5 Conclusions

In this thesis I have taken two distinct approaches to using molecular QTLs to understand complex trait genetics. In Chapter 2, I generated a map of eQTLs and caQTLs in iPSC-derived sensory neurons, and examined the overlap between these QTLs and the GWAS catalog. In Chapters 3 and 4, I used a large-scale public eQTL dataset to build the PRF scores model, which provides genome-wide scores of cell type-specific regulatory potential that can be used for fine-mapping at GWAS loci. These studies have focused on the goal of identifying causal variants and genes for complex traits, but have led to different insights into the challenges and opportunities for doing so. Many of my key results relate to how these model systems and tools can be effectively used or improved in the future.

5.1 Mapping QTLs and causal alleles in iPSC-derived cells

IPSC-derived cells are useful model systems for a number of reasons: they provide a renewable supply of specific cell types, thereby allowing multiple molecular assays to be performed; they are genetically matched to a specific donor; and they can be genetically engineered with CRISPR-Cas9 to investigate the causal effects of individual alleles. A challenge to their use is that the differentiated cells often appear immature, and so may not fully recapitulate the *in vivo* cells they are meant to model. Through a detailed analysis of the transcriptome in iPSC-derived sensory neurons, I encountered a second challenge: a fraction of differentiated cells appeared to arrive at a different cell state, being clearly non-neuronal. This heterogeneity, which varied across differentiations, presented a problem for QTL mapping, since it introduced additional, non-genetic variability in gene expression. More concerning was that variability was especially high for genes relating to neuronal differentiation and function. I explored both a standard linear model for QTL mapping and a model that incorporates allele-specific information. Because allele-specific signal is internally controlled, this model improved power across all genes, but especially so for highly variable genes.

To explore potential sources of differentiation variability, I compared metadata and transcriptomes between iPSCs and the sensory neurons derived from them. This led to the interesting observation that iPSCs grown in E8 medium differentiated to neurons more efficiently than iPSCs grown on feeder cells. Differential gene expression between these iPSC lines suggested that the cell culture differences impacted expression of pathways involved in suppressing neuronal differentiation, such as Wnt and TGF-β signalling. To a

certain extent, the iPSC culture conditions can be considered as an extended part of the differentiation protocol for iPSC-derived cells. Because iPSCs are now available from many cell banks, this suggests that in future studies researchers should examine carefully how factors relating to the iPSC lines used may influence the results. For example, it is not sufficient that treatment and control iPSC lines be isogenic - they should also have identical culture conditions for the results in differentiated cells to be comparable.

Epigenetic memory is another factor that may have contributed to heterogeneity across sensory neuron differentiations. The HIPSCI cell lines used were reprogrammed from fibroblasts derived from skin punch biopsies. Although all iPSC lines were pluripotent, as determined by differentiation to cell types of the three embryonic germ layers, some iPSC lines may have retained epigenetic marks that facilitated differentiation to a fibroblast-like cell fate, and the degree of epigenetic memory could differ across cell lines. Notably, the contaminating cells in sensory neuron cultures had characteristics of fibroblast gene expression, yet also had similarity with DRG expression. It is possible that, because differentiation timelines were kept short to use lab resources effectively, the cells simply did not have sufficient time to mature. Another possibility, suggested by Trapnell and colleagues, is that during directed differentiation, some cells fail to choose the desired differentiation path and thereby arrive at aberrant “dead end” cell fates (Cacchiarelli et al. 2017). The possibility of epigenetic memory has led some to suggest that iPSCs should have extensive epigenomic screening before being used for genome editing or differentiation (Grzybek et al. 2017). However, there remains debate on whether epigenetic memory in iPSCs is an important or a minor factor relative to other influences (Burrows et al. 2016). Since all HIPSCI lines had assays of global DNA methylation, it would be possible to investigate whether epigenetic marks at regulators of fibroblast identity are maintained in the iPSCs, and whether these correlate with sensory neuron differentiation outcomes.

