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

Luijk, R.

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

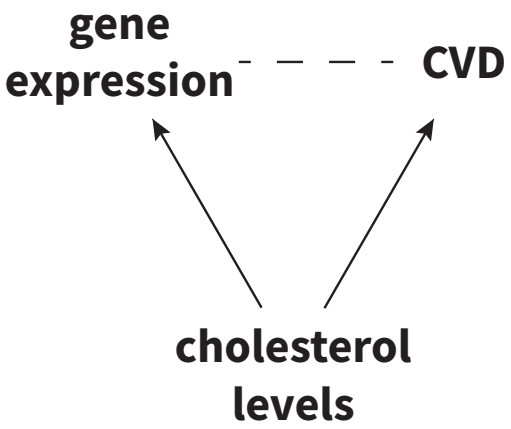

Figure 1.1: Depiction of the relationship between gene expression levels, a specific cardiovascular disease (CVD), and possible confounders, such as cholesterol levels. A correlation between gene expression and the CVD (dashed line) may lead to the conclusion of a possible causal relation between the 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

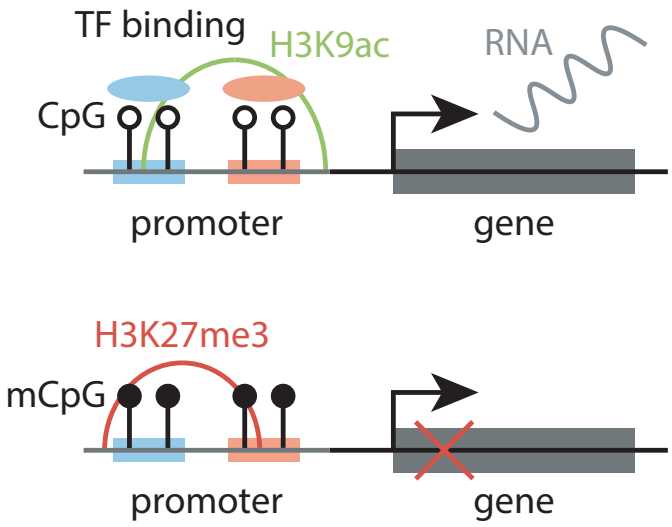

Figure 1.2: Simplified schematic overview of the interplay between epigenomic marks, transcription factor binding, and transcription. In the top panel, unmethylated CpGs and active histone marks (histone mark H3K9ac, in green) typically allow for the binding of transcription factors on their respective 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

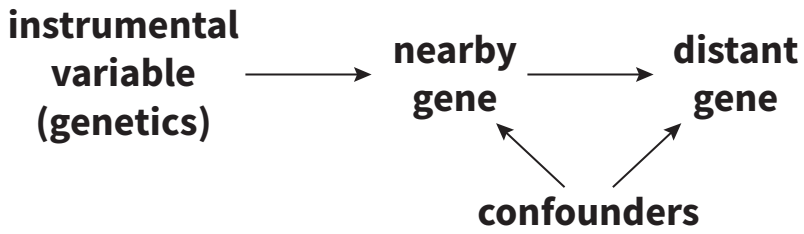

Figure 1.3: Underlying principle of Mendelian Randomization in genome research. A relationship between two genes (nearby gene, distant gene) is confounded by several factors. An instrumental variable is used as a way to manipulate the expression levels of a nearby gene, generating a proxy for its expression 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.