lacked an effect in males (Wald $P > 0.19$).

cell counts, age, and batch effects. We identified 4,504 individual variants representing 48 independent loci associated with X-chromosomal methylation in females (Wald $P < 5 \times 10^{-8}$, Figure 4.1 and Supplementary Fig. 1), each defined by the most strongly associated variant (as reflected by the lowest P-value), termed the sentinel variant. Of the 48 sentinel variants corresponding to these 48 loci, 44 were also associated with X-chromosomal methylation in males ($N = 1,398$, Supplementary Data 1, Supplementary Data 2; Wald $P < 1.1x10^{-6}$) indicating that the associations were unrelated to XCI. The four remaining variants did not show any indication for an effect in males (Wald $P > 0.19$) while they did show strong, widespread, and consistent same-direction effects across the X-chromosome in females (Supplementary Data 2, Supplementary Data 3, Supplementary Fig. 2). Formally testing for a genotype by sex interaction revealed significant interaction effects for three of the four variants. The rs140837774, rs139916287, and rs1736891 variants with evidence for an interaction ($P_{interaction} < 5.9 \times 10^{-4}$) mapped to the *SMCHD1/METTL4*, *TRIM6/HBG2*, and *ZSCAN9* loci, respectively, (see Methods). The remaining variant rs73937272 ($P_{interaction} = 0.88$) mapped near the *ZNF616* gene. Finally, we evaluated whether the effect of the autosomal loci was influenced by genetic variation on X, but this did not change the results (Supplementary Data 4, see Methods).

To establish the validity and stability of the analyses, we first investigated whether any of the associations were due to confounding by cellular heterogeneity. Therefore, we directly tested for an association between the four identified sentinel variants and the observed red and white blood cell counts. This did not result in any significant association (Supplementary Fig. 3). Furthermore, we determined that none of the four identified variants are among the variants known to affect blood composition [Orru et al., 2013; Roederer et al., 2015]. Vice versa, genetic variants known to affect blood cell counts also did not show an association with X-chromosomal methylation in our data (Supplementary Fig. 4, Supplementary Data 5). Re-testing the effects of the four sentinel variants while adjusting for nearby blood composition-associated SNPs (< 1Mb) did not influence the results (Supplementary Data 6, see Methods). Second, we addressed unknown confounding by including latent factors as covariates in our models, estimated in the methylation data using software for estimation and adjustment of unknown confounders in high-dimensional data (Wang et al. [2015], see Methods). Re-testing the four sentinel variants without these latent factors did not change the results (Supplementary Data 6). We conclude that the effects of the four variants identified in the discovery data are stable and not confounded by cellular heterogeneity or other, unknown, factors.

Finally, we tested the four sentinel variants in an additional 3,351 unrelated female samples (see Methods and Supplementary Data 7), and successfully replicated the rs140837774, rs139916287, and rs1736891 variants (Bonferroni corrected, $P_{adj} = 0.0096$, $P_{adj} = 2.4 \times 10^{-4}$, and $P_{adj} = 2.2 \times 10^{-3}$, respectively). The rs73937272 variant ($P_{adj} = 1$), which also lacked a sex-genotype interaction effect in the discovery set, was not replicated. In further analyses, we focussed on the three replicated loci.

Exploration of genetic loci with female-specific effects on X-methylation

The sentinel variant rs140837774 is an AATTG insertion/deletion variant (MAF = 0.49) on chromosome 18, located in intron 26 of *SMCHD1* (Supplementary Fig. 2), a gene known to be critically involved in XCI in mice [Blewitt et al., 2008; Chen et al., 2015; Daxinger et al., 2013; Gendrel et al., 2012, 2013]. In addition, *SMCHD1* mutations affect the methylation levels of the *D4Z4* repeat in humans, playing an important role in facioscapulohumeral dystrophy 2 (FSHD2, Lemmers et al. [2012]). To link rs140837774 to a nearby gene we performed a *cis*-eQTL analysis using RNA-seq data from the 1,867 females in the discovery set of our study (250Kb up- and downstream of the sentinel variant, Supplementary Data 8). We found that the deletion was strongly associated with decreased *SMCHD1* expression (Fisher's $P = 1.8 \times 10^{-10}$, regression coefficient = -0.13) and increased expression of the methyltransferase *METTL4*, albeit weaker (Wald $P = 4.9 \times 10^{-4}$, regression coefficient = 0.04). *METTL4* is a highly conserved gene [Breiling and Lyko, 2015; Falckenhayn et al., 2016], involved in the mRNA modification N^6-methyladenosine (reviewed in Gilbert et al. [2016]), which plays an important role in epigenetic regulation in mammals [Wu et al., 2016].

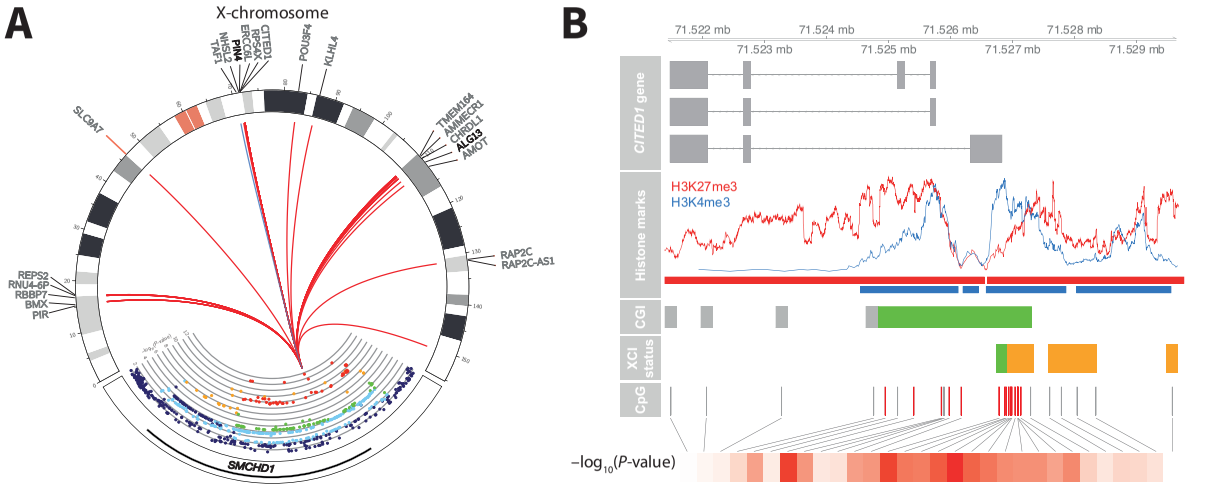

Figure 4.2: The *SMCHD1/METTL4* locus associates with DNA methylation at X-chromosomal and autosomal regions. **(A)** Plot showing the *SMCHD1/METTL4* locus and the effects it has on the X-chromosome. The colors in the *SMCHD1/METTL4* locus indicate LD (red: $R^2 \geq 0.8$; orange: $0.6 \leq R^2 < 0.8$; green: $0.4 \leq R^2 < 0.6$; light blue: $0.2 \leq R^2 < 0.4$; dark blue: $R^2 \leq 0.6 \leq 0.2$). The y-axis shows the –$\log_{10}$(P-value) of the association with overall X-chromosomal methylation. The line colors in the Circos plot indicate the direction of the effect (red: hypomethylation, blue: hypermethylation). The plot of the *SMCHD1/METTL4* locus mostly shows moderate to high LD and covers most of the *SMCHD1* gene. The deletion of rs140837774 is associated with both downregulation of *SMCHD1* and upregulation of *METTL4* (see main text). The deletion of rs140837774 is associated with hypomethylation at 98.2% of all associated CpGs (56 CpGs, red lines, mean effect size 1% per allele). Hypomethylation of CpGs near two genes known to escape XCI to varying degrees (PIN4 and ALG13, shown in bold) is associated with increased expression of these two genes. **(B)** Example of CpG island (CGI) in the CITED1 gene (first row) associated with the *SMCHD1/METTL4* locus. The CpGs associated with this locus (fifth row, indicated by red lines) are overrepresented in regions characterized by both active (blue) and repressive (red) histone marks (second row, red and blue bars, 2-fold enrichment, Fisher's $P = 1.5 \times 10^{-12}$; 6.9-fold enrichment, Fisher's $P = 1.5 \times 10^{-12}$), are often located in CpG islands (third row, green bar, 11.3-fold enrichment, Fisher's $P = 2.5 \times 10^{-14}$), and regions known to variably escape X-chromosome inactivation (fourth row, orange bars, 21.4-fold enrichment, Fisher's $P = 3.7 \times 10^{-18}$). The bottom row indicates the strength of the associations in terms of –$\log_{10}$(P-value) (dark red indicates strong associations).

The *SMCHD1/METTL4* variant was associated with altered methylation levels of 57 X-chromosomal CpGs in females (FDR < 0.05, Figure 4.2A, Supplementary Data 9). The deletion (the low *SMCHD1* expression allele) was associated with hypomethylation at 56 of those X-chromosomal CpGs (98.2%, binomial $P = 8.5 \times 10^{-13}$ Figure 4.2A), consistent with X-hypomethylation in female mice deficient for *SMCHD1* [Gendrel et al., 2013]. The mean effect size was 1% per rare allele (ranging from 0.27% to 2.34%, Supplementary Fig. 5), with the mean methylation values per CpG ranging from 2.6% to 55% (average methylation at these 56 CpGs is 23.6%).

Compared to all X-chromosomal CpGs in our data, the associated X chromosomal CpGs were strongly overrepresented in CpG islands (50 out of 57 CpGs, 11.3-fold enrichment, binomial $P = 2.5 \times 10^{-14}$, Figure 4.2B), in line with *SMCHD1*'s role in X-chromosomal CpG island methylation [Gendrel et al., 2012]. Data on chromatin marks in blood (Kundaje et al. [2015], see Methods) revealed a strong overrepresentation of the associated X-chromosomal CpGs in regions bivalently marked by the active histone mark H3K4me3 (47 CpGs, 8.2-fold enrichment, Fisher's $P = 1.5 \times 10^{-12}$), and the repressive mark H3K27me3 (38 CpGs, 6.9-fold enrichment, Fisher's $P = 1.5 \times 10^{-12}$), as compared to all X-chromosomal CpGs in our data. In agreement with this, we observed a 16.9-fold enrichment for CpGs overlapping bivalent/poised transcription start sites (TSSs) (35/57 CpGs, Fisher's $P = 4.3 \times 10^{-23}$) using predicted chromatin segmentations [Kundaje et al., 2015], possibly reflecting the mixed signals from both the active and inactive X chromosomes underlying these chromatin segmentations. Strikingly, annotation by the degree of escape for 489 TSSs in 27 different tissues, and specifically whole blood (Cotton et al. [2014], see Methods), revealed a strong overrepresentation of CpGs located near TSSs variably escaping XCI (22 CpGs, 21.4-fold enrichment, Fisher's $P = 3.7 \times 10^{-18}$, Figure 4.2B). Only a modest enrichment for associated CpGs in fully escaping XCI regions (15 CpGs, 4.2-fold enrichment, Fisher's $P = 4.5 \times 10^{-5}$) and an underrepresentation of associated CpGs in inactivated regions (7 CpGs, 28.6-fold depletion, Fisher's $P = 2.2 \times 10^{-23}$) was observed.

Further supporting a link with variable XCI, we observed that X-chromosomal CpGs were associated with differential expression of the nearby genes (<250Kb, see Methods) *ALG13* and *PIN4* (see Methods, Supplementary Data 10) both known to variably escape XCI [Zhang et al., 2013]. While a strong eQTL effect and a clear biologically relevant link with XCI mainly implies *SMCHD1* in X-chromosomal hypomethylation (insertion of rs140837774), an eQTL effect for *METTL4*, although slightly weaker, leaves open a possible role for *METTL4* in XCI, given its role in the mRNA modification N^6-methyladenosine.

Using both female and male samples ($N = 3,265$, Supplementary Fig. 6) to investigate associations of genetic variation at the *SMCHD1/METTL4* locus with autosomal methylation in trans (>5Mb), we found that the *SMCHD1/METTL4* variant was associated with 20 CpGs mapping to the *HOXD10*, *HOXC10*, and *HOXC11* genes of the HOXD and HOXC clusters located on chromosomes 2 and 12, as well as to the large protocadherin beta (PCDHβ) and gamma (PCDHγ) clusters on chromosome 5 (FDR < 0.05, Supplementary Fig. 7, Supplementary

Data 11), all known *SMCHD1* targets [Gendrel et al., 2013; Mould et al., 2013].

The second of the three sentinel variants, sentinel SNP rs139916287 (MAF = 0.07), is located in intron 4 of the *HBG2* gene on chromosome 11, in the β-globin locus (rs139916287, chromosome 11, Supplementary Fig. 2B). The rare allele of the sentinel variant was associated with decreased expression of both the *HBG2* and *TRIM6* genes (T allele; Wald $P = 5.3 \times 10^{-7}$, regression coefficient = -130.55; Wald $P = 9.8 \times 10^{-5}$, regression coefficient = -0.05; Supplementary Data 8), based on *cis*-eQTL mapping testing genes up to 250kb up-, and downstream of the sentinel variant (Supplementary Data 8). While *HBG2* showed higher expression levels and a stronger eQTL effect in our data, *TRIM6* has been shown to bind *XIST* [Chu et al., 2018], and contributes to the maintenance of pluripotency in mouse embryonic stem cells [Sato et al., 2012], making *TRIM6* a strong candidate for explaining our observations. Associating the *TRIM6/HBG2* variant with X-chromosomal methylation, we found 276 associated X-chromosomal CpG sites (FDR < 0.05, Figure 4.3A, Supplementary Data 9). The rare allele (T allele) was associated with hypomethylation at 258 of those CpGs (93.5%, binomial $P = 6.3 \times 10^{-47}$), where mean effect size at these 258 CpGs is 1.6% per T allele, ranging from 0.15% to 4.25% (Supplementary Fig. 5).

Similar to the *SMCHD1/METTL4* variant, associated CpGs were overrepresented in CGIs (199 CpGs, 4.2-fold enrichment, Fisher's $P = 1.6 \times 10^{-29}$), and enriched in regions characterized by H3K27me3 (208 CpGs, 10.5-fold enrichment, Fisher's $P = 3.5 \times 10^{-77}$) and H3K4me3 (Kundaje et al. [2015], 217 CpGs, 6.4-fold enrichment, Fisher's $P = 5.5 \times 10^{-46}$, Figure 4.3B). The associated CpGs were again strongly overrepresented in genomic regions variably escaping XCI in an external set of whole blood samples (Cotton et al. [2014], see Methods, 39 CpGs, 8.8-fold enrichment, Fisher's $P = 2.1 \times 10^{-20}$), to a lesser extent present in regions consistently escaping XCI (61 CpGs, 6.8-fold enrichment, Fisher's $P = 2.2 \times 10^{-23}$), and underrepresented in repressed regions (39 CpGs, 15.2-fold depletion, Fisher's $P = 1.9 \times 10^{-50}$).

In addition, many genes known to variably escape XCI [Zhang et al., 2013], annotated to CpGs associated with *TRIM6/HBG2* genetic variation (*ALG13*, *ATP6AP2*, *CXorf38*, *MED14*, *SMC1A*, *TBL1X*, Supplementary Data 10). Similar to *SMCHD1/METTL4*, these results suggest a role for the *TRIM6/HBG2* locus in variable escape from XCI.

The sentinel SNP of the third identified locus (rs1736891, MAF = 0.38) was associated with the expression of several nearby genes annotated as zinc fingers (Supplementary Data 8), but most strongly with downregulation of the expression of the nearby transcription factor *ZSCAN9* gene [Vaquerizas et al., 2009] located on chromosome 6, based on *cis*-eQTL mapping in our own data (Wald $P = 2.5 \times 10^{-49}$, Supplementary Data 8). The sentinel SNP was significantly associated with 19 X-chromosomal CpGs in females (FDR < 0.05, Supplementary Fig. 8, Supplementary Data 9), all located in the same CpG island: the high-expression A allele of rs1736891 was associated with mild hypomethylation of all 19 CpGs (Fisher's $P = 3.6 \times 10^{-5}$, mean effect size 1.3% per allele, Supplementary Fig.

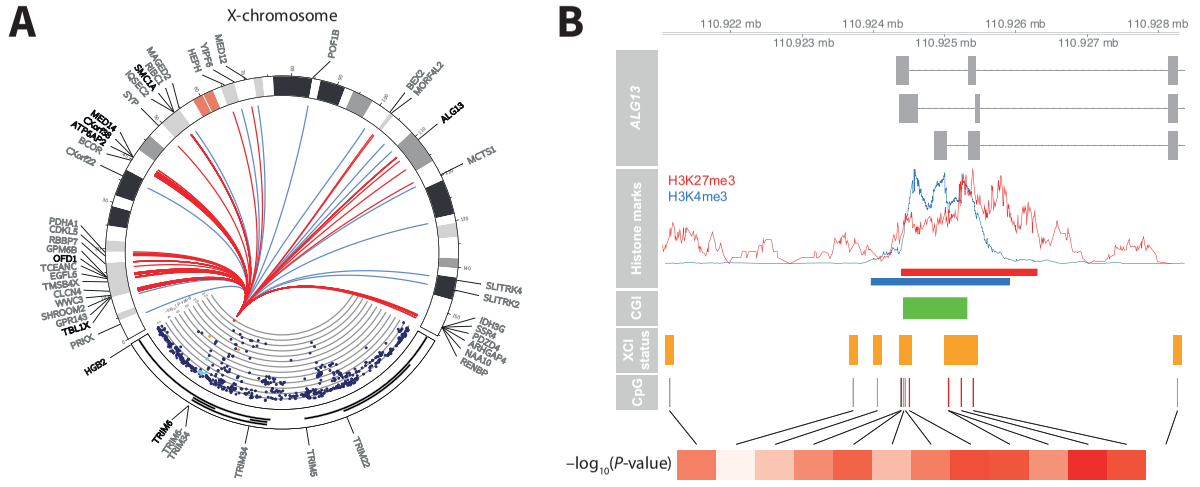

Figure 4.3: The *TRIM6/HBG2* locus is associated with DNA methylation at X-chromosomal regions. **(A)** Plot showing the *TRIM6/HBG2* locus and the widespread effects it has on the X-chromosome. The colors in the *TRIM6/HBG2* locus indicate LD (red: $R^2 \geq 0.8$; orange: $0.6 \leq R^2 < 0.8$; green: $0.4 \leq R^2 < 0.6$; light blue: $0.2 \leq R^2 < 0.4$; dark blue: $R^2 \leq 0.6 \leq 0.2$). The y-axis shows the $-\log_{10}(P\text{-value})$ of the association with overall X-chromosomal methylation. The line colors in the Circos plot indicate the direction of the effect (red: hypomethylation, blue: hypermethylation). The T allele of its sentinel variant rs139916287 is associated with upregulation of *HBG2*, downregulation of *TRIM6* (both shown in bold), and hypomethylation at 258 of the 276 associated CpGs (93.5%, red lines, mean effect size 1.6% per allele). X-chromosomal genes whose expression levels were associated with methylation levels of nearby CpGs are shown in bold. **(B)** Example of CpG island (CGI) in the *ALG13* gene associated with the *TRIM6/HBG2* locus. The enrichments of CpGs in certain genomic regions are similar to those found for the *SMCHD1/METTL4* locus. Most notably, the associated CpGs are also overrepresented in regions known to variably escape X-chromosome inactivation (fourth row, orange bars, 8.8-fold enrichment, Fisher's $P = 2.1 \times 10^{-20}$).