One solution to the problem of heterogeneity in iPSC-derived cells is to improve differentiation protocols. This is not straightforward, because many permutations are possible for the growth factors and inhibitors used, as well as for the timing and duration of their addition. Single-cell sequencing at specific time points during differentiation is likely to be highly informative, as it can identify factors that appear to be responsible for decisions at cell fate branch points. Indeed, this has been done at single time points to identify regulators of differentiation to myeloid and lymphoid lineages (Papalexi and Satija 2017), and at multiple time points for reprogramming fibroblasts to myotubes (Cacchiarelli et al. 2017). Another solution to differentiation heterogeneity is to use automated marker-based sorting of the iPSC-derived cells to enrich for the desired cell type. However, if the contaminating cells

have appreciable gene expression similarity to the target cell type, as we found for some of the single fibroblast-like cells, this approach will at best deplete the population of contaminating cells, but will not eliminate them. In addition, suitable markers are only available for some cell types.

Turning more directly to single-cell sequencing approaches could sidestep some of the problems of differentiation heterogeneity for detecting genetic effects. For example, cells from multiple donors can be pooled and differentiated together, and following scRNA-seq, each cell can be traced back to its donor based on the pattern of common variants in expressed transcripts. Pooling should greatly improve power, as it controls for variability due to differentiation. Because pooling also reduces the amount of cell culture work needed, it could also enable larger sample sizes for the same cost. Large-scale iPSC banks will be essential resources in such efforts.

An original goal of the sensory neurons study was to identify genetic variants and genes that influence sensory neuron electrophysiology, and potentially chronic pain. As we discovered, electrophysiological profiles in iPSC-derived sensory neurons were highly variable across single cells, and with our relatively small sample size we did not have sufficient power to detect associations between gene expression and electrophysiological phenotypes. An alternative approach reported recently, PatchSeq, controlled for variability across neuronal cells by first measuring electrophysiological properties by patch clamping, followed by sequencing the same single cells (Bardy et al. 2016). By linking neuron functional properties with gene expression, the authors discovered genes that distinguish highly functional mature neurons from those with less mature functional states.

Using the large set of sensory neuron eQTLs, sQTLs, and caQTLs that we identified, we used LD to detect overlaps with GWAS catalog associations. This revealed a handful of “positive control” overlaps, such as between an SNCA eQTL and a Parkinson’s disease association, as well as novel overlaps that point to specific genes for some GWAS loci. It is not clear that sensory neurons are the relevant cell type for these traits; however, the presence of certain clear overlaps, such as the *TNFRSF1A* association with multiple sclerosis, indicates that some trait-relevant gene regulatory mechanisms are shared across neurons, or perhaps shared more broadly across cell types. Although we detected a large number of sensory neuron caQTLs, we found fewer clear cases of caQTL overlap with GWAS traits where neurons seem likely to be relevant. A possible explanation is that chromatin QTLs are common, and that only a fraction actually influence gene expression

and complex traits. It may also be that, because we only had 31 ATAC-seq samples, we had low power to detect many true genetic effects on chromatin.

It must be kept in mind that many of the overlaps we reported may not represent true colocalisation between causal variants for the molecular and GWAS traits. Because molecular QTLs are relatively common across the genome, many overlaps based on LD are expected by chance. More sophisticated statistical colocalisation methods, such as coloc (Giambartolomei et al. 2014), can use summary statistics to distinguish whether a given overlap is more consistent with shared or distinct causal variants. Applying coloc would likely have strengthened the evidence for overlap of some causal variants, while eliminating others. Another caveat is that we did not experimentally validate any of our novel QTL-GWAS overlaps.

The results from our sensory neurons study, as well as those from the NextGen consortium, indicate some of the future challenges in identification of causal variants for complex traits. Based on our ability to detect eQTLs across genes with different levels of expression variability, we estimated that 40 to 80 iPSC-derived cell lines would be needed to detect a difference between alleles for a single gene regulatory variant of moderate effect size. This is consistent with NextGen consortium results: Warren et al. performed 136 differentiations to hepatocyte-like cells across 68 cell lines from individuals recruited based on their genotypes, and marginally detected (p=0.04) an effect of rs12740374 on *SORT1* expression (Warren et al. 2017). Of note, they reported a median differentiation efficiency to HLCs of just 10%, and their expression assays were on sorted HLC populations. When applying CRISPR-Cas9 in one parental cell line to generate deletions covering rs12740374, they used 38 differentiations to show a difference in *SORT1* expression between wild-type and deletion cell lines. The smaller number of differentiations needed in the gene-editing experiment may reflect the fact that a single cell line was used, and so variability due to both genetic background and iPSC generation were eliminated. However, it is also possible that the deletion alleles had a larger effect size on average than the naturally-occurring SNP alleles.

