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

Luijk, R.

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.