5). There was an overlap of in the CpGs associated with the sentinel variants of the *ZSCAN9* and the *SMCHD1/METTL4* locus, although the two loci are located on different chromosomes (chromosomes 6 and 18, respectively, Supplementary Fig. 2). These associations were statistically independent from each other (*i.e.*, additive), as all identified loci were identified using conditional analyses (see Methods). Specifically, 17 out of 19 CpGs (89.5%) were also targeted by the *SMCHD1/METTL4* locus, and all 19 CpGs show consistent opposite effects for both loci (Supplementary Fig. 9). Similar to the *SMCHD1/METTL4* locus, the *ZSCAN9* locus also associated with autosomal DNA methylation in trans (>5Mb). However, none of the autosomal CpGs overlapped between the two loci (Supplementary Data 11).

Discussion

Here, we identify three autosomal genetic loci with female-specific effects on X-chromosomal methylation in humans (*SMCHD1/METTL4*, *TRIM6/HBG2*, and *ZSCAN9*), all of which were associated with altered expression of autosomal genes *in cis*. Furthermore, all three loci were consistently associated with mild hypomethylation of CpGs overrepresented in CpG islands of X-chromosomal regions known to variably escape XCI in whole blood [Cotton et al., 2014; Zhang et al., 2013]. The former finding extended to 26 other tissues [Cotton et al., 2014], suggesting a cross-tissue genetic basis for variable escape from XCI. We observed a striking underrepresentation of affected CpGs in fully inactivated CGIs, which may be due to the tightly regulated nature of these regions. Methylation of these CpGs may be impervious to the impact of autosomal genetic variation or effects may be substantially weaker requiring much larger data sets to detect them.

While most of the previous work on XCI was done using mouse models and established a critical role for *SMCHD1* in XCI [Blewitt et al., 2008; Daxinger et al., 2013; Gendrel et al., 2013], we here confirm the role of the *SMCHD1/METTL4* locus in XCI in humans and highlight its impact on variable escape from XCI. This phenomenon has not been previously described in mice, perhaps due to the lack of genetic variability in the often inbred mice, leading to less (variable) escape from XCI than occurs in humans [Carrel and Willard, 2005; Cotton et al., 2013]. We also observed associations of the *SMCHD1/METTL4* locus with known autosomal *SMCHD1* targets [Gendrel et al., 2013; Mould et al., 2013], most notably the protocadherin clusters [Mason et al., 2017]. Interestingly, similar to the X-chromosome, the expression of the clustered protocadherin genes is stochastic and mono-allelic [Chess, 2005], suggesting a common mechanism.

In addition to the *SMCHD1/METTL4* locus, our results indicated a role for the *TRIM6/HBG2* locus in XCI. *TRIM6* is a strong candidate to influence female X-chromosome methylation because it was reported to bind *XIST* [Chu et al., 2018] and is involved in *MYC* and *NANOG* regulation [Sato et al., 2012]. Similarly, our data suggest a role for the *ZSCAN9* locus in variable escape from XCI, as it affects a single CpG island that is also targeted by the *SMCHD1/METTL4* locus. While this does suggest a role for the two loci in the same pathway, the effects on the

X-chromosome were statistically independent.

Given the biological consistency of the findings presented here, and the replication thereof in an independent set of samples, our data support a role of autosomal genetic variants in regulating Xi methylation in particular at variably escaping regions. However, to definitely demonstrate causality, unequivocally identify the responsible genes, and provide precise insight into the exact underlying mechanisms, *in vitro* experiments are needed. Importantly, a population genomics approach, like ours, will reveal effects on both XCI establishment and maintenance, which occur during different developmental stages and may involve different molecular pathways. At this point, the exact role of the *SMCHD1/METTL4*, *TRIM6/HBG2* and *ZSCAN9* loci during these processes remain to be determined. Therefore, it will be crucial to design experiments that can discriminate between an effect during the establishment and maintenance phases.

In conclusion, variable escape from XCI in humans has a genetic basis and we identified three autosomal loci, one previous implicated in XCI in mice and two new loci, that influence regions that are susceptible to variable escape from XCI by controlling X-chromosomal DNA methylation or correlated epigenetic marks.

Methods

Discovery cohorts

The Biobank-based Integrative Omics Study (BIOS) Consortium comprises six Dutch biobanks: Cohort on Diabetes and Atherosclerosis Maastricht (CODAM, van Greevenbroek et al. [2011]), LifeLines-DEEP (LLD, Tigchelaar et al. [2015]), Leiden Longevity Study (LLS, Schoenmaker et al. [2005]), Netherlands Twin Registry (NTR, Boomsma et al. [2002]), Rotterdam Study (RS, Hofman et al. [2013]), Prospective ALS Study Netherlands (PAN, Huisman et al. [2011]). The data that were analyzed in this study came from 3,265 unrelated individuals (Supplementary Data 1). Genotype data, DNA methylation data, and gene expression data were measured in whole blood for all samples. In addition, sex, age, measured cell counts (lymphocytes, neutrophils, monocytes, eosinophils, basophils, and red blood cell counts), and information on technical batches were obtained from the contributing cohorts. The Human Genotyping facility (HugeF, Erasmus MC, Rotterdam, The Netherlands, `http://www.glimdna.org`) generated the methylation and RNA-sequencing data and supplied information on technical batches.

Genotype data were generated within each cohort. Details on the genotyping and quality control methods have previously been detailed elsewhere (LLD: Tigchelaar et al. [2015]; LLS: Deelen et al. [2014a]; NTR: Lin et al. [2016]; RS: Hofman et al. [2013]; PAN: Huisman et al. [2011]).

For each cohort, the genotype data were harmonized towards the Genome of the Netherlands (GoNL, The Genome of the Netherlands Consortium et al.

[2014]) using Genotype Harmonizer [Deelen et al., 2014b] and subsequently imputed per cohort using Impute2 [Howie et al., 2009] and the GoNL reference panel (v5, The Genome of the Netherlands Consortium et al. [2014]). We removed SNPs with an imputation info-score below 0.5, a HWE P-value $< 10^{-4}$, a call rate below 95%, or a minor allele frequency smaller than 0.01. These imputation and filtering steps resulted in 7,545,443 SNPs that passed quality control in each of the datasets.

A detailed description regarding generation and processing of the gene expression data can be found elsewhere [Zhernakova et al., 2017]. Briefly, total RNA from whole blood was deprived of globin using Ambion's GLOBIN clear kit and subsequently processed for sequencing using Illumina's Truseq version 2 library preparation kit. Paired-end sequencing of 2x50bp was performed using Illumina's Hiseq2000, pooling 10 samples per lane. Finally, read sets per sample were generated using CASAVA, retaining only reads passing Illumina's Chastity Filter for further processing. Data were generated by the Human Genotyping facility (HugeF) of ErasmusMC (The Netherlands, see URLs). Initial QC was performed using FastQC (v0.10.1), removal of adaptors was performed using cutadapt (v1.1, Martin [2011]), and Sickle (v1.2, Joshi Fass, J. [2011]) was used to trim low quality ends of the reads (minimum length 25, minimum quality 20). The sequencing reads were mapped to human genome (HG19) using STAR (v2.3.0e, Dobin et al. [2013]).

To avoid reference mapping bias, all GoNL SNPs (`http://www.nlgenome.nl/?page_id=9`) with MAF > 0.01 in the reference genome were masked with N. Read pairs with at most 8 mismatches, mapping to as most 5 positions, were used.

Gene expression quantification was determined using base counts (for a detailed description, see Zhernakova et al. [2017]). The gene definitions used for quantification were based on Ensembl version 71. For data analysis, we used reads per kilobase per million mapped reads (RPKM), and only used protein coding genes with sufficient expression levels (median RPKM > 1), resulting in a set of 10,781 genes. To limit the influence of any outliers still present in the data, the data were transformed using a rank-based inverse normal transformation within each cohort.

The Zymo EZ DNA methylation kit (Zymo Research, Irvine, CA, USA) was used to bisulfite-convert 500 ng of genomic DNA, and 4 μl of bisulfite-converted DNA was measured on the Illumina HumanMethylation450 array using the manufacturer's protocol (Illumina, San Diego, CA, USA). Preprocessing and normalization of the data were done as described earlier [Tobi et al., 2015]. Removal of ambiguously mapped probes or probes containing known common genetic variants [Chen et al., 2013] were removed, followed by quality control (QC) using MethylAid's default settings [van Iterson et al., 2014], investigating methylated and unmethylated signal intensities, bisulfite conversion, hybridization, and detection P-values. Filtering of individual beta-values was based on detection P-value ($P < 0.01$), number of beads available (≤ 2) or zero values for signal intensity. Normalization was done using Functional Normalization [Fortin et al., 2014] as implemented

in the minfi R package [Aryee et al., 2014], using five principal components extracted using the control probes for normalization. All samples or probes with more than 5% of their values missing were removed, based on the QC performed using MethylAid. The final dataset consisted of 440,825 probes measured in 3,265 samples. Lastly, similar to the RNA-sequencing data, the methylation data were also transformed using a rank-based inverse normal transformation within each cohort, to limit the influence of any remaining outliers while removing any systematic differences in mean methylation between cohorts.

Replication cohorts

Accessible Resource for Integrative Epigenomics Studies (ARIES) Samples were drawn from the Avon Longitudinal Study of Parents and Children (ALSPAC, Fraser et al. [2013]; Boyd et al. [2013]). Blood from 1022 mother-child pairs (children at three time points and their mothers at two time points) were selected for analysis as part of Accessible Resource for Integrative Epigenomic Studies (ARIES, `http://www.ariesepigenomics.org.uk/`).

Written informed consent has been obtained for all ALSPAC participants. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Genotyping and methylation measurements have been previously described [Gaunt et al., 2016; Min et al., 2017].

We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

This work was supported by the UK Medical Research Council; Wellcome (`www.wellcome.ac.uk`; [grant number 102215/2/13/2 to ALSPAC]); the University of Bristol to ALSPAC; the UK Economic and Social Research Council (`www.esrc.ac.uk`; [ES/N000498/1] to CR) and the UK Medical Research Council (`www.mrc.ac.uk`; grant numbers [MC_UU_12013/1, MC_UU_12013/2 to JLM, CR]).

Exeter, Schizophrenia Phase 1 The University College London case-control sample has been described elsewhere [Datta et al., 2008; Hannon et al., 2016] but briefly comprises of unrelated ancestrally matched schizophrenia cases recruited from NHS mental health services and controls from the United Kingdom. Each control subject was interviewed to confirm that they did not have a personal history of an RDC defined mental disorder or a family history of schizophrenia, bipolar disorder, or alcohol dependence. UK National Health Service multicentre and local research ethics approval was obtained and all subjects signed an approved consent form after reading an information sheet. Details of DNA methylation and genetic data generation, processing, quality control and normalisation can be found in the original EWAS manuscript [Hannon et al., 2016].

Exeter, Schizophrenia Phase 2 The Aberdeen case-control sample has been described elsewhere [The International Schizophrenia Consortium, 2008; Hannon et al., 2016] but briefly contains schizophrenia cases and controls who have self-identified as born in the British Isles (95% in Scotland). Controls were volunteers recruited through general practices in Scotland. Practice lists were screened for potentially suitable volunteers by age and sex and by exclusion of subjects with major mental illness or use of neuroleptic medication. Volunteers who replied to a written invitation were interviewed using a short questionnaire to exclude major mental illness in individual themselves and first degree relatives. All cases and controls gave informed consent. The study was approved by both local and multiregional academic ethical committees. Details of DNA methylation and genetic data generation, processing, quality control and normalisation can be found in the original EWAS manuscript [Hannon et al., 2016].

Cooperative health research in the Region of Augsburg Study (KORA F4) The KORA study (Cooperative health research in the Region of Augsburg) consists of independent population-based samples from the general population living in the region of Augsburg, Southern Germany. Written informed consent has been given by each participant and the study was approved by the local ethical committee. The dataset comprised individuals from the KORA F4 survey (all with genotyping and methylation data available) conducted during 2006–2008. The KORA study was initiated and financed by the Helmholtz Zentrum München - German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research has been supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ.

Twins UK The TwinsUK cohort was established in 1992 to recruit monozygotic and dizygotic twins [Moayyeri et al., 2013]. More than 80% of participants are healthy female Caucasians (age range from 16 to 98 years old). The cohort includes more than 13,000 twin participants from all regions across the United Kingdom, and many have had multiple visits over the years. The TwinsUK cohort has been used in many epidemiological studies and is representative of the general UK population for a wide range of diseases and traits [Andrew et al., 2001].

TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union (EU), and the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility, and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

Rotterdam Study The Rotterdam Study (RS) is a large prospective, population-based cohort study aimed at assessing the occurrence of and risk factors for chronic (cardiovascular, endocrine, hepatic, neurological, ophthalmic, psychiatric, dermatological, oncological, and respiratory) diseases in the elderly [Ikram et al., 2017]. The study comprises 14,926 subjects in total, living in the well defined

Ommoord district in the city of Rotterdam in the Netherlands. In 1989, the first cohort, Rotterdam Study-I (RS-I) comprised of 7,983 subjects with age 55 years or above. In 2000, the second cohort, Rotterdam Study-II (RS-II) was included with 3,011 subjects who had reached an age of 55 or over in 2000. In 2006, the third cohort, Rotterdam Study-III (RS-III) was further included with 3,932 subjects with age 45 years and above. Each participant gave an informed consent and the study was approved by the medical ethics committee of the Erasmus University Medical Center, Rotterdam, the Netherlands.

The generation and management of the Illumina 450K methylation array data (EWAS data) for the Rotterdam Study was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. The EWAS data was funded by the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and by the Netherlands Organization for Scientific Research (NWO; project number 184021007) and made available as a Rainbow Project (RP3; BIOS) of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). We thank Mr. Michael Verbiest, Ms. Mila Jhamai, Ms. Sarah Higgins, Mr. Marijn Verkerk, and Lisette Stolk PhD for their help in creating the methylation database. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

Identifying female-specific genetic effects influencing X-chromosomal DNA methylation in the discovery cohorts

To identify autosomal genetic variants influencing DNA methylation anywhere on the X-chromosome we applied a two-step approach [Luijk et al., 2015] using 1,867 female samples from the replication cohorts for which both genotype data and methylation data were available. We first fitted linear models to test for an association between each autosomal SNP i and each of 10,286 X-chromosomal CpGs j individually, correcting for known covariates M (cell counts, cohort, age, technical batches - *e.g.*, sample plate and array position) and unknown confounding by including latent factors U, estimated using *cate* [Wang et al., 2015], where the eigenvalue difference method implemented in *cate* suggested an optimal number of three latent factors:

$$y_j = \beta_{ij} x_i + \gamma M + \delta U \tag{4.1}$$

For each autosomal genetic variant i, this approach yields 10,286 P-values p_{ij}. Next, we combined all 10,286 P-values corresponding to one genetic variant i into one overall P-value P_i using the Simes procedure [Simes, 1986], yielding 7,545,443 P-values P_i, one for each autosomal genetic variant tested.

This overall P-value per SNP indicates if an autosomal SNP influences DNA methylation *anywhere* on the X-chromosome, reducing this analysis to a GWAS for X-chromosomal DNA methylation. SNPs with an overall P-value $< 5 \times 10^{-8}$ were deemed significantly associated with X-chromosomal DNA methylation.

To identify independent effects among the identified variants, we performed iterative conditional analyses. We re-ran the entire above procedure, correcting for the strongest associated sentinel variant, as determined by the lowest overall P-value.

$$y_j = \beta_{ij}x_i + \gamma M + \delta U + \beta_{topSNP}x_{sentinel} \tag{4.2}$$

Having identified a new top SNP at the same genome-wide significance level of $P < 5x10^{-8}$, we again re-did our analysis, now correcting for two top SNPs. We repeated this process until no new independent effects were identified, which was after 47 such iterations, thus yielding 48 sentinel variants, corresponding to 48 different loci.

Next, to establish the female-specificity of the identified loci on X-chromosomal methylation, we aimed to validate the 48 identified loci in 1,398 males from the discovery cohorts for which the same genotype and methylation data were available. Any locus also having an effect in males would then mean that particular locus was not female specific. To do this, we tested the sentinel variant per locus found in females in the exact same way as we did in females, but also testing all SNPs within 1 Mb correlated to the sentinel variant ($R^2 \geq 0.8$ in males). A locus with any SNP having an overall P-value < 0.05 in males was not considered to be female-specific, yielding four loci with four corresponding sentinel variants.

Replication of sentinel variants associated with female-specific X-chromosomal DNA methylation in the replication cohorts

To replicate the four identified sentinel variants, we used an independent sample of 3,351 females from 5 different replication cohorts (see section Description of replication cohorts), all having genotype and 450k methylation data available. Similar to the discovery phase, each of four sentinel variants x_i was associated with all X-chromosomal CpGs y_j in each replication cohort k:

$$y_{jk} = \beta_{ijk}x_{ik} \tag{4.3}$$

each yielding a test-statistic t_{ijk}. We then combined the test-statistics corresponding to each genetic variant i and CpG j between each cohort k using Stouffer's weighted Z-method (discussed in Liptak [1958]), resulting in one overall Z-score Z_{ij} for each variant-CpG pair i, j:

$$Z_{ij} = \frac{\sum_k w_k t_{ijk}}{\sqrt{\sum_k w_k^2}} \tag{4.4}$$

where w_k indicates the sample size for replication cohort k. Converting each overall Z-score Z_{ij} to a P-value P_{ij}, we again used the Simes' procedure[Simes,

1986] to calculate one overall P-value P_i per genetic variant i, representing the statistical evidence for an association with any X-chromosomal CpG in the replication cohorts.

Local (*cis*) expression QTL mapping

In order to map the identified sentinel variants associated with female-specific X-chromosomal methylation to nearby genes, we employed *cis*-eQTL mapping in the discovery cohorts, where we associated the genotypes of a genetic variant with the expression levels of genes j *in cis* (< 250Kb). Similar to the *trans*-meQTL mapping for chromosome X, we corrected for known covariates M (*i.e.*, cell counts, cohort, age, technical batches), and unknown confounding U using *cate*, using an optimal number of latent factors to include, as suggested by *cate*:

$$y_j = \beta_{ij} x_i + \gamma M + \delta U \tag{4.5}$$

Next, we performed the Bonferroni correction to the corresponding P-values p_{ij} to identify genes associated with the genetic variant.

Associating X-chromosomal CpGs with changes in the expression of nearby genes

To identify genes associated with DNA methylation of nearby CpGs (< 250Kb), we used a similar model as for *trans*-meQTL and *cis*-eQTL mapping. We associated methylation levels of CpGs x_i with the observed expression values of a gene y_j using a linear model, correcting for covariates M (*i.e.*, cell counts, cohort, age, technical batches):

$$y_j = \beta_{ij} x_i + \gamma M \tag{4.6}$$

The Bonferroni correction was used to determine significant CpG-gene pairs.