In sum, it is clear that determining causal alleles for human complex traits remains difficult, due to the modest effect sizes of the alleles, and to the variable efficiency of differentiation and limited maturity of derived cells. Allele-specific analyses, when possible, are one means of controlling for sample-to-sample expression variability. A number of new approaches based on single-cell RNA-seq offer even more powerful solutions. Sequencing single cells at time points along the differentiation course will elucidate genes that determine different cell

fates; pooling cells from multiple individuals will control for non-genetic variability; and comparing cell functional states to single-cell transcriptomes will enable linking cell phenotypes with gene expression.

5.2 Towards better predictive models of gene regulation

Molecular QTLs are one of the most useful types of evidence of variant functionality, since they directly link genetic variation with an intermediate phenotype: eQTLs link regulatory regions to genes, and chromatin accessibility QTLs can aid fine-mapping by providing a strong prior on the locations of putative causal variants. However, QTLs also have limitations. As with GWAS, identifying the causal variant for an association is extremely difficult. In addition, because gene regulation can be cell type- and context-specific, the absence of an observed effect for a variant could simply mean that the relevant condition has not been tested. Efforts are underway to produce QTL maps in additional cell types and conditions, but the combinatorial space to explore is vast. Lastly, QTLs can only be detected at genetic variants of appreciable allele frequency; yet, we may be interested in the regulatory effects of rare or *de novo* variants. For these reasons, molecular QTL maps provide only one piece of the puzzle to identify causal alleles for complex traits.

Because the experiments necessary to demonstrate the molecular effects of a single causal variant are so challenging, it is important to first bring all available data to bear on the problem to identify strong candidate variants. Yet, so many types of genomic and epigenomic data have now been generated across various cell types that manual, heuristic approaches to identify relevant data are no longer sufficient. We developed PRF scores to simplify this task, bringing together a large number of annotations in a rigorous framework that assigns transparent, cell type-specific scores for each variant. In developing PRF scores, we found that the quantitative values of annotations, such as histone modifications and chromatin accessibility, can improve predictions of eQTL locations. We found that imputed data from the Roadmap Epigenomics project largely improves predictions, while having the advantage of being available across all of the tissues profiled. We also found that some annotations have different levels of informativeness based on a variant's distance to the gene in question. With 38 annotations in an integrated model, PRF scores showed slightly better performance than CADD and GWAVA in classifying likely causal eQTL variants in GTEx tissues.

A key advantage of PRF scores is that they can be used to determine prior probabilities of influencing gene expression for a set of variants. We used this feature to apply PRF scores to fine-map associations across six GWAS. There were a number of clear examples where PRF scores identified a good candidate causal variant for the association. However, these still represented a small minority of all associations. More often, PRF scores moderately changed posterior probabilities across credible set variants, but even with a breakdown of variant scores into annotation contributions, no clear molecular mechanism was indicated. Although the credible set of variants was reduced in size on average with PRF score fine-mapping, we found that this also occurred with scores randomized across variants. This suggests that reduction in credible set size is not a reliable indicator of fine-mapping performance.

PRF scores model one type of variant functionality - the likelihood that a variant influences expression of a particular nearby gene. The fact that PRF scores are gene-specific can be both a strength and a weakness. A strength is that, when a variant is prioritized due to having a high PRF score, it implicates a specific gene as the one most likely to be associated. A weakness is the difficulty of handling the multitude of scores for each variant; as a simplification for fine-mapping, we used a variant's maximum PRF score across genes. With up to tens of genes within 1 Mb of a variant, and over a hundred Roadmap epigenomes, computing PRF scores is also computationally intensive.

In principle, the different gene-PRF scores for a variant could be used to determine the relative probability that the variant influences expression of each of those genes. However, the primary annotation in our model for distinguishing the target gene is TSS distance. As a result, PRF scores provide little information on potential distal regulatory interactions, and in most cases, PRF score fine-mapping did not change the gene most strongly implicated for GWAS associations. This is at odds with recent reports suggesting that most disease-associated enhancers contact genes beyond the nearest one in the genome (Mumbach et al. 2017). It is not clear to what extent this represents a true discrepancy between the regulatory architecture of eQTLs and GWAS associations, because the target genes for eQTLs are known, whereas in most cases those for GWAS are not. Once a larger number of causal genes are identified for GWAS associations, likely in the near future, we will be better able to answer this question.