Identifying genetic variants influencing autosomal DNA methylation

To identify long-range effects (> 5Mb) of a genetic variant on DNA methylation at autosomal CpGs, we performed *trans*-meQTL mapping using all 3,265 samples for which both genotype data and methylation data were available, as we expected these effects to be present in both women and men (Supplementary Fig. 6). For any genetic variant i and CpG j, we fitted a linear model correcting for known covariates M (cell counts, cohort, age, technical batches), and unknown confounding U using *cate*:

$$y_j = \beta_{ij} x_i + \gamma M + \delta U \tag{4.7}$$

The FDR was controlled within each set of corresponding P-values p_{ij}, to obtain a list of associated CpGs for a genetic variant i.

Testing epistatic effects

To test if the identified autosomal loci have any epistatic effects on X-chromosomal DNA methylation, we corrected the analysis for X-chromosomal *cis*-meQTLs. We first mapped *cis*-meQTLs (< 250Kb) on the X-chromosome by testing all nearby genetic variants for an effect on any of the X-chromosomal CpGs associated with one of the three autosomal loci. For any genetic variant i and CpG j, we fitted a linear model correcting for known covariates M (cell counts, cohort, age, technical batches):

$$y_j = \beta_{ij}^X x_i + \gamma M \tag{4.8}$$

We corrected for multiple testing using the Bonferroni procedure, selecting CpGs harboring *cis*-meQTLs. Next, we re-tested the effects of the autosomal loci on the X-chromosomal CpGs, but this time correcting for the strongest cis-SNP.

$$y_j = \beta_{ij}^X x_i^X + \beta_{ij}^{auto} x_i^{auto} + \gamma M \tag{4.9}$$

Annotations and enrichment tests

CpGs were annotation using UCSC Genome Browser [Kent et al., 2002], histone marks and chromatin states data from the Blueprint Epigenome data [Martens and Stunnenberg, 2013], transcription factor binding site (TFBS) data from the Encode Project [Consortium et al., 2012], and data on regions escaping X-inactivation [Cotton et al., 2014]. All annotations were done based on the location of the CpG site using HG19/GRCh37.

The CpG island (CGI) track from the UCSC Genome Browser was used to map CpGs to CGIs. Shores were defined as the flanking 2 kb regions. All other regions were defined as non-CGI.

We obtained Epigenomics Roadmap ChIP-seq data on histone marks measured in blood-related cell types (the GM12878 lymphoblastoid cell line, the K562 leukemia cell line, and monocytes). We selected five different histone marks for which data measured in both men and women were available (H3K4me3, H3K4me1, H3K9me3, H3K27me3, H3K27ac). A CpG was said to overlap with any histone mark if it did so in any the of the data sets.

We obtained Epigenomics Roadmap data on the 16 predicted core chromatin states data in blood-related cell types (the GM12878 lymphoblastoid cell line, the K562 leukemia cell line, and monocytes). A CpG was said to overlap with any chromatin state if it did so in any of the available data sets for that histone mark. Likewise, we obtained transcription factor binding data from the Encode Project, using blood-related cell types only (GM08714, GM10847, GM12878, GM12892, GM18505, GM18526, GM18951, GM19099, GM19193).

The degree of escape from X-inactivation for 632 transcription start sites (TSS) has previously been established in 27 different tissues [Cotton et al., 2014]. Within each tissue, each TSS was said to fully escape XCI, variably escape XCI, or be subject to XCI. We mapped each X-chromosomal CpG to the nearest such TSS, annotating each CpG with the accompanying scores for each of the 27

tissues. CpGs not in the vicinity of any such TSS (>10kb, 4,698 CpGs) were left unannotated.

In order to determine the enrichment of CpGs for any of the described genomic contexts, we used Fisher's exact test, where the used all X-chromosome CpGs as the background set.

Data availability

Raw data were submitted to the European Genome-phenome Archive (EGA) under accession EGAS00001001077 [https://www.ebi.ac.uk/ega/studies/EGAS00001001077].

References

Andrew, T. et al. [2001]. Are Twins and Singletons Comparable? A Study of Disease-related and Lifestyle Characteristics in Adult Women, *Twin Research* **4**(06): 464–477.

Aryee, M. J. et al. [2014]. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays, *Bioinformatics* **30**(10): 1363–1369.

Blewitt, M. E. et al. [2008]. SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation, *Nature Genetics* **40**(5): 663–669.

Boomsma, D. I. et al. [2002]. Netherlands Twin Register: A Focus on Longitudinal Research, *Twin Research* **5**(5): 401–406.

Boyd, A. et al. [2013]. Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children, *International Journal of Epidemiology* **42**(1): 111–127.

Breiling, A. and Lyko, F. [2015]. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond, *Epigenetics & Chromatin* **8**: 24.

Brinkman, A. B. et al. [2006]. Histone modification patterns associated with the human X chromosome, *EMBO Rep* **7**(6): 628–634.

Carrel, L. and Willard, H. F. [1999]. Heterogeneous gene expression from the inactive X chromosome: An X-linked gene that escapes X inactivation in some human cell lines but is inactivated in others, *Proceedings of the National Academy of Sciences* **96**(13): 7364–7369.

Carrel, L. and Willard, H. F. [2005]. X-inactivation profile reveals extensive variability in X-linked gene expression in females, *Nature* **434**(7031): 400–404.

Chen, K. et al. [2015]. Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation, *Proc Natl Acad Sci U S A* **112**(27): E3535–44.

Chen, Y. A. et al. [2013]. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray, *Epigenetics* **8**(2): 203–209.

Chess, A. [2005]. Monoallelic expression of protocadherin genes., *Nature Genetics* **37**(2): 120–121.

Chu, C. et al. [2018]. Systematic Discovery of Xist RNA Binding Proteins, *Cell* **161**(2): 404–416.

Consortium, Dunham, I. et al. [2012]. An integrated encyclopedia of DNA elements in the human genome, *Nature* **489**(7414): 57–74.

Cotton, A. M. et al. [2013]. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome, *Genome Biol* **14**(11): R122.

Cotton, A. M. et al. [2014]. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation, *Human Molecular Genetics* **24**(6): 1528–1539.

Datta, S. R. et al. [2008]. A threonine to isoleucine missense mutation in the pericentriolar material 1 gene is strongly associated with schizophrenia, *Molecular Psychiatry* **15**: 615.

Daxinger, L. et al. [2013]. An ENU mutagenesis screen identifies novel and known genes involved in epigenetic processes in the mouse, *Genome Biol* **14**(9): R96.

Deelen, J. et al. [2014a]. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age, *Human Molecular Genetics* **23**(16): 4420–4432.

Deelen, P. et al. [2014b]. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration, *BMC Research Notes* **7**(1): 901.

Dobin, A. et al. [2013]. STAR: ultrafast universal RNA-seq aligner, *Bioinformatics* **29**(1): 15–21.

Falckenhayn, C. et al. [2016]. Comprehensive DNA methylation analysis of the Aedes aegypti genome, *Scientific Reports* **6**: 36444.

Fortin, J. P. et al. [2014]. Functional normalization of 450k methylation array data improves replication in large cancer studies, *Genome Biol* **15**(12): 503.

Fraser, A. et al. [2013]. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort, *International Journal of Epidemiology* **42**(1): 97–110.

Galupa, R. and Heard, E. [2015]. X-chromosome inactivation: new insights into cis and trans regulation, *Current Opinion in Genetics & Development* **31**: 57–66.

Gaunt, T. R. et al. [2016]. Systematic identification of genetic influences on methylation across the human life course, *Genome Biol* **17**: 61.

Gendrel, A.-V. et al. [2012]. Smchd1-Dependent and -Independent Pathways Determine Developmental Dynamics of CpG Island Methylation on the Inactive X Chromosome, *Developmental Cell* **23**(2): 265–279.

Gendrel, A. et al. [2013]. Epigenetic Functions of Smchd1 Repress Gene Clusters on the Inactive X Chromosome and on Autosomes, *Molecular and Cellular Biology* **33**(16): 3150–3165.

Gilbert, W. V., Bell, T. A. and Schaening, C. [2016]. Messenger RNA modifications: Form, distribution, and function, *Science* **352**(6292): 1408 LP – 1412.

Hannon, E. et al. [2016]. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation, *Genome Biology* **17**(1): 176.

Heard, E. et al. [2001]. Methylation of Histone H3 at Lys-9 Is an Early Mark on the X Chromosome during X Inactivation, *Cell* **107**(6): 727–738.

Hofman, A. et al. [2013]. The Rotterdam Study: 2014 objectives and design update, *European Journal of Epidemiology* **28**(11): 889–926.

Howie, B. N., Donnelly, P. and Marchini, J. [2009]. A Flexible and Accurate Genotype Imputation Method for the

Next Generation of Genome-Wide Association Studies, *plos genetics* **5**(6).

Huisman, M. H. et al. [2011]. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology, *J Neurol Neurosurg Psychiatry* **82**(10): 1165–1170.

Ikram, M. A. et al. [2017]. The Rotterdam Study: 2018 update on objectives, design and main results, *European Journal of Epidemiology* **32**(9): 807–850.

Joshi Fass, J., N. [2011]. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33).

Kent, W. J. et al. [2002]. The Human Genome Browser at UCSC, *Genome Res* **12**(6): 996–1006.

Kundaje, A. et al. [2015]. Integrative analysis of 111 reference human epigenomes, *Nature* **518**(7539): 317–330.

Lemmers, R. J. et al. [2012]. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2, *Nature Genetics* **44**(12): 1370–1374.

Lin, B. D., Willemsen, G., Abdellaoui, A., Bartels, M., Ehli, E. A., Davies, G. E., Boomsma, D. I. and Hottenga, J. J. [2016]. The Genetic Overlap Between Hair and Eye Color, *Twin Research and Human Genetics* **19**(6): 595–599.

Liptak, T. [1958]. On the combination of independent tests, *Magyar Tud Akad Mat Kutato Int Kozl* **3**: 171–197.

Luijk, R. et al. [2015]. An alternative approach to multiple testing for methylation QTL mapping reduces the proportion of falsely identified CpGs, *Bioinformatics* **31**(3): 340–345.

Lyon, M. F. [1961]. Gene Action in the X-chromosome of the Mouse (Mus musculus L.), *Nature* **190**(4773): 372–373.

Martens, J. H. A. and Stunnenberg, H. G. [2013]. BLUEPRINT: mapping human blood cell epigenomes, *Haematologica* **98**(10): 1487–1489.

Martin, M. [2011]. Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal* **17**(1): 10.

Mason, A. G. et al. [2017]. SMCHD1 regulates a limited set of gene clusters on autosomal chromosomes, *Skeletal Muscle* **7**(1): 12.

Massah, S. et al. [2014]. Epigenetic characterization of the growth hormone gene identifies SmcHD1 as a regulator of autosomal gene clusters, *PLoS One* **9**(5): e97535.

Min, J. et al. [2017]. Meffil: efficient normalisation and analysis of very large DNA methylation samples, *Doi.Org* p. 125963.

Moayyeri, A. et al. [2013]. Cohort Profile: TwinsUK and Healthy Ageing Twin Study, *International Journal of Epidemiology* **42**(1): 76–85.

Mould, A. W. et al. [2013]. Smchd1 regulates a subset of autosomal genes subject to monoallelic expression in addition to being critical for X inactivation, *Epigenetics Chromatin* **6**(1): 19.

Nozawa, R. S. et al. [2013]. Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway, *Nat Struct Mol Biol* **20**(5): 566–573.

Orru, V. et al. [2013]. Genetic variants regulating immune cell levels in health and disease, *Cell* **155**(1): 242–256.

Peeters, S. B., Cotton, A. M. and Brown, C. J. [2014]. Variable escape from X-chromosome inactivation: identifying factors that tip the scales towards expression, *Bioessays* **36**(8): 746–756.

Plath, K. et al. [2003]. Role of histone H3 lysine 27 methylation in X inactivation, *Science* **300**(5616): 131–135.

Roederer, M. et al. [2015]. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis, *Cell* **161**(2): 387–403.

Sato, T., Okumura, F., Ariga, T. and Hatakeyama, S. [2012]. TRIM6 interacts with Myc and maintains the pluripotency of mouse embryonic stem cells, *J Cell Sci* **125**(Pt 6): 1544–1555.

Schoenmaker, M. et al. [2005]. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study, *European Journal of Human Genetics* .

Sharp, A. J., Stathaki, E., Migliavacca, E., Brahmachary, M., Montgomery, S. B., Dupre, Y. and Antonarakis, S. E. [2011]. DNA methylation profiles of human active and inactive X chromosomes, *Genome Res* **21**(10): 1592–1600.

Simes, R. J. [1986]. An Improved Bonferroni Procedure for Multiple Tests of Significance, *Biometrika* **73**(3): 751–754.

The Genome of the Netherlands Consortium et al. [2014]. Whole-genome sequence variation, population structure and demographic history of the Dutch population, *Nature Genetics* **46**(8): 818–825.

The International Schizophrenia Consortium [2008]. Rare chromosomal deletions and duplications increase risk of schizophrenia, *Nature* **455**(7210): 237–241.

Tigchelaar, E. F. et al. [2015]. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics, *BMJ Open* **5**(8): e006772.

Tobi, E. W., Slieker, R. C., Stein, A. D., Suchiman, H. E., Slagboom, P. E., van Zwet, E. W., Heijmans, B. T. and Lumey, L. H. [2015]. Early gestation as the critical time-window for changes in the prenatal environment to affect the adult human blood methylome, *Int J Epidemiol* **44**(4): 1211–1223.

van Greevenbroek, M. M. J. et al. [2011]. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and

general inflammation (the CODAM study), *European Journal of Clinical Investigation* **41**(4): 372–379.

van Iterson, M. et al. [2014]. MethylAid: visual and interactive quality control of large Illumina 450k datasets, *Bioinformatics* **30**(23): 3435–3437.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. and Luscombe, N. M. [2009]. A census of human transcription factors: function, expression and evolution, *Nat Rev Genet* **10**(4): 252–263.

Wang, J., Zhao, Q., Hastie, T. and Owen, A. B. [2015]. Confounder Adjustment in Multiple Hypothesis Testing, *ArXiv e-prints* .

Wu, T. P. et al. [2016]. DNA methylation on N6-adenine in mammalian embryonic stem cells, *Nature* **532**(7599): 329–333.

Yang, F., Babak, T., Shendure, J. and Disteche, C. M. [2010]. Global survey of escape from X inactivation by RNA-sequencing in mouse, *Genome Res* **20**(5): 614–622.

Yasukochi, Y. et al. [2010]. X chromosome-wide analyses of genomic DNA methylation states and gene expression in male and female neutrophils, *Proc Natl Acad Sci U S A* **107**(8): 3704–3709.

Zhang, Y. et al. [2013]. Genes that escape X-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving, *Mol Biol Evol* **30**(12): 2588–2601.

Zhernakova, D. V. et al. [2017]. Identification of context-dependent expression quantitative trait loci in whole blood, *Nature Genetics* **49**(1): 139–145.

5 GENOME-WIDE IDENTIFICATION OF DIRECTED GENE NETWORKS USING LARGE-SCALE POPULATION GENOMICS DATA

René Luijk, K.F. Dekkers, M. van Iterson, W. Arindrarto, A. Claringbould, P. Hop, BIOS Consortium, D.I. Boomsma, C.M. van Duijn, M.M. 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, E.W. van Zwet

Nature Communications, **9**(1) (2018)

Abstract

Identification of causal drivers behind regulatory gene networks is crucial in understanding gene function. Here, we develop a method for the large-scale inference of gene-gene interactions in observational population genomics data that are both directed (using local genetic instruments as causal anchors, akin to Mendelian Randomization) and specific (by controlling for linkage disequilibrium and pleiotropy). Analysis of genotype and whole-blood RNA-sequencing data from 3,072 individuals identified 49 genes as drivers of downstream transcriptional changes (Wald $P < 7 \times 10^{-10}$), among which transcription factors were overrepresented (Fisher's $P = 3.3 \times 10^{-7}$). Our analysis suggests new gene functions and targets, including for *SENP7* (zinc-finger genes involved in retroviral repression) and *BCL2A1* (target genes possibly involved in auditory dysfunction). Our work highlights the utility of population genomics data in deriving directed gene expression networks. A resource of *trans*-effects for all 6,600 genes with a genetic instrument can be explored individually using a web-based browser.

Introduction

Identification of the causal drivers underlying regulatory gene networks may yield new insights into gene function [Stuart et al., 2003; de la Fuente, 2010], possibly leading to the disentanglement of disease mechanisms characterized by transcriptional dysregulation [Lee and Young, 2013]. Gene networks are commonly based on the observed co-expression of genes. However, such networks show only undirected relationships between genes, which makes it impossible to pinpoint the causal drivers behind these associations. Adding to this, confounding (*e.g.* due to demographic and clinical characteristics, technical factors, and batch effects [Bruning et al., 2016; van Iterson et al., 2017]) induces spurious correlations between the expression of genes. Correcting for all confounders may prove difficult as some may be unknown [McGregor et al., 2016]. Residual confounding then leads to very large, inter-connected co-expression networks that do not reflect true biological relationships. To address these issues, we exploited recent developments in data analysis approaches that enable the inference of causal relationships through the assignment of directed gene-gene associations in population-based transcriptome data using genetic instruments [Gamazon et al., 2015; Gusev et al., 2016; Zhu et al., 2016] (GIs). Analogous to Mendelian Randomization (MR, Davey Smith and Hemani [2014]; Evans and Davey Smith [2015]), the use of genetics provides an anchor from where directed associations can be identified. Moreover, GIs are free from any non-genetic confounding. Related efforts have used similar methods to identify additional genes associated with different phenotypes, either using individual level data [Gamazon et al., 2015; Gusev et al., 2016] or using publicly available eQTL and GWAS catalogues [Zhu et al., 2016]. However, these efforts have not systematically taken linkage disequilibrium (LD) and pleiotropy (a genetic locus affecting multiple genes) into account. As both may lead to correlations between GIs, we aimed to improve upon these methods in order to minimize the influence of LD and pleiotropy, and would detect the actual driver genes. This possibly induces non-causal relations [Solovieff et al., 2013], precluding the identification of the specific causal gene involved when not accounted for LD and pleiotropy.

Here, we combine genotype and expression data of 3,072 unrelated individuals from whole blood samples to establish a resource of directed gene networks using genetic variation as an instrument. We use local genetic variation in the population to capture the portion of expression level variation explained by nearby genetic variants (local genetic component) of gene expression levels, successfully identifying a predictive genetic instrument (GI) for the observed gene expression of 6,600 protein-coding genes. These GIs are then tested for an association with potential target genes *in trans*. Applying a robust genome-wide approach that corrects for linkage disequilibrium and local pleiotropy by modelling nearby GIs as covariates, we identify 49 index genes each influencing up to 33 target genes (Bonferroni correction, Wald $P < 7 \times 10^{-10}$). Closer inspection of examples reveals that coherent biological processes underlie these associations, and we suggest new gene functions based on these newly identified target genes, *e.g.* for *SENP7* and *BCL2A1*. An interactive online browser allows researchers to

look-up specific genes of interest (see URLs).

Results

Establishing directed associations in transcriptome data

We aim to establish a resource of index genes that causally affect the expression of target genes *in trans* using large-scale observational RNA-sequencing data. However, causality cannot be inferred from the correlation between the observed expression measurements of genes, and therefore is traditionally addressed by experimental manipulation. Furthermore, both residual and unknown confounding can induce correlation between genes, possibly yielding to extensive correlation networks that are not driven by biology. To establish causal relations between genes, we assume a structural causal model [Pearl, 2009] describing the relations between genes and using their genetic components, the local genetic variants predicting their expression, as genetic instruments [Davey Smith and Hemani, 2014] (GIs). To be able to conclude the presence of a causal effect of the index gene on the target gene, the potential influence of linkage disequilibrium (LD) and pleiotropic effects have to be taken into account, as they may cause GIs of neighbouring genes to be correlated (Figure 5). This is done by blocking the so-called back-door path [Pearl, 2009] from the index GI through the genetic GIs of nearby genes to the target gene by correcting the association between the GI and target gene expression for these other GIs. Note that this path cannot be blocked by adjusting for the observed expression of the nearby genes, as this may introduce collider bias, resulting in spurious associations.

To assign directed relationships between the expression of genes and establish putative causality, the first step in our analysis approach was to identify a GI for the expression of each gene, reflecting the local genetic component. To this end, we used data on 3,072 individuals with available genotype and gene expression data (Supplementary Data 1), measured in whole blood, where we focused on at least moderately expressed (see Methods) protein-coding genes (10,781 genes, Supplementary Fig. 1). Using the 1,021 samples in the training set (see Methods), we obtained a GI consisting of at least 1 SNP for the expression of 8,976 genes by applying lasso regression to nearby genetic variants while controlling for known (cohort, sex, age, white and red blood cell counts) and unknown covariates [Wang et al., 2015] (see Methods). Adding distant genetic variants to the prediction model has been shown to add very little predictive power [Gamazon et al., 2015] and would have induced the risk of including long-range pleiotropic effects.

The strength of the GIs was evaluated using the 2,051 samples in the test set (see Methods). Taking LD and local pleiotropy into account by including the GIs of neighbouring genes ($<$ 1 Mb, Figure 5), we identified 6,600 sufficiently strong GIs having at least partly specific predictive ability (Supplementary Fig. 2a) for the expression its corresponding index gene (F-statistic > 10, Supplementary Fig. 1, Supplementary Data 2). To evaluate the effects of these

6,600 GIs on target gene expression, we tested for an association of each of 6,600 GIs with all of 10,781 expressed, protein-coding genes *in trans* (> 10Mb, Supplementary Fig. 2b). To have maximum statistical power we used all 3,072 samples, as opposed to only using the 2,051 samples from the test set, as the results from both analyses showed very similar results (Supplementary Fig. 3). First, this analysis was done without accounting for LD and local pleiotropy (*i.e.*, correcting for neighbouring LD, Figure 5). This genome-wide analysis resulted in 401 directed associations between 134 index genes and 276 target genes after adjustment for multiple testing using the Bonferroni correction (Wald $P < 7 \times 10^{-10}$, Figure 5.2, Supplementary Data 3). Among them were 134 index genes affecting the expression of 1 to 33 target genes *in trans* (3.2 genes on average, median of 1 gene), totalling 276 identified target genes. As expected, the resulting networks contained many instances where the same target gene was influenced by multiple neighbouring index genes, hindering the identification of the causal gene (65 such instances). Repeating the analysis for the 134 identified index genes, but corrected for LD and local pleiotropy by including the GIs of neighbouring genes (< 1Mb) resulted in the identification of specific directed effects for 49 index genes on 144 target genes, totalling 156 directed associations (Wald $P < 7 \times 10^{-10}$, Figure 5.2), where the number of target genes affected by an index gene varied from 1 to 33 (Supplementary Data 8, 3.2 genes on average, median of 1 gene). The number of target genes associated with multiple neighbouring index genes drops from 65 to 2, underscoring the importance of correction for LD and pleiotropy. As this set of 156 directed associations is free from LD and local pleiotropy, and possibly reflect truly causal relations, we use these in further analyses.

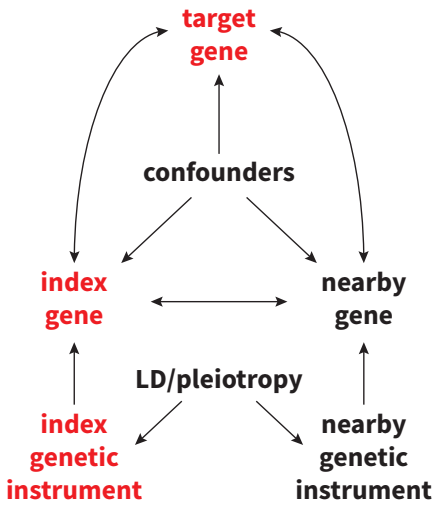

Figure 5.1: Diagram showing the presumed relations between each variable. A directed arrow indicates the possibility of a causal effect. For instance, the index genetic instrument represents nearby SNPs with a possible effect on the nearby gene (analogous to *cis*-eQTLs). A double arrow means the possibility of a causal effect in either direction. The index gene, for example, could have a causal effect on the target gene, or vice versa. We aim to assess the presence of a causal effect of the index gene on the target gene using genetic instruments (GIs) that are free of non-genetic confounding. To do this, we must block the back-door path from the index GI through the GIs of nearby genes to the target

gene. This back-door path represents
linkage disequilibrium and local pleiotropy and is precluded by correcting for the GIs of nearby genes. Correction for observed gene expression (either of the index gene or of nearby genes) does not block this back-door path, but instead possibly leads to a collider bias, falsely introducing a correlation between the index GI and the target gene.

Validity and stability of the analyses

To ensure the validity and stability of the analyses, we compared out methodology to earlier work and performed several checks regarding common challenges inherent to these analyses and the assumptions underlying them. First, we compared our approach to previously described approaches [Gamazon et al., 2015; Gusev et al., 2016] by applying these to our data. Each approach consists of a method to create GIs, and a model used to test for *trans*-effects. First, we used all methods to create GIs (lasso, elastic net, BLUP, and BSLMM), and investigated their predictive power of the index gene (see Methods). The methods that used feature selection (our method, lasso, and elastic net) showed similar predictive ability. Less predictive power was observed for methods using all nearby genetic variants (BLUP, BSLMM, Supplementary Fig. 4). Identifying *trans*-effects showed a lower number of *trans*-effects identified for the BLUP and BSLMM methods (Supplementary Fig. 5), possibly as a result of their less predictive GIs (Supplementary Fig. 4). In addition, as this *trans*-model does not take LD into account, a large number of target genes are associated with the GIs of many neighboring index genes (Supplementary Fig. 5).

To investigate how well our proposed *trans*-model is able to control for LD, and to evaluate the statistical power of this model, we performed a simulation study investigating several scenarios (Supplementary Fig. 6, see Methods for details on the simulation of the data). Overall, the simulations show high power to detect a true causal effect of the GI of the index gene on the target gene, where the correlation between GI and index gene, and between index gene and target gene contribute most to an increased power. The presence of correlated GIs of nearby genes plays a smaller role. Under the null hypothesis (*i.e.*, when a neighboring gene influences the target gene, and not the tested index gene, see blue and purple lines), the uncorrected analysis will indeed lead to false positives (indicated by higher power), while the corrected analysis will indeed lead to false positives in 5% of the tests performed, indicating LD is indeed corrected for. The simulation confirms that our approach is more specific in identifying the causal gene than its competitors.

By design, the GIs should be independent of most confounding factors, but confounding may still occur if genetic variants directly affect blood composition, leading to spurious associations. While we have already explicitly corrected for known white and red blood cell counts, we also evaluated the association of the 49 GIs with these cell counts, and found that none of the 49 GIs were significantly related to any observed cell counts (Supplementary Fig. 7a). In addition, all 156 directed associations remained significant after further adjustment for nearby

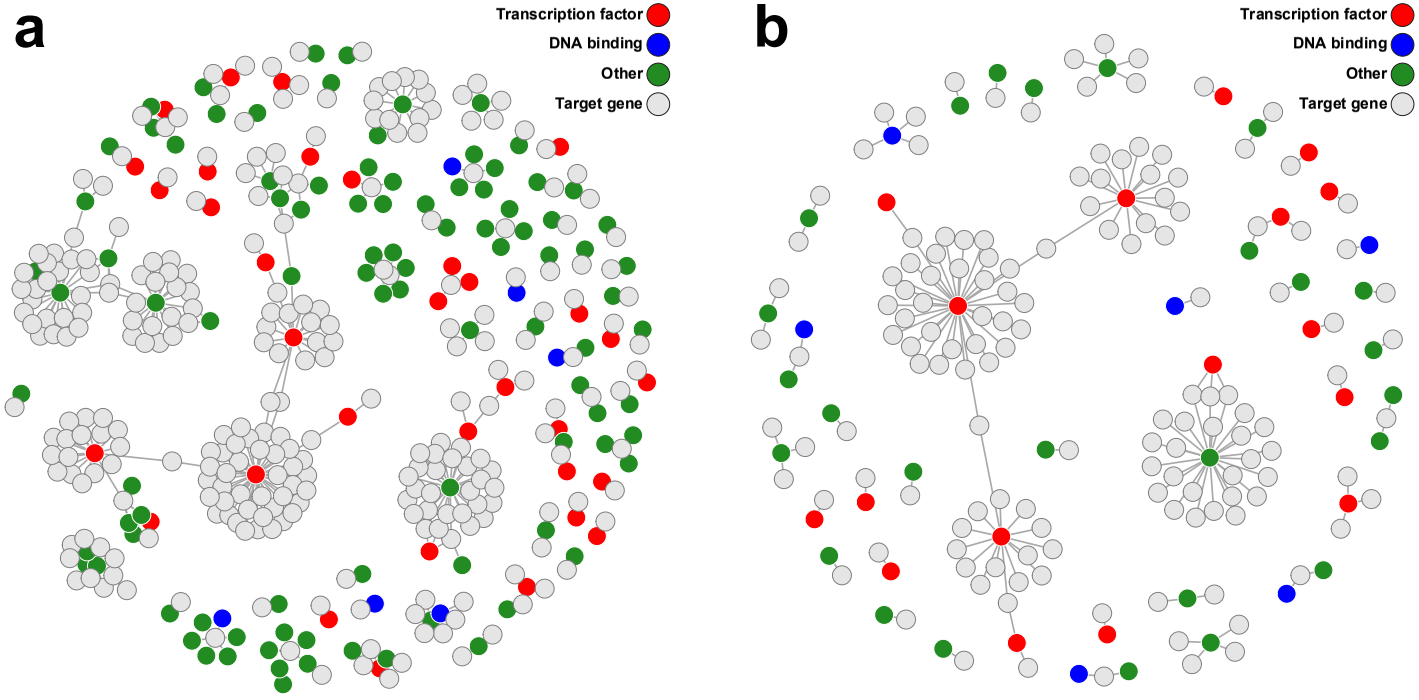

Figure 5.2: Gene networks showing the directed gene–gene association between genes. Panels show the associations when not taking LD and local pleiotropy into account **(a)** and when these are corrected for **(b)**. Index genes identified as a transcription factor are indicated by red circles. Blue circles indicate index genes with DNA binding properties, but are not a known transcription factor. Green circles indicate other index genes. Light grey circles indicate target genes. The uncorrected analysis shows 134 index genes (coloured circles) influencing 276 target genes, where several neighbouring index genes seemingly influencing the same target gene, which is reflective of a shared genetic component of those index genes. Specifically, 65 target genes are associated with multiple index genes which lie in close proximity to one another. The number of index genes drop sharply from 134 to 49 (2.7-fold decrease) when do taking LD and local pleiotropy into account. The number of target genes also drops, from 276 to 144 (1.9-fold decrease).

genetic variants (< 1Mb) reported to influence blood composition [Orru et al., 2013; Roederer et al., 2015] (Supplementary Fig. 7b).

To combat any unknown residual confounding and possibly gain statistical power, we added five latent factors to our models, estimated from the observed expression data using *cate* [Wang et al., 2015] (see Methods). We re-tested the 156 identified associations without these factors to evaluate the model sensitivity, showing similar results with slightly attenuated test statistics (Supplementary Fig. 7c). This indicates that our analysis was not influenced by unknown confounding and confirmed the independence of GIs from non-genetic confounding, but did help in reducing the noise in the data, leading to increased statistical power.

Next, to validate the GIs of the 49 index genes, we compared the SNPs constituting the GIs of the 49 index genes associated with target gene expression with previous *cis*-eQTL mapping efforts. While similar sets of genes may be identified using a *cis*-eQTL approach, the utility of using multi-SNP GIs over single-SNP GIs (akin to *cis*-eQTLs) is shown in the increased predictive ability of multi-SNP GIs (Supplementary Fig. 7d), and thus in the number of predictive GIs. Only 4,910 single-SNP GIs were predictive of its corresponding index gene (F-statistic > 10), compared to 6,600 multi-SNP instrumental variables. Of the 49 index genes corresponding to the 49 GIs, 47 genes (96.1%) were previously identified as harbouring a *cis*-eQTL in large subset of the whole blood transcriptome data we analysed here (2,116 overlapping samples), using an independent analysis strategy [Zhernakova et al., 2017]. Almost all of the corresponding GIs (98%, 46 GIs) were strongly correlated with the corresponding eQTL SNPs ($R^2 > 0.8$). Similarly, 26 of the 49 index genes (53%) were also reported as having a *cis*-eQTL effect in a much smaller set of whole blood samples ($N_{GTEx} = 338$) part of GTEx [GTEx Consortium, 2017], 23 of which also correlated strongly with the corresponding eQTL-SNPs ($R^2 > 0.8$). When considering all tissues in the GTEx project, we found 48 of 49 index genes were identified as harbouring a *cis*-eQTL in any of the 44 tissues measured.

Afterwards, we compared our identified effects with *trans*-eQTLs identified earlier in whole-blood samples [Joehanes et al., 2017]. First, we found 97 target genes identified here (67%) overlapped with those found by Joehanes et al., 19 of which had their corresponding GI and lead SNP in close proximity (< 1Mb, Supplementary Fig. 8), suggesting that the effects are indeed mediated by the index gene assigned using our approach. Testing for a *cis*-eQTL of those SNPs identified by Joehanes et al. on the nearby index genes, we found all 19 index genes indeed had at least one nearby lead SNP that influenced its expression (Wald $P < 6 \times 10^{-4}$, Supplementary Data 4).

As a last check, we investigated potential mediation effects of each of the 49 GIs by observed index gene expression using the Sobel test [Sobel, 1982] (Figure 5). This method is based on the notion that the effect of a GI on target gene expression should diminish when correcting for the mediator observed index gene expression. However, small effect sizes and considerable noise in both mediator and outcome lead to low statistical power to detect mediated effects [Fritz and MacKinnon, 2007]. Nevertheless, we found 105 of 156 significant directed associations (67%) to show evidence for mediation (Bonferroni correction for 156

tests: Wald $P < 3.1 \times 10^{-4}$; Supplementary Data 5).

Exploration of directed networks

To gain insight in the molecular function of the 49 index genes affecting target gene expression, we used Gene Ontology (GO) to annotate our findings. The set of 49 index genes was overrepresented in the GO terms DNA Binding (Fisher's $P = 5 \times 10^{-8}$) and Nucleic Acid Binding (Fisher's $P = 2.8 \times 10^{-5}$, Supplementary Data 6), with 43.8% (21 genes) and 47.9% (23 genes) of genes overlapping with those gene sets, respectively. In line with this finding, we found a significant overrepresentation of transcription factors (17 genes; odds ratio = 5.7, Fisher's $P = 3.3 \times 10^{-7}$) using a manually curated database of transcription factors [Vaquerizas et al., 2009]. We note that such enrichments are expected a priori and hence indirectly validate our approach. Of interest, several target genes of two transcription factors overlapped with those identified in previous studies [Jiang et al., 2007] (*IKZF1*: 27% of its target genes, 4 genes; *PLAGL1*: 15% of its target genes, 5 genes).

Finally, we explore the biological processes that are revealed by our analysis for several index genes that either are known transcription factors [Vaquerizas et al., 2009] or affect many genes *in trans*. While these results are limited to Bonferroni-significant directed associations (Wald $P < 7 \times 10^{-10}$, correcting for all possible combinations of the 6,600 index genes and 10,781 target genes), researchers can interactively explore the whole resource using a dedicated browser (see URLs).

We identified 25 target genes to be affected *in trans* by sentrin/small ubiquitin-like modifier (SUMO)-specific proteases 7 (*SENP7*, Figure 5.3, Figure 5.4, Supplementary Data 8), significantly expanding on the five previously suspected target genes resulting from an earlier expression QTL approach [Lemire et al., 2015]. Increased *SENP7* expression resulted in the upregulation of all but one of the target genes (96%). Remarkably, 23 of the 25 target genes affected by *SENP7* are zinc finger protein (ZFP) genes located on chromosome 19. More specifically, 18 target genes are located in a 1.5Mb ZFP cluster mapping to 19q13.43 (Figure 5.3). ZFPs in this cluster are known transcriptional repressors, particularly involved in the repression of endogenous retroviruses [Lukic et al., 2014]. Parallel to this, *SENP7* has also been identified to promote chromatin relaxation for homologous recombination DNA repair, specifically through interaction with chromatin repressive KRAB-Association Protein (*KAP1*, also known as *TRIM28*). *KAP1* had already been implicated in transcriptional repression, especially in epigenetic repression and retroviral silencing [Fasching et al., 2015], although *KAP1* had no predictive GI (F-statistic $= 4.9$). Therefore, it has been speculated *SENP7* may also play a role in retroviral silencing [Garvin et al., 2013]. Given the widespread effects of *SENP7* on the transcription of chromosome 19-linked ZFPs involved in retroviral repression [Lukic et al., 2014], it corroborates a role of *SENP7* in the repression of retroviruses, specifically through regulation of this ZFP cluster. *SENP7* is not a TF and does not bind DNA, but considering it is a SUMOylation enzyme, it possibly has its effect on the ZFP

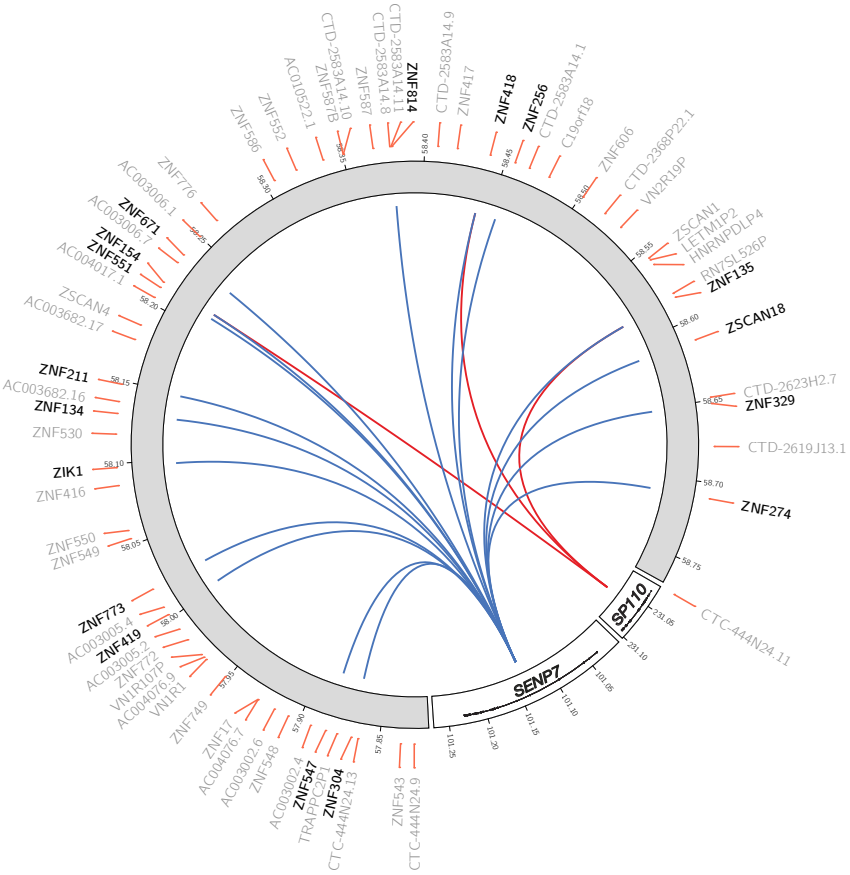

Figure 5.3: *SENP7* (chromosome 3) and *SP110* (chromosome 2) affect a zinc finger cluster located on chromosome 19. Many of these genes are involved in retroviral repression, among others. Blue lines indicate a positive association (upregulation), red lines indicate a negative association (downregulation). Colouring indicates consistent opposite effects of *SENP7* and *SP110* on their shared target genes.

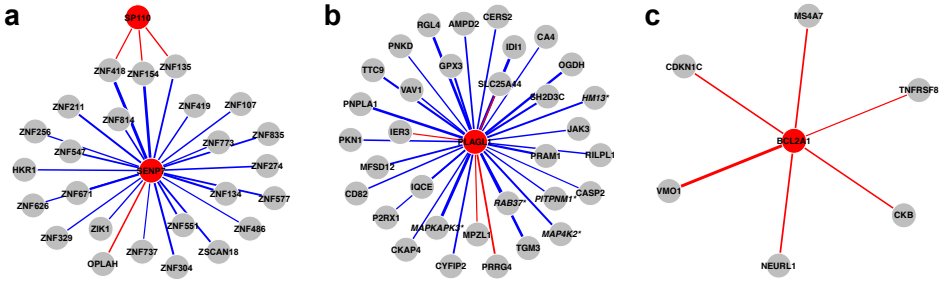

Figure 5.4: Identified target genes for different index genes. Panels show target genes for *SENP7* and *SP110* **(a)**, *PLAGL1* **(b)**, and *BCL2A1* **(c)**. Starred and italic gene names indicate previously reported target genes (Supplementary Data 7). Blue and red lines indicate positive and negative associations, respectively; line thickness indicates strength of the association.

cluster through deSUMOylation of *KAP1* (Li et al., 2007).

In our genome-wide analysis, we found that the transcription factor SP110 nuclear body protein (*SP110*) influences three zinc finger proteins (Figure 5.3, Figure 5.4). During viral infections in humans, *SP110* has been shown to interact with the Remodelling and Spacing Factor 1 (*RSF1*) and Activating Transcription Factor 7 Interacting Protein (*ATF7IP*), suggesting it is involved in chromatin remodelling *Cai2011*. Interestingly, all three of the genes targeted by *SP110* are also independently influenced by *SENP7*, although *SP110* shows opposite effects (Supplementary Fig. 9), and are located in the same ZFP gene cluster on chromosome 19. This overlap of target genes supports the previous suggestion that *SP110* is involved in the innate antiviral response [Lee et al., 2013], presumably through regulation of the same ZPF cluster regulated by *SENP7*.

The index gene with the most identified target gene effects *in trans* is Pleiomorphic Adenoma Gene-Like 1 (*PLAGL1*, also known as *LOT1*, *ZAC*). *PLAGL1* is a transcription factor and affected 33 genes, 29 of which are positively associated with *PLAGL1* expression (88%, Figure 5.4). *PLAGL1* is part of the imprinted *HYMAI/ZAC1* locus, which has a crucial role in fetal development and metabolism [Varrault et al., 2006]. This locus, and overexpression of *PLAGL1* specifically, has been associated with transient neonatal diabetes mellitus [Cai et al., 2011] (TNDM) possibly by reducing insulin secretion [Hoffmann and Spengler, 2012]. *PLAGL1* is known to be a transcriptional regulator of PACAP-type I receptor [Ciani et al., 1999] (*PAC1-R*). *PACAP*, in turn, is a regulator of insulin secretion [Yada et al., 1998]. In line with these findings, we found several target genes to be involved in metabolic processes. Most notably, we identified *MAPKAPK3* (MK3) and *MAP4K2* to be upregulated by *PLAGL1*, previously identified as *PLAGL1* targets [Marbach et al., 2016], and both part of the mitogen-activated protein kinase (MAPK) pathway. This pathway has been observed to be upregulated in type II diabetic patients (reviewed in Frojdo et al. [2009]). In addition, inhibition of *MAPKAP2* and *MAPKAP3* in obese,

insulin-resistant mice has been shown to result in improved metabolism [Ozcan et al., 2015], in line with the association between upregulation of *PLAGL1* and the development of TNDM. Furthermore, *PLAGL1* may be implicated in lipid metabolism and obesity through its effect on *IDI1*, *PNPLA1*, *JAK3*, and *RAB37* expression [Chang et al., 2013; Mishra et al., 2015; Vock et al., 2008; Xu et al., 2013]. While not previously established as target genes, they are in line with the proposed role of *PLAGL1* in metabolism [Valente et al., 2005].

Increased expression of Bcl-related protein A1 (*BCL2A1*) downregulated all five identified target genes (Figure 5.4). *BCL2A1* encodes a protein part of the B-cell lymphoma 2 (*BCL2*) family, an important family of apoptosis regulators. It has been implicated in the development of cancer, possibly through the inhibition of apoptosis (reviewed in Vogler [2012]). One target gene, *NEURL1*, is known to cause apoptosis [Teider et al., 2010], in line with its strong negative association with *BCL2A1* expression. Similarly, *CDKN1C* was also downregulated by *BCL2A1*, and implicated in the promotion of cell death [Vlachos et al., 2007; Yan et al., 1997]. However, little is known about the strongest associated target gene, *VMO1* (Wald $P = 1.5 \times 10^{-8}$). It has been implicated in hearing, due to its highly abundant expression in the mouse inner ear [Peters et al., 2007], where *BCL2A1* may have a role in the development of hearing loss through apoptosis, since cell death is a known contributor to hearing loss in mice [Tadros et al., 2008]. In line with its role in the inhibition of apoptosis, *BCL2A1* overexpression has a protective effect on inner ear mechanosensory hair cell death in mice [Cunningham et al., 2004]. Lastly, the target gene *CKB* has also been implicated in hearing impairment in mice [Shin et al., 2007] and Huntington's disease [Lin et al., 2011], further suggesting a role of *BCL2A1* in auditory dysfunction.

Discussion

In this work, we report on an approach that uses population genomics data to generate a resource of directed gene networks. Our genome-wide analysis of whole-blood transcriptomes yields strong evidence for 49 index genes to specifically affect the expression of up to 33 target genes *in trans*. We suggest previously unknown functions of several index genes based on the identification of new target genes. Researchers can fully exploit the utility of the resource to look up *trans*-effects of a gene of interest using an interactive gene network browser (see URLs).

The identified directed associations provide improved mechanistic insight into gene function. Many of the 49 index genes affecting target gene expression are established transcription factors (TFs), or are known for having DNA binding properties, an anticipated observation supporting the validity of our analysis. The identification of non-TFs will in part relate to the fact that the effect of an index gene may regulate the activity of TFs, for example by post-translational modification. This is illustrated by *SENP7* that we observed to concertedly affect the expression of zinc finger protein genes involved in the repression of retroviruses, likely by deSUMOylation of the transcription factor *KAP1* [Li et al.,

2007]. Other mechanistic insights that can be distilled from these results include the potential involvement of *BCL2A1* in auditory dysfunction, conceivably through the regulation of apoptosis.

While observational gene expression data can be used to construct gene co-expression networks [Lin et al., 2011], which is sometimes complemented with additional experimental information [Marbach et al., 2016], such an approach lacks the ability to assign causal directions. Experimental approaches using CRISPR-cas9 coupled with single-cell technology [Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016] are in principle able to demonstrate putative causality at a large scale, but only *in vitro*, while the advantage of observational data is that it reflects *in vivo* situations. These experimental approaches currently rely on extensive processing of single-cell data that is associated with high technical variability [Adamson et al., 2016], complicating the construction of specific gene-gene associations. In addition, off-target effects of CRISPR-cas9 cannot be excluded [Schaefer et al., 2017], potentially influencing the interpretation of these experiments. Finally, such efforts are currently limited in the number of genes tested [Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016], whereas we were able to perform a genome-wide analysis. Hence, experimental and population genomics approaches are complementary in identifying causal gene networks.

Traditional *trans*-eQTL studies aim to find specific genetic loci associated with distal changes in gene expression [Joehanes et al., 2017; Westra et al., 2013]. The limitation of this approach is that they are not designed to assign the specific causal gene responsible for the *trans*-effect because they do not control for LD and local pleiotropy (a genetic locus affecting multiple nearby genes). Hence, our approach enriches *trans*-eQTL approaches by specifying which index gene induces changes in target gene expression. However, it does not detect *trans*-effects independent of effects on local gene expression. In addition, identification of the causal path using a *trans*-eQTL approach is difficult to establish. Testing for mediation through local changes in expression [Pierce et al., 2014; Yang et al., 2017] may be limited in statistical power, as these approaches are designed to only test the mediation effect of one lead SNP [Pierce et al., 2014]. In addition, they too do not correct for pleiotropy or LD, possibly leading to several identified *cis*-genes mediating a *trans*-eQTL.

Related analysis methods were recently used to infer associations between gene expression and phenotypic outcomes (instead of gene expression as we did here). Two studies first constructed multi-marker GIs in relatively small sample sets to then apply these GIs in large datasets without gene expression data [Gamazon et al., 2015; Gusev et al., 2016]. A different, summary-data-based Mendelian randomization (SMR) approach identifies genes associated with complex traits based on publicly available GWAS and eQTL catalogues [Zhu et al., 2016]. However, neither of these approaches take LD or pleiotropic effects into account, leading to many neighbouring, non-specific effects [Gamazon et al., 2015; Gusev et al., 2016; Zhu et al., 2016]. We show that correcting for LD and local pleiotropy will aid in the identification of the causal gene, as opposed to the identification of multiple, neighbouring genes, analogous to fine mapping

in GWAS. Furthermore, the use of eQTL and GWAS catalogues are usually the result of genome-wide analyses, where only statistically significant variants are taken into account. Here, we use the full genetic landscape surrounding a gene, thereby maximizing the predictive ability of expression measurements by our GIs [Gamazon et al., 2015]. While we have used our genome-wide approach to identify directed gene networks, we note this method may also be used to annotate trait-associated variants with potential target genes, either by using individual level data [Gamazon et al., 2015; Gusev et al., 2016], or by using SMR [Zhu et al., 2016].

The analysis approach presented here relies on using GIs of expression of an index gene as a causal anchor, an approach akin to Mendelian randomization [Davey Smith and Hemani, 2014]. While GIs could provide directionality to bi-directional associations in observational data, genetic variation generally explains a relatively small proportion of the variation in expression (Supplementary Fig. 2a). The GIs for index gene expression identified here are no exception, significantly limiting statistical power of similar approaches [Brion et al., 2013; Freeman et al., 2013]. Increased sample sizes and improvement on the prediction of index gene expression will help in identifying more target genes.

Our current analysis strategy aims for causal inference, obviating LD and pleiotropic effect by correcting for the GIs of nearby genes. However, we only corrected for GIs of genes within 1 Mb of the current index gene, leaving the possibility of pleiotropic effects beyond this threshold. For example, the GI of an index gene may influence both the expression of the index gene and another gene, located outside of the 1 Mb window, where the induced changes in that genes' expression are the causal factor of the identified target genes. A related problem arises when a shared genetic component between neighbouring index genes causes all of them to associate with a single distant target gene, hindering the identification of the index gene responsible for the induced *trans*-effect. By correcting for the GI of nearby genes, these potentially biologically relevant effects are lost (Figure 5).

In conclusion, we present a genome-wide approach that identifies causal effects of gene expression on distal transcriptional activity in population genomics data and showcase several examples providing new biological insights. The resulting resource is available as an interactive network browser that can be utilized by researchers for look-ups of specific genes of interest (see URLs).

Methods

Cohorts

The Biobank-based Integrative Omics Study (BIOS, Additional SI1) Consortium comprises six Dutch biobanks: Cohort on Diabetes and Atherosclerosis Maastricht [van Greevenbroek et al., 2011] (CODAM), LifeLines-DEEP [Tigchelaar et al., 2015] (LLD), Leiden Longevity Study [Schoenmaker et al., 2005] (LLS), Netherlands Twin Registry [Boomsma et al., 2002] (NTR), Rotterdam Study

[Hofman et al., 2013] (RS), Prospective ALS Study Netherlands [Huisman et al., 2011] (PAN). The data that were analysed in this study came from 3,072 unrelated individuals (Supplementary Data 1). Genotype data and gene expression data were measured in whole blood for all samples. In addition, sex, age, and cell counts were obtained from the contributing cohorts. The Human Genotyping facility (HugeF, Erasmus MC, Rotterdam, The Netherlands, `http://www.blimdna.org`) generated the RNA-sequencing data.

Genotype data

Genotype data were generated within each cohort (LLD: Tigchelaar et al. [2015]; LLS: Deelen et al. [2014a]; NTR: Lin et al. [2016]; RS: Hofman et al. [2013]; PAN: Huisman et al. [2011]).

For each cohort, the genotype data were harmonized towards the Genome of the Netherlands [The Genome of the Netherlands Consortium et al., 2014] (GoNL) using Genotype Harmonizer [Deelen et al., 2014b] and subsequently imputed per cohort using Impute2 [Howie et al., 2009] and the GoNL reference panel [The Genome of the Netherlands Consortium et al., 2014] (v5). We removed SNPs with an imputation info-score below 0.5, a HWE $P < 10^{-4}$, a call rate below 95%, or a minor allele frequency smaller than 0.01. These imputation and filtering steps resulted in 7,545,443 SNPs that passed quality control in each of the datasets.

Gene expression data

Total RNA from whole blood was deprived of globin using Ambion's GLOBIN clear kit and subsequently processed for sequencing using Illumina's Truseq version 2 library preparation kit. Paired-end sequencing of 2x50bp was performed using Illumina's Hiseq2000, pooling 10 samples per lane. Finally, read sets per sample were generated using CASAVA, retaining only reads passing Illumina's Chastity Filter for further processing. Data were generated by the Human Genotyping facility (HugeF) of ErasmusMC (The Netherlands, see URLs). Initial QC was performed using FastQC (v0.10.1), removal of adaptors was performed using *cutadapt* [Martin, 2011] (v1.1), and *Sickle* [Joshi Fass, J., 2011] (v1.2) was used to trim low quality ends of the reads (minimum length 25, minimum quality 20). The sequencing reads were mapped to human genome (HG19) using *STAR* [Dobin et al., 2013] (v2.3.0e).

To avoid reference mapping bias, all GoNL SNPs (`http://www.nlgenome.nl/?page_id=9`) with MAF > 0.01 in the reference genome were masked with N. Read pairs with at most 8 mismatches, mapping to as most 5 positions, were used.

Gene expression quantification was determined using base counts [Zhernakova et al., 2017]. The gene definitions used for quantification were based on Ensembl version 71, with the extension that regions with overlapping exons were treated as separate genes and reads mapping within these overlapping parts did not count towards expression of the normal genes.

For data analysis, we used counts per million (CPM), and only used protein coding genes with sufficient expression levels (median $\log(\text{CPM}) > 0$), resulting in a set of 10,781 genes. To limit the influence of any outliers still present in the data, the data were transformed using a rank-based inverse normal transformation within each cohort.

Constructing a local genetic instrument for gene expression

We started by constructing genetic instruments (GIs) for the expression of each gene in our data. We first split up the genotype and RNA-sequencing data in a training set (one-third of all samples, $N_{train} = 1,021$) and a test set (two-thirds of all samples, $N_{test} = 2,051$), making sure all cohorts and both sexes were evenly distributed over the train and test sets (57% female), as well as an even distribution of age (mean = 56, sd = 14.8). Using the training set only, we built a GI for each gene j separately that best predicts its expression levels using lasso, using nearby genetic variants only (either within the gene or within 100kb of a gene's TSS or TES), while correcting for both known (cohort, sex, age, cell counts) and unknown covariates:

$$y_j = \mathbf{D}_j^T \beta_j + \mathbf{C}^T \gamma + \epsilon \tag{5.1}$$

where y_j is the gene expression for gene j, $\mathbf{D}_j^T$ the scaled matrix with dosage values for the nearby genetic variants with its corresponding regression coefficients β_j, $\mathbf{C}$ the matrix of scaled known and unknown covariates and their corresponding regression coefficients γ, and the vector or residuals ϵ. Estimation of the unknown covariates was done by applying *cate* [Wang et al., 2015] to the observed expression data, leading to 5 unknown latent factors used. Those factors, together with the known covariates, were left unpenalized. To estimate the optimal penalization parameter λ, we used five-fold cross-validation as implemented in the R package *glmnet* [Friedman et al., 2010]. The obtained GI for index gene j consisted of a weighted linear combination of the dosage values of the selected nearby genetic variants, weighted by the obtained regression coefficients β_j, to obtain GI_j for index gene j:

$$GI_j = \mathbf{D}_j^T \beta_j \tag{5.2}$$

where GI_j is a vector of values. We then evaluated its predictive ability in the test set by employing Analysis of Variance (ANOVA) to evaluate the added predictive power of the GI over the covariates and neighbouring GIs (within 1Mb), as reflected by the F-statistic ($F > 10$).

Earlier work related to establishing putative causal relations between gene expression and phenotypic traits [Gamazon et al., 2015; Gusev et al., 2016] shows overlap with our proposed method, but also some distinct differences. First, none of them attempt to account for pleiotropy. Furthermore, two earlier studies [Gamazon et al., 2015; Gusev et al., 2016] have both used a single top eQTL SNP as a GI, or have used all nearby genetic variants, without feature selection [Gusev et al., 2016]. While not performing feature selection at all may improve

the predictive ability over our method, it may also induce pleiotropy or LD. This may especially be the case since the authors have used a 1Mb window around a gene, and have not corrected for pleiotropy or LD. The other study [Gamazon et al., 2015] has indeed used feature selection using elastic net, which also leads to sparse models, albeit slightly less sparse than our proposed method.

Testing for *trans*-effects

Using linear regression, we tested for an association between each GI j and the expression of potential target genes k *in trans* (> 10Mb), while correcting for known (cohort, sex, age, red and white blood cell counts) and unknown covariates, as well as GIs of nearby genes (< 1Mb):

$$y_k = GI_j\phi_j + \mathbf{C}^T\gamma + \mathbf{G}_j^T + \epsilon \tag{5.3}$$

where we test for the significance of the regression coefficient ϕ, and $\mathbf{G}_j$ represents the GIs of index genes near the current index gene j. Missing observations in the measured red blood cell count (RBC) and white blood cell counts (WBC) were imputed using the R package *pls*, as described earlier [van Iterson et al., 2017]. Any inflation or bias in the test-statistics was estimated and corrected for using the R package *bacon* [van Iterson et al., 2017]. Correction for multiple testing was done using Bonferroni (Wald $P < 7x10^{-10}$). The resulting networks were visualized using the R packages *network* and *ndtv*.

Enrichment analyses

Functional analysis of gene sets was performed for GO Molecular Function annotations using DAVID [Huang da et al., 2009], providing a custom background consisting of all genes with a predictive GI ($F > 10$). Fisher's exact test was employed to specifically test for an enrichment of transcription factors using manually curated database of transcription factors [Vaquerizas et al., 2009].

Simulation study

Simulating data of genetic instruments (GIs), their corresponding gene expression measurements, and a target gene was done as follows:

- Generate two normally distributed, correlated genetic instruments, where the correlation between the different GIs represents LD/pleiotropy. We used 5 different values for the correlation r_{GI} as estimated in our data, corresponding to the minimum absolute correlation in our identified effects, the 25th, 50th, 75th percentile, and the maximum value.
- Generate the index gene expression by creating a new normally distributed variable correlated to the index GI. Again, we used 5 different values for the correlation $r_{GI,index}$, using estimations from our data, corresponding to the minimum absolute correlation in our identified effects, the 25th, 50th, 75th percentile, and the maximum value.

- Similarly, generate the nearby gene expression by creating a new normally distributed variable correlated to the nearby GI. Here, we also used 5 values for the correlation $r_{GI,nearby}$ corresponding to the minimum absolute correlation in our identified effects, the 25th, 50th, 75th percentile, and the maximum value.
- Lastly, generate the target gene by creating a new normally distributed variable correlated to either the index gene ($r_{index,target}$), or the nearby gene ($r_{nearby,target}$), depending on the hypothesis tested (Supplementary Fig. 6). We again used different values for these correlations.

We have simulated two scenarios (see figure below), corresponding to the alternative and null hypotheses:

- The GI of the index gene causally influences its corresponding index gene, which influences the target gene (Supplementary Fig. 6a).
- The GI of a nearby gene causally influences its corresponding gene, which influences the target gene (Supplementary Fig. 6b).

For both scenarios, we have tested the effect of the index GI (β_{index}) on the target gene y, both corrected for LD by including the GI of the nearby gene GI_{nearby},

$$y = \beta_{index} GI_{index} + \beta_{nearby} GI_{nearby} + \epsilon \tag{5.4}$$

and without correcting for LD:

$$y = \beta_{index} GI_{index} + \epsilon \tag{5.5}$$

For each set of different settings (*i.e.*, different correlations among the different variables), this lead to the testing of four models, two for each scenario (Supplementary Fig. 6). Repeating this analysis 500 times for each unique set of settings, we then were able to estimate the power of each model by calculating the proportion of times the P-value was smaller than 0.05:

$$\frac{1}{500} \sum_{i=1}^{500} I(P_i < 0.05) \tag{5.6}$$

URLs

Look-ups can be performed using our interactive gene network browser at `http://bbmri.researchlumc.nl/NetworkBrowser/`. Data were generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (`http://www.glimDNA.org`). Webpages of participating cohorts: LifeLines, `http://lifelines.nl/lifelines-research/general`; Leiden Longevity Study, `http://www.healthy-ageing.nl/` and `http://www.leidenlangleven.nl/`; Netherlands Twin Registry, `http://www.tweelingenregister.org/`; Rotterdam Studies, `http://www.erasmusmc.nl/epi/research/The-Rotterdam-Study/`; Genetic Research in Isolated Populations program, `http://www.epib.nl/research/geneticepi/research.html#gip`; CODAM study, `http://www.carimmaastricht.nl/`; PAN study, `http://www.alsonderzoek.nl/`.

Data availability

Raw data were submitted to the European Genome-phenome Archive (EGA) under accession EGAS00001001077 [https://www.ebi.ac.uk/ega/studies/EGAS00001001077].

Code availability

R code is available from `https://git.lumc.nl/r.luijk/DirectedGeneNetworks`. This repository describes the main analyses done.

Acknowledgments

This research was financially supported by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO, numbers 184.021.007 and 184.033.111). Samples were contributed by LifeLines, the Leiden Longevity Study, the Netherlands Twin Registry (NTR), the Rotterdam Study, the Genetic Research in Isolated Populations program, the Cohort on Diabetes and Atherosclerosis Maastricht (CODAM) study and the Prospective ALS study Netherlands (PAN). We thank the participants of all aforementioned biobanks and acknowledge the contributions of the investigators to this study. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

References

Adamson, B. et al. [2016]. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response, *Cell* **167**(7): 1867–1882 e21.

Boomsma, D. I. et al. [2002]. Netherlands Twin Register: A Focus on Longitudinal Research, *Twin Research* **5**(5): 401–406.

Brion, M. J., Shakhbazov, K. and Visscher, P. M. [2013]. Calculating statistical power in Mendelian randomization studies, *Int J Epidemiol* **42**(5): 1497–1501.

Bruning, O. et al. [2016]. Confounding Factors in the Transcriptome Analysis of an In-Vivo Exposure Experiment, *PLoS One* **11**(1): e0145252.

Cai, L., Wang, Y., Wang, J. F. and Chou, K. C. [2011]. Identification of proteins interacting with human SP110 during the process of viral infections, *Med Chem* **7**(2): 121–126.

Chang, P. A. et al. [2013]. Identification of human patatin-like phospholipase domain-containing protein 1 and a mutant in human cervical cancer HeLa cells, *Mol Biol Rep* **40**(10): 5597–5605.

Ciani, E., Hoffmann, A., Schmidt, P., Journot, L. and Spengler, D. [1999]. Induction of the PAC1-R (PACAP-type I receptor) gene by p53 and Zac, *Molecular Brain Research* **69**(2): 290–294.

Cunningham, L. L., Matsui, J. I., Warchol, M. E. and Rubel, E. W. [2004]. Overexpression of Bcl-2 prevents neomycin-induced hair cell death and caspase-9 activation in the adult mouse utricle in vitro, *J Neurobiol* **60**(1): 89–100.

Davey Smith, G. and Hemani, G. [2014]. Mendelian randomization: genetic anchors for causal inference in epidemiological studies, *Hum Mol Genet* **23**(R1): R89–98.

de la Fuente, A. [2010]. From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases, *Trends in Genetics* **26**(7): 326–333.

Deelen, J. et al. [2014a]. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age, *Human Molecular Genetics* **23**(16): 4420–4432.

Deelen, P. et al. [2014b]. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration, *BMC Research Notes* **7**(1): 901.

Dixit, A. et al. [2016]. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens, *Cell* **167**(7): 1853–1866 e17.

Dobin, A. et al. [2013]. STAR: ultrafast universal RNA-seq aligner, *Bioinformatics* **29**(1): 15–21.

Evans, D. M. and Davey Smith, G. [2015]. Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality, *Annual Review of Genomics and Human Genetics* **16**(1): 327–350.

Fasching, L. et al. [2015]. TRIM28 represses transcription of

endogenous retroviruses in neural progenitor cells, *Cell Rep* **10**(1): 20–28.

Freeman, G., Cowling, B. J. and Schooling, C. M. [2013]. Power and sample size calculations for Mendelian randomization studies using one genetic instrument, *Int J Epidemiol* **42**(4): 1157–1163.

Friedman, J., Hastie, T. and Tibshirani, R. [2010]. Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software* **33**(1).

Fritz, M. S. and MacKinnon, D. P. [2007]. Required Sample Size to Detect the Mediated Effect, *Psychological science* **18**(3): 233–239.

Frojdo, S., Vidal, H. and Pirola, L. [2009]. Alterations of insulin signaling in type 2 diabetes: a review of the current evidence from humans, *Biochim Biophys Acta* **1792**(2): 83–92.

Gamazon, E. R. et al. [2015]. A gene-based association method for mapping traits using reference transcriptome data, *Nature Genetics* **47**(9): 1091–1098.

Garvin, A. J. et al. [2013]. The deSUMOylase SENP7 promotes chromatin relaxation for homologous recombination DNA repair, *EMBO Rep* **14**(11): 975–983.

GTEx Consortium [2017]. Genetic effects on gene expression across human tissues, *Nature* **550**: 204.

Gusev, A. et al. [2016]. Integrative approaches for large-scale transcriptome-wide association studies, *Nature Genetics* **48**(3): 245–252.

Hoffmann, A. and Spengler, D. [2012]. Transient neonatal diabetes mellitus gene Zac1 impairs insulin secretion in mice through Rasgrf1, *Mol Cell Biol* **32**(13): 2549–2560.

Hofman, A. et al. [2013]. The Rotterdam Study: 2014 objectives and design update, *European Journal of Epidemiology* **28**(11): 889–926.

Howie, B. N., Donnelly, P. and Marchini, J. [2009]. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies, *plos genetics* **5**(6).

Huang da, W., Sherman, B. T. and Lempicki, R. A. [2009]. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc* **4**(1): 44–57.

Huisman, M. H. et al. [2011]. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology, *J Neurol Neurosurg Psychiatry* **82**(10): 1165–1170.

Jaitin, D. A. et al. [2016]. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq, *Cell* **167**(7): 1883–1896 e15.

Jiang, C., Xuan, Z., Zhao, F. and Zhang, M. Q. [2007]. TRED: a transcriptional regulatory element database, new entries and other development, *Nucleic Acids Research* **35**(Database issue): D137–D140.

Joehanes, R. et al. [2017]. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies, *Genome Biology* **18**(1): 16.

Joshi Fass, J., N. [2011]. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33).

Lee, M. N. et al. [2013]. Identification of regulators of the innate immune response to cytosolic DNA and retroviral infection by an integrative approach, *Nat Immunol* **14**(2): 179–185.

Lee, T. I. and Young, R. A. [2013]. Transcriptional regulation and its misregulation in disease, *Cell* **152**(6): 1237–1251.

Lemire, M. et al. [2015]. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci, *Nat Commun* **6**: 6326.

Li, X. et al. [2007]. Role for KAP1 serine 824 phosphorylation and sumoylation/desumoylation switch in regulating KAP1-mediated transcriptional repression, *J Biol Chem* **282**(50): 36177–36189.

Lin, B. D., Willemsen, G., Abdellaoui, A., Bartels, M., Ehli, E. A., Davies, G. E., Boomsma, D. I. and Hottenga, J. J. [2016]. The Genetic Overlap Between Hair and Eye Color, *Twin Research and Human Genetics* **19**(6): 595–599.

Lin, Y. S. et al. [2011]. Dysregulated brain creatine kinase is associated with hearing impairment in mouse models of Huntington disease, *J Clin Invest* **121**(4): 1519–1523.

Lukic, S., Nicolas, J. C. and Levine, A. J. [2014]. The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses, *Cell Death Differ* **21**(3): 381–387.

Marbach, D. et al. [2016]. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases, *Nat Methods* **13**(4): 366–370.

Martin, M. [2011]. Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal* **17**(1): 10.

McGregor, K. et al. [2016]. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies, *Genome Biol* **17**: 84.

Mishra, J. et al. [2015]. Role of Janus Kinase 3 in Predisposition to Obesity-associated Metabolic Syndrome, *J Biol Chem* **290**(49): 29301–29312.

Orru, V. et al. [2013]. Genetic variants regulating immune cell levels in health and disease, *Cell* **155**(1): 242–256.

Ozcan, L. et al. [2015]. Treatment of Obese Insulin-Resistant Mice With an Allosteric MAPKAPK2/3 Inhibitor Lowers Blood Glucose and Improves Insulin Sensitivity, *Diabetes* **64**(10): 3396–3405.

Pearl, J. [2009]. *Causality: Models, Reasoning, and Inference*, 2nd edn, Cambridge University Press, New York.

Peters, L. M. et al. [2007]. Signatures from tissue-specific MPSS libraries identify transcripts preferentially expressed in the mouse inner ear, *Genomics* **89**(2): 197–206.

Pierce, B. L. et al. [2014]. Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians, *PLOS Genetics* **10**(12): e1004818.

Roederer, M. et al. [2015]. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis, *Cell* **161**(2): 387–403.

Schaefer, K. A. et al. [2017]. Unexpected mutations after CRISPR-Cas9 editing in vivo, *Nat Meth* **14**(6): 547–548.

Schoenmaker, M. et al. [2005]. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study, *European Journal of Human Genetics* .

Shin, J. B. et al. [2007]. Hair bundles are specialized for ATP delivery via creatine kinase, *Neuron* **53**(3): 371–386.

Sobel, M. E. [1982]. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models, *Sociological Methodology* **13**: 290.

Solovieff, N. et al. [2013]. Pleiotropy in complex traits: challenges and strategies, *Nature reviews. Genetics* **14**(7): 483–495.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. [2003]. A gene-coexpression network for global discovery of conserved genetic modules, *Science* **302**(5643): 249–255.

Tadros, S. F., D'Souza, M., Zhu, X. and Frisina, R. D. [2008]. Apoptosis-related genes change their expression with age and hearing loss in the mouse cochlea, *Apoptosis* **13**(11): 1303–1321.

Teider, N. et al. [2010]. Neuralized1 causes apoptosis and downregulates Notch target genes in medulloblastoma, *Neuro Oncol* **12**(12): 1244–1256.

The Genome of the Netherlands Consortium et al. [2014]. Whole-genome sequence variation, population structure and demographic history of the Dutch population, *Nature Genetics* **46**(8): 818–825.

Tigchelaar, E. F. et al. [2015]. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics, *BMJ Open* **5**(8): e006772.

Valente, T., Junyent, F. and Auladell, C. [2005]. Zac1 is expressed in progenitor/stem cells of the neuroectoderm and mesoderm during embryogenesis: differential phenotype of the Zac1-expressing cells during development, *Dev Dyn* **233**(2): 667–679.

van Greevenbroek, M. M. J. et al. [2011]. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study), *European Journal of Clinical Investigation* **41**(4): 372–379.

van Iterson, M. et al. [2017]. Controlling bias and inflation in

epigenome- and transcriptome-wide association studies using the empirical null distribution, *Genome Biol* **18**(1): 19.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. and Luscombe, N. M. [2009]. A census of human transcription factors: function, expression and evolution, *Nat Rev Genet* **10**(4): 252–263.

Varrault, A. et al. [2006]. Zac1 regulates an imprinted gene network critically involved in the control of embryonic growth, *Dev Cell* **11**(5): 711–722.

Vlachos, P., Nyman, U., Hajji, N. and Joseph, B. [2007]. The cell cycle inhibitor p57(Kip2) promotes cell death via the mitochondrial apoptotic pathway, *Cell Death Differ* **14**(8): 1497–1507.

Vock, C., Doring, F. and Nitz, I. [2008]. Transcriptional regulation of HMG-CoA synthase and HMG-CoA reductase genes by human ACBP, *Cell Physiol Biochem* **22**(5-6): 515–524.

Vogler, M. [2012]. BCL2A1: the underdog in the BCL2 family, *Cell Death Differ* **19**(1): 67–74.

Wang, J., Zhao, Q., Hastie, T. and Owen, A. B. [2015]. Confounder Adjustment in Multiple Hypothesis Testing, *ArXiv e-prints* .

Westra, H. J. et al. [2013]. Systematic identification of trans eQTLs as putative drivers of known disease associations, *Nature Genetics* **45**(10): 1238–1243.

Xu, D., Yin, C., Wang, S. and Xiao, Y. [2013]. JAK-STAT in lipid metabolism of adipocytes, *JAKSTAT* **2**(4): e27203.

Yada, T. et al. [1998]. Autocrine Action of PACAP in Islets Augments Glucose-Induced Insulin Secretiona, *Annals of the New York Academy of Sciences* **865**(1 VIP, PACAP, A): 451–457.

Yan, Y. et al. [1997]. Ablation of the CDK inhibitor p57Kip2 results in increased apoptosis and delayed differentiation during mouse development, *Genes & Development* **11**(8): 973–983.

Yang, F. et al. [2017]. Identifying cis -mediators for trans -eQTLs across many human tissues using genomic mediation analysis, *Genome research* pp. 1–13.

Zhernakova, D. V. et al. [2017]. Identification of context-dependent expression quantitative trait loci in whole blood, *Nature Genetics* **49**(1): 139–145.

Zhu, Z. et al. [2016]. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets, *Nature Genetics* **48**(5): 481–487.

6 | Discussion

Summary of results

In this thesis, we aimed to better understand how genetic variation affect the processes underlying health and disease, as trait-associated genetic variants are often located in non-coding regions. This hampers their interpretability, and has prompted the exploration of their effects on transcriptional regulation, a process that is crucial in the development of common and complex diseases [Lee and Young, 2013]. To do this, we have used a variety of omics data in a large collection of individuals from the general population. Using these data, we have investigated the local and distal effects of genetic variants on other molecular phenotypes, such as gene expression levels and DNA methylation levels of CpG sites, and the underlying mechanisms. This has resulted in a framework enabling the exploration of causal hypotheses about transcriptional regulation using genetics as a causal anchor. The approaches used in this thesis have yielded insight into transcriptional (dys)regulation and several underlying mechanisms. This will be helpful in better understanding how transcriptional regulation contributes to complex phenotypes related to health and disease, such as common diseases.

As a first step in investigating transcriptional regulation, local effects of genetic variants on other molecular phenotypes have often been investigated, most notably gene expression levels and DNA methylation levels. In **chapters 2** and **3**, we look into the local (*cis*) effects on DNA methylation levels of nearby CpG sites (methylation QTL mapping; meQTL). First, in **chapter 2**, we explored local (*cis*) methylation quantitative trait loci (meQTL) mapping from both a biological and methodological perspective. In *cis*-meQTL mapping, the interest is often in the identification of CpG sites under the influence of genetic variation. To achieve this, a list of significant SNP-CpG pairs is typically compiled with a false discovery rate of 5%, meaning 95% of all SNP-CpG pairs will be true positives. However, it is commonly implied that 95% of the CpG sites occurring in that list are also truly associated with some genetic variant. However, as not all CpG sites may occur equally often in this list, this claim does not hold true. We developed an alternative approach that specifically aims to provide a list of CpG sites under the influence of genetic variation, where the FDR among individual CpG sites is controlled at 5%. Simulations provided evidence for the effectiveness of our method, and showed the other, commonly used method, indeed leads to increased false discovery rates. By this new approach we identified CpG sites whose methylation levels are truly influenced by local genetic variation. From a biological perspective, we observe that the genetic variant most strongly associated with a particular CpG site is often in close proximity to that CpG site (<50kb), much closer than the search windows usually employed (100kb-500kb windows).

In **chapter 3** we used a much larger set of whole-blood samples for *cis*-meQTL mapping, identifying a *cis*-meQTL for a third of all interrogated CpG sites (<250kb). Many of these CpG sites exhibited relatively high variation in the observed methylation levels as indicated by the variance, despite the effect sizes typically being small, not more than several percentage points. In addition

to *cis*-meQTL mapping, we aimed to explore the downstream effects of genetic variants known to affect biological traits. As such variants are often located in non-coding regions, their functional consequences are often unknown. Hence, we used meQTL mapping to identify their effects on DNA methylation levels *in trans* (*i.e.*, investigating long-range effects, >5Mb). We show that one third of all interrogated variants affect DNA methylation levels at multiple CpG sites *in trans*. In addition, many of these genetic variants are associated with the expression levels of nearby transcription factors, which may not necessarily be the nearest gene [Peeters et al., 2014], while affecting methylation levels of CpG sites known to be located in binding sites of that particular transcription factor. Several of these variants had *trans*-effects on methylation levels of CpG sites across all chromosomes, suggesting these variants affect large regulatory networks. In addition, many of the genetic variants also affected the gene expression levels of nearby transcription factors. Interestingly, many of the affected CpG sites were located in known binding sites of that particular transcription factor. These findings hint at a possible mechanism underlying the *trans*-effects: a genetic variant could alter transcription factor levels, which in turn affects target gene expression. At the same time, we also overlaid physical interchromosomal contacts in blood-related cell types [Rao et al., 2014] with the *trans*-meQTLs. Using information on the three-dimensional conformation of the chromatin, we identified an overrepresentation of variant-CpG site pairs, each located on a different chromosome, that are located in regions known to physically interact with each other, if only momentarily. This presents a second way in which these *trans*-meQTLs could come about, and is reminiscent of the physical proximity of the genetic variant and target CpG site of *cis*-meQTLs. Hence, it may be possible that the underlying mechanism is similar to that underlying *cis*-meQTLs, where sequence-specific binding of transcription factors could induce changes in, among others, DNA methylation [Do et al., 2017]. Together, these results showcase the utility of (*trans*-)meQTL mapping in identifying downstream effects of trait-associated genetic variants whose functional consequences are not known.

In **chapter 4**, we used similar methodology as in the previous chapters to investigate X-chromosome inactivation (XCI), a phenomenon that has only been extensively studies in mice. XCI a process by which one of two copies of the X-chromosome is silenced in order to achieve dosage equality between males and females, and where DNA methylation seems to be heavily involved [Sharp et al., 2011; Yasukochi et al., 2010]. Using *trans*-meQTL mapping in both female and male samples separately, and only using autosomal genetic variants, we were able to identify three autosomal genetic loci having female-specific effects on X-chromosome DNA methylation. All three loci were also associated with changes in the expression levels of these nearby genes (*i.e.*, *cis*-eQTLs), suggesting the *trans*-meQTL effects were mediated by these *cis*-eQTL effects. This seemingly confirms the mechanism underlying *trans*-effects posited in **chapter 3**. More interestingly, however, was the observation that many of the affected X-chromosome CpG sites were located near, and associated with genes known to escape XCI to varying degrees. Previous studies have shown that genes escaping XCI are often related to mental impairment [Zhang et al., 2013]. Not only does this suggest there is a

genetic component to variable escape from XCI in humans, and that this genetic basis might be related to other biological phenotypes, such as mental impairment.

In **chapter 5**, we aimed to establish directed gene-gene interactions using observational data, as dysregulation of these gene networks often underlies common diseases. However, directly correlating the expression levels of two genes is hampered by confounding factors and reverse causation. Inter-relating all genes will therefore result in large, inter-connected gene networks. To circumvent these issues, we used a Mendelian Randomization (MR)-type approach [Davey Smith and Hemani, 2014], which explicitly uses genetics as a causal anchor to discover these directed gene networks. Strong correlations among genetic variants (*i.e.*, linkage disequilibrium; LD) and pleiotropy commonly cause several neighboring genes to be causally linked with the same distal gene. In this chapter, we adjusted for these two factors, in an effort to identify the causal driver of the association. This approach has led to the identification of genes that plausibly have a causal effect on the expression levels of up to 33 genes *in trans*. This approach illustrates the utility of MR in suggesting causal relationships regarding gene regulation, possibly prioritizing which genes to look into first, *e.g.*, in an experimental setting or in different tissues.

Nature's experiment

The molecular QTL mapping approaches described and applied in this thesis are instrumental in making a first step towards better understanding transcriptional regulation, the main aim of this thesis. However, this approach is unable to inter-relate non-genetic molecular phenotypes, such as gene expression levels or DNA methylation levels, which is often plagued by confounding and other biases, such as reverse causation. **Chapter 5** describes how we used Mendelian Randomization-type (MR) approach to establish directed gene-gene associations.

This approach is based on the observation that many genes have a strong local genetic component, an observation that is used to build genetic instruments predictive of their expression levels. In that regard, MR is somewhat analogous to a randomized controlled trial (RCT). Instead of the treatment being randomized over the study participants, MR uses the random distribution of alleles from parents to offspring. The genetic instrument could be thought of as the exposure that manipulates the gene expression levels, similar to a treatment altering the risk factor for a disease. There are some distinct differences between MR and a RCT, however [Nitsch et al., 2006]. Most notably, a randomized treatment can usually only be investigated for short-term effects on an intermediary phenotype, whereas a randomized genotype may have long-term effects.

As genetic variation is usually free from any non-genetic confounding, any association between the genetic instrument and a potential target gene is possibly causal. Using this approach, we identified directed gene-networks. Many of the potentially causal genes were again known transcription factors or known to be involved in chromatin remodeling, making it likely they were indeed the drivers behind the associations. Furthermore, we identified several previously

unknown target genes, despite earlier *trans*-eQTL efforts. These findings make MR a powerful approach, as it is able to simulate an RCT, while circumventing the practical and ethical limitations of human experimentation. Fairly recent developments in experimental approaches (*i.e.*, CRISPR-cas9) are complementary in establishing causality on a large scale [Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016]. However, these approaches rely on in vitro experimentation, whereas MR can be done in vivo and on a whole-organism level, and their results could be difficult to interpret given that CRISPR-cas9 may have off-target effects [Schaefer et al., 2017].

Even though there is great utility in MR, the method itself is not without limitations. From a statistical perspective, MR often suffers from low statistical power, especially when applied to inherently variable omics data. Measurement error, technical batch effects, environmental factors, and other, possibly even unknown factors contributing to the measurement variability that obscure the small effect sizes genetic variants usually have on any molecular trait. Because the identified effects using MR depend on the strength of the genetic instrument, the statistical power to detect *trans*-effects is limited. A possible solution is to either measure the same variable in a single large set of samples, or combine different existing datasets, like we have done in this thesis.

Secondly, linkage disequilibrium (LD) could hinder isolating a causal driver behind an association. Similar to in a GWAS, local LD will inevitably cause several correlated genetic instruments to be associated with the same target phenotype, as is the case in the directed gene-gene associations described in **chapter 5**. As a solution, we have corrected for the genetic instruments of nearby genes (within 1Mb). Correcting for local LD may help, but could also nullify all associations, making it impossible to pinpoint the causal driver behind the association on the basis of statistical evidence alone. Even when local LD is accounted for, long-range LD could still lead to the false conclusion of causality, when in reality another gene is responsible for the association.

Lastly, pleiotropy could have the same effect as LD on the statistical analysis, resulting in the association of several genes with the same target gene or another outcome. From a biological perspective, however, they do indeed differ. In the case of LD, genetic variation influences both the index gene, as well as the target gene. Pleiotropy causes one genetic instrument to independently influence both the driver gene, as well as the target gene. Extensive checks and additional analyses could minimize the risk of marking an association as directed or even causal, but more in-depth investigation of the results using external data may give even more confidence in these results.

From polygenic to omnigenic

The methods in this thesis that use genetic variation on a genome-wide scale, such as molecular QTL mapping and Mendelian Randomization-type approaches, have the potential to generate many hypotheses for further research into transcriptional

regulation. This may especially be relevant considering only a small fraction of all genes in the genome have been directly investigated in a dedicated paper [Stoeger et al., 2018]. In fact, only 2,000 out roughly 19,000 known genes have been directly studied in a paper dedicated to a single gene. However, these papers do make up 90 percent of all scientific research in their respective fields in recent years. Data-driven approaches using naturally occurring genetic variation could circumvent the practical and ethical limitations of experimentation in either humans or animals, and even perform many *in vivo* "experiments" simultaneously. This has the potential to greatly increase the number of genes directly investigated. This is especially true given that multiple omics datasets are becoming increasingly available in larger sample sizes.

Ironically, increased sample sizes may also bring about some new challenges. For molecular QTL mapping efforts, for example, there is a seemingly ever-increasing number of identified *cis*-, and *trans*-effects across the entire genome that go along with increased sample sizes, and large meta-analyses should only add to this. For MR, limited statistical power may still limit the number of identified associations, but in return allows for the formulation causal hypotheses about statistical associations. As mentioned at the beginning of this thesis, this should provide better insight into the genetic underpinnings of transcriptional regulation. In the case of molecular QTL mapping, however, it could be that more and more genetic variants are associated with many other molecular phenotypes. This has been one of the hindering factors when trying to formulate causal hypotheses about how genetics influences transcriptional (dys)regulation in health and disease.

At the beginning of this thesis, we mentioned the thought-provoking idea called the omnigenic model [Boyle et al., 2017]. The authors hypothesize that a limited number of core disease-related genes could account for a modest to moderate amount of inter-individual variation. However, many, if not all, genes expressed in a disease-relevant tissue could influence a phenotype through small, possibly indirect effects on these core disease-related genes. This is analogous to the transcription factors identified to be causally related to their target genes *in trans* (**chapters 3 and 5**), which could be the core genes. LD and pleiotropy then cause nearby genes to also be associated with the same target gene, even though they might play a less important role in regulating target gene expression.

If this model holds true, then this has far-reaching consequences for genomics research. As long as one has sufficient statistical power, one would be able to identify a plethora of *cis*-, and *trans*-effects for virtually all genes. Prioritizing these markers for elaborate functional studies into the relation of the variants with the trait or disease endpoint would then be the next challenge, distinguishing between core genes having a direct, causal effect and those that have a secondary, indirect effect. Furthermore, one would have to establish how much of the phenotypic variance is explained by these core effects. Lastly, taking into account the possible tissue-specificity would be imperative, as the importance of each effect may be tissue-specific. For example, large-scale efforts like the GTEx project have already established tissue-specific effects of genetic variants on gene expression values, both *in cis* and *in trans* [GTEx Consortium, 2017] (GTEx

Consortium, 2017). Such efforts could already give an indication which genes are under genetic control in which tissues, helping to prioritize which genes to look into.

Observation, replication, and experimentation: a case for triangulation

Replicating the results from earlier studies, *i.e.*, confirming the outcomes, is necessary for the validity of the result and the scientific method in general. In this thesis, we replicated many of the *trans*-meQTLs (**chapter 3**) and genetic effects on XCI (**chapter 4**) in independent sets of samples. However, recent concerns regarding replication were raised [Munafò and Davey Smith, 2018], where the authors argue that while replication of results is laudable and desirable, it is not sufficient. Any flaw in the original analysis will be repeated in a replication study, thereby yielding the same, possibly skewed, result. Routine replication then, so they argue, has the potential to only make matters worse, as consistent replication of a result may make it seem as an established truth, even when the finding is false. Instead, the authors propose *triangulation* – using several distinct lines of evidence to verify a result [Munafò and Davey Smith, 2018], rather than replication, should be the main focus when judging the validity of scientific research.

Indeed, using different methodologies, each with their own merits and faults, should be employed to provide a definitive answer to a research question. Hence, we have attempted to experimentally validate the results in **chapter 4** by knocking down several of the genes identified there, and investigating changes in X-chromosome DNA methylation in the female RPE cell line. While knockdown was generally successful, we did not observe the same changes in X-methylation (data not shown). Such changes in methylation would indicate a re-activated inactivate X-chromosome (Xi). This observation is similar to those made in mouse embryonic fibroblasts [Gdula et al., 2018], where no changes in Xi expression were found upon knockdown the same gene as in **chapter 4**. One possible reason could be that the gene under study is involved in XCI initiation, whereas these experiments would only show effects on XCI maintenance. This implies that timing, too, could play an important role in experimental validation of a result obtained using data-analytics. Our population genomics approach would be able to pick up on these effects, as they would reveal effects in both stages of XCI. In addition, tissue-specificity might have hindered the replication, as the knock-down experiment was done in in an RPE cell line, not whole-blood. Lastly, if escape from XCI results from an interplay between several genes, then this would not be replicated by only knocking down a single gene. As a result, experimental validation, especially in this particular case, possibly requires meticulous, large-scale, and lengthy experiments.

Systematically validating the *trans*-meQTLs (**chapter 3**) and directed gene-gene associations (**chapter 5**) using experiments would be even more challenging, due to the sheer number of leads generated in these chapters. Together, these two issues highlight the difficulty of reproducing findings using different research

modalities, and conceivably requires a great deal of time and resources. As a result, publishing the findings on a single research question will take lots of time and resources, which may not be feasible. Still, other research modalities should be explored, as well as a variety of relevant tissues. Specifically, lab experiments do fulfil an important role in science, as they provide a relatively controlled environment in which to manipulate a variable, like gene expression levels. As a result, this manipulation virtually guarantees a causal relationship between two variables.

Another research modality that could further address some of the limitations of the cohort studies described in this thesis is a longitudinal study. In such a study, several features are repeatedly measured, *i.e.*, at different time points. For example, genes identified in this thesis could be investigated over time, allowing changes in their expression to be associated with a relevant biological phenotype, such as target gene expression or another biological trait. This is especially a powerful approach when such data is gathered in several tissues, or when an exposure is administered. Such exposures could be clinical exposures, but also lifestyle interventions.

Summing up, whenever practical and ethical limitations allow, it is imperative to employ several methodologies to validate the results obtained from population genomics studies, as this ensures that the research findings that seem to be true, actually are true. At the same time, though, it is important to keep in mind that one may fail to validate the results obtained using a population genomics approach using another modality, such as in vitro experiments. While this could mean the initial result is false, it could also mean the design of the follow-up experiment was insufficient to reproduce the first observation. Making this distinction is important, but hard, and could mean that for a specific research question, any one approach, such as a population genomics approach, is the only feasible option.

Final remarks

The challenges of investigating transcriptional regulation using the next wave of omics studies are laid out. In order to start getting a better understand of the transcriptional (dys)regulation underlying complex biological traits, including common diseases, combining efforts to generate larger sample sizes will be necessary, though not sufficient. In addition to larger sample sizes, measuring different omics data in the same set of individuals (*e.g.*, epigenomics, transcriptomics, proteomics, metabolomics, among others), will allow researchers to get a complete picture of the mechanisms underlying a biological trait. Following these up using different research modalities will be imperative.

In doing so, working towards causal hypotheses using MR-type approaches should be a priority, ultimately providing a mechanistic insight into how a trait-associated genetic variant affects a phenotypic outcome through effects on different omics layers. While we have attempted to control for local LD and pleiotropy, if the omnigenic model holds true, this may be more difficult than

previously appreciated.

Overcoming all of these challenges is no easy feat. One ought to be cautious when interpreting small effect sizes, and take heed when routinely replicating previous findings. Ultimately, coming at research questions from multiple angles will be essential to provide a valid answer to a research question. However, other research modalities, such as lab experiments, are not always feasible due to ethical considerations or different practical issues. As such, MR-type approaches applied to large-scale datasets should take a leading role in exploring transcriptional regulation in a variety of tissues, aiming to ultimately uncovering the mechanisms underlying health and disease.

References

Adamson, B. et al. [2016]. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response, *Cell* **167**(7): 1867–1882 e21.

Boyle, E. A. et al. [2017]. An Expanded View of Complex Traits: From Polygenic to Omnigenic, *Cell* **169**(7): 1177–1186.

Davey Smith, G. and Hemani, G. [2014]. Mendelian randomization: genetic anchors for causal inference in epidemiological studies, *Hum Mol Genet* **23**(R1): R89–98.

Dixit, A. et al. [2016]. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens, *Cell* **167**(7): 1853–1866 e17.

Do, C., Shearer et al. [2017]. Genetic-epigenetic interactions in cis: A major focus in the post-GWAS era, *Genome Biology* **18**(1).

Gdula, M. R. et al. [2018]. The non-canonical SMC protein SmcHD1 antagonises TAD formation on the inactive X chromosome.

GTEx Consortium [2017]. Genetic effects on gene expression across human tissues, *Nature* **550**: 204.

Jaitin, D. A. et al. [2016]. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq, *Cell* **167**(7): 1883–1896 e15.

Lee, T. I. and Young, R. A. [2013]. Transcriptional regulation and its misregulation in disease, *Cell* **152**(6): 1237–1251.

Munafò, M. R. and Davey Smith, G. [2018]. Robust research needs many lines of evidence, *Nature* **553**(7689): 399–401.

Nitsch, D. et al. [2006]. Limits to causal inference based on mendelian randomization: A comparison with randomized controlled trials, *American Journal of Epidemiology* **163**(5): 397–403.

Peeters, S. B., Cotton, A. M. and Brown, C. J. [2014]. Variable escape from X-chromosome inactivation: identifying factors that tip the scales towards expression, *Bioessays* **36**(8): 746–756.

Rao, S. S. P. et al. [2014]. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping, *Cell* **159**(7): 1665–1680.

Schaefer, K. A. et al. [2017]. Unexpected mutations after CRISPR-Cas9 editing in vivo, *Nat Meth* **14**(6): 547–548.

Sharp, A. J., Stathaki, E., Migliavacca, E., Brahmachary, M., Montgomery, S. B., Dupre, Y. and Antonarakis, S. E. [2011]. DNA methylation profiles of human active and inactive X chromosomes, *Genome Res* **21**(10): 1592–1600.

Stoeger, T. et al. [2018]. Large-scale investigation of the reasons why potentially important genes are ignored, *PLOS Biology* **16**(9): e2006643.

Yasukochi, Y. et al. [2010]. X chromosome-wide analyses of genomic DNA methylation states and gene expression in male and female

neutrophils, *Proc Natl Acad Sci U S A* **107**(8): 3704–3709.

Zhang, Y. et al. [2013]. Genes that escape X-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving, *Mol Biol Evol* **30**(12): 2588–2601.

NEDERLANDSE SAMENVATTING

Moleculaire epidemiologie en transcriptionele regulatie

Epidemiologie is een wetenschappelijke discipline die zich bezig houdt met het bestuderen van veelvoorkomende ziekten en de oorzaken die daaraan ten grondslag liggen. Moleculaire epidemiologie is een deeldiscipline die specifiek ingaat op de moleculaire veranderingen die een rol spelen bij de verschillende aspecten omtrent ziekten. De vergaarde moleculaire informatie betreft vaak verscheidene aspecten van het genoom die een rol spelen bij transcriptionele regulatie. Dit proces bepaalt de mate waarin een gen geactiveerd wordt, wanneer dit gebeurt, en in welk weefsel.

Voor verschillende ziekten is al ontdekt dat misregulatie van genen gecorreleerd is met de ontwikkeling van deze ziekten. Deze correlaties zijn gevonden in grote genoombrede associatiestudies, die de verbanden tussen miljoenen genetische varianten en een ziekte onderzoeken, en zo genen vinden die mogelijk bij de ontwikkeling van deze ziekte betrokken zijn. Het gaat hierbij veelal om complexe ziekten die ontstaan door een samenspel van verschillende genen. Mede doordat deze ziekten niet eenduidige oorzaken hebben, geven de verbanden tussen genetische varianten en de aanwezigheid of afwezigheid van een ziekte nog niet direct een volledig beeld van welke genen er precies bij betrokken zijn, of hoe deze zich tot elkaar verhouden.

Andere informatie omtrent transcriptionele regulatie kan hierbij helpen, zoals het transcriptoom (expressieniveaus van verschillende genen), en het epigenoom. Epigenetische modificaties zijn moleculaire dimmers op het DNA. Door het veranderen van de toegangelijkheid van het DNA zelf, beïnvloeden ze de mate waarin genen afgeschreven kunnen worden, en dus ook hun expressieniveaus. Een belangrijke epigenetische dimmer die in dit proefschrift bestudeerd wordt, is DNA methylatie, waarbij een bepaald molecuul (een methylgroep) op het DNA geplaatst wordt. Hoewel de relatie tussen methylatieniveaus en expressieniveaus complex is, kan het over algemeen worden gezegd dat hogere methylatieniveaus leidt tot lagere activiteit van het gen. Het relateren van veranderingen in expressie-, en methylatieniveaus aan elkaar en aan

genetische markers van ziekte, kunnen een beter beeld geven welke netwerken van genen betrokken zijn bij de ontwikkeling van ziekten.

Deze aanpak is echter ook niet vrij van beperkingen. Een verband tussen een ziekte en de expressieniveaus van een gen betekent niet direct dat er ook een oorzakelijk verband is. In dit proefschrift proberen we deze oorzakelijk verbanden wel te leggen, om zo beter te begrijpen hoe transcriptionele regulatie werkt, en mogelijk tot ziekte kan leiden.

Van correlatie tot causatie

Oorzakelijke verbanden in de regulatie van genen onderling is helaas erg lastig aan te tonen. Experimentele manipulatie van het expressieniveau van genen is vaak noodzakelijk om te bewijzen dat er een causaal verband bestaat. Dit wordt vaak gebruikt om eerdere correlaties die bij genoombrede associatiestudies zijn gevonden te bevestigen in een laboratorium. Het is echter niet altijd haalbaar om alle mogelijke aanwijzingen uit deze studies op deze wijze te onderzoeken. Soms is het aantal te onderzoeken aanwijzingen simpelweg te groot, is het ethisch onverantwoord om deze experimenten in mensen uit te voeren, of heeft een experimentele manipulatie onverwachte neveneffecten.

Gelukkig is het mogelijk om, gebruik makende van observationele data, tot op zekere hoogte uitspraken te doen over causaliteit, en de noodzakelijke, tijdrovende experimenten beter te prioriteren. Quantitative trait loci (QTL) mapping kan hierbij een eerste stap in de goede richting zijn. Hierbij worden verbanden tussen genetische varianten, verspreid over het gehele genoom (genoombreed) en de methylatieniveaus van CpG-dinucleotiden gezocht. CpG-dinucleotiden zijn locaties op het genoom waarbij de twee basen (of “letters”) C en G elkaar opvolgen. Een vaak onderzochte vorm van DNA methylatie komt voor op deze locaties. Dergelijke verbanden helpen om gevonden associaties tussen specifieke genetische varianten en ziekten verder te onderzoeken.

Een volgende stap is het leggen van causale verbanden in de regulatie van verschillende genen onderling (gen-gen interacties). Dit kan met het gebruik van zogeheten Mendelian Randomization technieken, waardoor we met redelijke zekerheid oorzakelijke verbanden tussen de expressieniveaus van genen in mensen kunnen aanwijzen, zonder hierbij experimenten uit te voeren. Dit soort analyses vormen een belangrijke extra stap in de interpretatie van de resultaten van genoombrede associatiestudies.

In dit proefschrift pogen we middels deze technieken een eerste stap te nemen richting het vinden van bewijs voor causaliteit met betrekking tot transcriptionele regulatie. Het uiteindelijke doel is om hierdoor verder te komen dan het simpelweg duiden van correlaties, en in plaats daarvan causale hypothesen te kunnen opstellen over transcriptionele netwerken.

We beginnen in **hoofdstuk 2** met een methodologische bijdrage aan het zoeken naar verbanden tussen genoombrede patronen van genetische varianten en CpG-dinucleotiden die bij elkaar in de buurt op het genoom liggen, methylatie-QTLs

genoemd, waarbij we gebruik maken van een kleine openbare dataset. We laten zien dat een veel gebruikte multiple testing-strategie een hoog aantal CpG-dinucleotiden foutief aanduidt als beïnvloed door lokale (*cis*) genetische variatie, en ontwikkelen een nieuwe methode die dit voorkomt. Verder concluderen we dat *cis*-meQTLs nog lokaler zijn dan voorheen gedacht, en voor het vinden lokale effecten doorgaans niet verder gekeken hoeft te worden dan 50kb.

In **hoofdstuk 3** maken we gebruik van een grotere dataset van 3.841 Nederlandse individuen om de effecten van 6.111 geselecteerde genetische varianten te relateren aan alle verder weggelegen CpG-dinucleotiden (*trans*). De varianten werden geselecteerd omdat zij in genoombrede associatestudies (GWAS) geassocieerd bleken met één of meerdere ziekten of andere complexe fenotypen. Van de 6.111 varianten kunnen we er 1.907 relateren aan meerdere verder weg gelegen CpG-dinucleotiden. Veel van deze varianten bleken daarbij ook de expressie van nabijgelegen transcriptiefactoren te beinvloeden en de CpG-dinucleotiden die in een bindingssite van de desbetreffende transcriptiefactor liggen. Dit leidt tot onze eerste hypothese: genetische varianten brengen veranderingen teweeg in de expressieniveaus van verder weggelegen genen door de expressieniveaus van nabijgelegen transcriptiefactoren te beïnvloeden. Tot slot stellen we dat een derde van alle onderzochte CpG-dinucleotiden onder invloed staat van nabijgelegen (*cis*) genetische varianten, veel meer dan aanvankelijk gedacht.

In **hoofdstuk 4** onderzoeken we welke diverse rollen epigenetische regulatie speelt bij X-chromosomale inactivatie (XCI), het proces waarbij één van de twee X-chromosomen in vrouwelijke zoogdieren wordt "uitgezet". We veronderstellen dat genetische varianten die alleen in vrouwen de X-chromosomale methylatie beïnvloeden betrokken moeten zijn bij DNA methylatie en XCI. Een drietal van zulke varianten worden geïdentificeerd en gerepliceerd, en zijn dus vermoedelijk betrokken bij XCI. De aangedane CpG-dinucleotiden liggen voornamelijk in de buurt van X-chromosomale genen die veelal ontsnappen aan XCI, wat suggereert dat er een genetische basis is voor dit verschijnsel. Vervolgens onderzoeken we de effecten van de genetische varianten op de expressieniveaus van nabijgelegen genen en wijzen verschillende genen toe aan iedere variant die derhalve gedacht worden verantwoordelijk te zijn voor de veranderingen op het X-chromosoom. Twee van de drie aangewezen genen zijn nog niet eerder geïmpliceerd in XCI en kunnen dus nieuwe aanwijzingen zijn voor het reguleren van XCI via DNA methylatie of aanverwante epigenomische veranderingen.

Tot slot proberen we in **hoofdstuk 5** uitspraken te doen over welke genen veranderingen teweegbrengen in de expressieniveaus van andere genen. Hiermee gaan we met het gebruik van een gemodificeerde Mendelian Randomization-analyse voorbij aan de QTL mapping, hoewel we nog steeds genetische variatie als causaal anker gebruiken om deze hypotheses te kunnen opstellen. Hierbij proberen we de correlaties tussen de verschillende genetische instrumenten, alsook pleiotropische effecten tegen te gaan om zo één gen aan te kunnen wijzen als causale driver. Net als in **hoofdstuk 3** blijkt ook hier dat transcriptiefactoren vaker dan verwacht verantwoordelijk lijken voor veranderde expressieniveaus van andere *in cis* en *trans* gelegen genen. De resulterende catalogus van

gen-geninteracties leverden reeds nieuwe biologische inzichten op en zouden daarnaast de basis kunnen vormen voor vervolgonderzoek omtrent de causale drivers.

Conclusie

Vele ziekten worden veroorzaakt door een verstoring van transcriptionele regulatie. Bij het ontstaan van veelvoorkomende, complexe aandoeningen zijn vaak grote aantallen genen betrokken. Het is van groot belang om de transcriptionele regulatie tussen genen onderling te onderzoeken en met name waar het genen betreft waarvan een relatie met ziekte al is aangetoond. Door verschillende beperkingen is het echter lastig om causale relaties te leggen tussen bijvoorbeeld de expressieniveaus van verschillende genen. Tezamen zijn de verschillende hoofdstukken in dit proefschrift voorbeelden van hoe genetische variatie gebruikt kan worden om met observationele data toch uitspraken te kunnen doen over deze oorzakelijke verbanden. De resultaten uit dit proefschrift helpen hierbij door een beter begrip van transcriptionele (dis)regulatie te geven, terwijl de gebruikte methoden in het algemeen gebruikt kunnen worden om ook met andere type databronnen soortgelijke analyses uit te voeren.

PUBLICATIONS

R. Luijk, J. J. Goeman, E. P. Slagboom, B. T. Heijmans, and E. W. van Zwet. An alternative approach to multiple testing for methylation QTL mapping reduces the proportion of falsely identified CpGs. *Bioinformatics*, **31**(3):340-345 (2015)

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. 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. Stehouwer, C.J. van der Kallen, C.G. Schalkwijk, L.H. van den Berg, E.W. van Zwet, H. Mei, Y. Li, M. Lemire, T.J. Hudson, BIOS Consortium, P.E. Slagboom, C. Wijmenga, J.H. Veldink, M.M. van Greevenbroek, C.M. van Duijn, D.I. Boomsma, A. Isaacs, R. Jansen, J.B. van Meurs, P.A.C. 't Hoen, L. Franke, B.T. Heijmans. Disease variants alter transcription factor levels and methylation levels of their binding sites. *Nature Genetics*, **49**(1):131-138 (2017)

R. Luijk, H. Wu, C.K. Ward-Caviness, E. Hannon, E. Carnero-Montoro, J.L. Min, P. Mandaviya, M. Müller-Nurasyid, H. Mei, S.M. van der Maare, BIOS Consortium, C. Relton, J. Mill, M. Waldenberger, J.T. Bell, R. Jansen, A. Zhernakova, L. Franke, P.A.C. 't Hoen, D.I. Boomsma, C.M. van Duijn, M.M.J. van Greevenbroek, J.H. Veldink, C. Wijmenga, J. van Meurs, L. Daxinger, P.E. Slagboom, E.W. van Zwet, B.T. Heijmans. Autosomal genetic variation is associated with DNA methylation in regions variably escaping X-chromosome inactivation. *Nature Communications*, **9**(1) (2018)

R. Luijk, K.F. Dekkers, M. van Iterson, W. Arindrarto, A. Claringbould, P. Hop, BIOS Consortium, D.I. Boomsma, C.M. van Duijn, M.M. 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, E.W. van Zwet. Genome-wide identification of directed gene networks using large-scale population genomics data, *Nature Communications*, **9**(1) (2018)

H.W. van Steenbergen, **R. Luijk**, R. Shoemaker, B.T. Heijmans, T.W. Huizinga, A.H. van der Helm-van Mil. Differential methylation within the major histocompatibility complex region in rheumatoid arthritis: a replication study, *Rheumatology*, **53**(12), 2317-2318 (2014)

M. van Iterson, E.W. Tobi, R.C. Slieker, W. den Hollander, **R. Luijk**, P.E. Slagboom, B.T. Heijmans. MethylAid: visual and interactive quality control of large Illumina 450k datasets, *Bioinformatics*, **30**(23), 3435-3437 (2014)

R.C. Slieker, M. van Iterson, **R. Luijk**, M. Beekman, D.V. Zhernakova, M.H. Moed, H. Mei, M. van Galen, P. Deelen, M. Bonder, A. Zhernakova, A.G. Uitterlinden, E.F. Tigchelaar, C.D.A. Stehouwer, C.G. Schalkwijk, C.J.H. van der Kallen, A. Hofman, D. van Heemst, E.J. de Geus, J. van Dongen, J. Deelen, L.H. van den Berg, J. van Meurs, R. Jansen, P.A. C. 't Hoen, L. Franke, C. Wijmenga, J.H. Veldink, M.A. Swertz, M.M.J. van Greevenbroek, C.M. van Duijn, D.I. Boomsma, BIOS consortium, P.E. Slagboom, B.T. Heijmans. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms, *Genome Biology*, **17**(191) (2016)

E.W. Tobi, R.C. Slieker, **R. Luijk**, K.F. Dekkers, A.D. Stein, K.M. Xu, BIOS Consortium, P.E. Slagboom, E.W. van Zwet, L.H. Lumey, B.T. Heijmans. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood, *Science Advances*, **4**(1) (2018)

M.A. Siemelink, S.W. van der Laan, S. Haitjema, I.D. van Koeverden, J. Schaap, M. Wesseling, S.C.A. de Jager, M. Mokry, M. van Iterson, K.F. Dekkers, **R. Luijk**, H. Foroughi Asl, T. Michoel, J.L.M. Björkegren, E. Aavik, S. Ylä-Herttuala, G.J. de Borst, F.W. Asselbergs,, H. el Azzouzi, H.M. den Ruijter, B.T. Heijmans, G. Pasterkamp. Smoking is associated to DNA methylation in atherosclerotic carotid lesions, *Circulation: Genomic and Precision Medicine* (2018)

J.L. Min, G. Hemani, E. Hannon, K.F. Dekkers, J. Castillo-Fernandez, **R. Luijk**, E. Carnero-Montoro *et al.*. Genomic and phenomic insights from an atlas of genetic effects on DNA methylation. Manuscript submitted for publication.

* Contributed equally

CURRICULUM VITÆ

René Luijk werd geboren op 6 juli 1988 te Leiden, Zuid-Holland, maar groeide op in Sassenheim, Zuid-Holland, waar hij in 2005 de HAVO afrondde. Een propedeuse van de lerarenopleiding Engels van de Hogeschool van Amsterdam gaf toegang tot de bachelor Psychologie aan de Universiteit Leiden, welke hij in 2009 afrondde.

Hij vervolgde zijn opleiding met de master Statistical Science, ook in Leiden. Deze sluit hij af met zijn scriptie The group lasso in the proportional hazards model with an application to multiply imputed high-dimensional data, onder supervisie van Prof. Dr. Jelle J. Goeman en Prof. Dr. Hein Putter van de afdeling Medische Statistiek & Bio-informatica.

Eveneens binnen het LUMC begon hij zijn promotieonderzoek, zowel binnen de sectie Medische Statistiek als de sectie Moleculaire Epidemiologie, beide onderdeel van de afdeling Biomedical Data Sciences, ditmaal onder dagelijkse supervisie van Dr. Bas T. Heijmans en Dr. Erik W. van Zwet, en algemene begeleiding van promotor Prof. Dr. P. Eline Slagboom. Dit onderzoek richtte zich op de ontwikkeling en toepassing van methoden voor het ontrafelen van de genetische grondslag van transcriptionele regulatie.

René vervolgde hierna buiten de wetenschap zijn carrière als Data Scientist in het digitale domein.

NAWOORD

Volgens het welbekende cliché gaat het niet om de bestemming, maar om de reis ernaartoe. Mijns inziens zijn de mensen die men tijdens deze reis tegenkomt het belangrijkste onderdeel daarvan.

Allereerst mijn ouders, die mij de vrijheid hebben gegeven om mijn doelen en dromen na te jagen. Jullie hebben mij daarin nooit gepusht en altijd onvoorwaardelijk gesteund, en waar nodig een hart onder de riem gestoken. De aanmoedigingen om door te zetten, ook al heb ik regelmatig aan mijn eigen kunnen getwijfeld, zijn hierbij van onschatbaar belang geweest.

Binnen het LUMC heb ik veel verschillende mensen mogen ontmoeten die mij ieder op hun manier verder hebben geholpen. Bas en Erik, jullie hebben het beste in mij boven gebracht met jullie kritische blik, en scherp oog voor zelfs de kleinste details. Eline, jouw visie op de communicatie van wetenschap en het kunnen plaatsen van onderzoek in een groter geheel heeft mij veel geleerd over het communiceren van resultaten voor een breder publiek dan het veld waarin je werkt. Jelle, ook de discussies met jou gedurende deze jaren hebben hier sterk aan bijgedragen.

De collega's van MolEpi. Met vele verschillende expertises binnen één afdeling is er altijd wel iemand die een helpende hand kan bieden. Dit, gepaard met een immer gezellige atmosfeer heeft altijd een prettige werkomgeving opgeleverd.

Roderick, bedankt dat je mij wegwijs hebt gemaakt binnen de voor mij nieuwe wereld van epigenetica en BioConductor. Dit is een zeer belangrijke start gebleken voor de rest van mijn tijd in het LUMC. Koen, ook jij hebt een belangrijk deel gehad in mijn inwijding in de voor mij veelal onbekende nieuwe terminologie. De rest van de vaste epigenetica-club - bestaande uit Elmar, Maarten - hebben mij altijd uitgedaagd beter werk te leveren.

Geiten van de R&S 102, al tijdens onze tijd op de Rijn- en Schiekade hebben alle gezellige borrels en tripjes de nodige afleiding en steun geboden voor het studerende leven. Dit is gelukkig niet veranderd, en is dan ook vaak een welkome afwisseling geweest op het promoverende leven.

Over time, the collective known as the Ballers has changed in its composition, but never in its level of gezelligheid. Although you differ quite a bit personality-wise, together you're an impressively intelligent collective, making it ever interesting and fun to hang out with you. Through many discussions on all sorts of topics you have given me new perspectives on

the world, and have truly enriched my life. Thank you for all the late-night gezelligheid, even though it should be clear by now I'm not much of an evening person.

Alexander L. DeSouza, you seem to be in a league of your own. You continue to show me what passion, a drive to excel, and a seemingly relentless work ethic can do when chasing your goals. Thank you for being here to support me during my defense as my paranymph, and supplying me with trivia you read in your favorite news outlet.

Johannes Everardus Wilhelmus Jong, al zo'n 22 jaar ben je een belangrijk deel in mijn leven. Gedurende deze tijd heb ik altijd veel lol en steun ontvangen die het studerende leven, en ook de afgelopen jaren, makkelijk hebben gemaakt. Je bent er bij alle hoogte- en dieptepunten bijgeweest, dus kon je ook vandaag niet ontbreken als mijn paranimf.

Na kraju, ali ne manje važno, moja najveća ljubav Gorana. Posjeduješ strpljivost koja se ponekad čini neograničena. Oduvijek si mi podržala tijekom ovog - da budem iskren - (pre)dugog poduhvata, čak iako vjerojatno nije lako živjeti sa mnom. Nikad neću moći ti reći koliko si mi važna. Samo mogu ti reći *hvala*, i radujem se biti zajedno za ostatak našeg balkanskog života.