New data sources may improve our ability to relate regulatory regions to genes. Promoter-capture Hi-C is being used across many cell types to identify DNA regions that interact with gene promoters. While the presence of an interaction is weaker evidence of a causal

relationship than if an eQTL were present, these data are easier to acquire, are not limited by LD, and may be especially informative for distal interactions. Another information source tying regulatory regions to genes comes from exploiting correlations across cell types between gene expression and epigenetic activity at regulatory regions (Hait et al. 2017), with the assumption that correlated activity patterns indicate an interaction. A third source of evidence comes when both non-coding and rare coding variant associations for the same phenotype occur at a locus; this implies that the non-coding association likely acts through the same gene as the coding variants. To the extent that these data identify causal distal regulatory interactions, when integrated into a PRF score model, the gene-specific nature of the scores may become a greater benefit.

There remains considerable potential for improving a PRF score model by including annotations with nucleotide-level resolution. A number of machine-learning methods have been developed to relate DNA sequence with the likelihood of influencing a particular molecular phenotype, usually chromatin accessibility or TF binding. Massively parallel reporter assays (MPRAs) with nucleotide resolution are also being developed. A recent method, called HiDRA, uses dense tiling of candidate sequences across regions of open chromatin in a reporter assay, and when combined with machine-learning across overlapping fragments, suggests critical nucleotides for regulatory element function (Wang et al. 2017). Because investigators are unlikely to apply a large range of methods to prioritize variants in their own datasets, these methods are excellent candidates as input to an integrative model.

5.3 The future of fine-mapping

Fine-mapping methods that rely on statistical signal only, such as FINEMAP (Benner et al. 2016), are valuable because they can rapidly assess different potential configurations of multiple causal variants, and are unbiased by our imperfect knowledge of genomic annotations and function. However, they have two main drawbacks: they cannot resolve causal variants at regions of very high LD, and their results depend greatly on having a population LD reference panel closely matched in ancestry with the study population. They also provide no input on potential causal mechanisms leading to the association.

Even though PRF scores show some utility for fine-mapping, I think that a model integrating multiple data sources can do much better. The goal of fine-mapping is not simply to identify the causal variant for an association, but to identify a causal mechanism. For this, we need allele-specific, nucleotide-resolution predictions indicating what molecular events

(transcription factor binding, RNA structure, splicing, etc.) are likely to be altered by a given variant, in which direction, and with effect on which genes. With this granularity and specificity of information, the predictions would be more likely to bring specific variants to the fore at association loci. They would also provide researchers with the information needed to evaluate how plausible each variant's potential causal mechanism is for the trait in question, while making clear the experiments necessary to validate them.

Developing such a model will be challenging for a number of reasons. The first is the diversity of molecular mechanisms that influence complex traits. Protein-coding variants can alter protein structure, post-translational modifications, interactions, catalytic sites, localisation, splicing, and degradation. Even in the minority of cases where a coding variant is present, the evidence for its causality must be weighed against potential regulatory variants in LD. At a greater number of loci, regulatory variants are implicated. Although identification of likely gene regulatory regions has greatly improved over the past decade, we still have a limited understanding of the "regulatory code". Regulatory variants can influence phenotypes via well-established mechanisms, such as by altering TF binding sites, with downstream effects on promoter interactions and gene expression. However, they can also act in many other ways, such as by affecting transcript stability, splicing, ribosome pausing, expression of non-coding RNAs, DNA methylation, and by changing large chromosomal domains.

Another challenge is that regulatory variants may affect one or more genes, and these may be distal. So far no single assay or information source can identify the target genes of regulatory regions at high resolution genome-wide. Whereas changes to protein structure have an effect across all cell types, changes to gene regulation are more likely to be cell type or context-specific. Linking regulatory regions to genes is likely to be an endeavour where we gradually accumulate evidence from multiple sources over time, including from eQTL studies, Hi-C experiments, correlations across cell types and contexts, and increasingly from gene editing experiments at individual loci.

The next challenge for integrative regulatory predictions and fine-mapping methods is to incorporate supervised training with multiple datasets that cover different regulatory mechanisms. With PRF scores we used steady-state eQTLs as a training dataset to identify annotations that predict gene regulation. Other methods have similarly used single datasets for supervised training, such as chromatin accessibility (D. Lee et al. 2015; J. Zhou and Troyanskaya 2015; Kelley, Snoek, and Rinn 2016), eQTLs (Ioannidis et al. 2017; Y. Li and Kellis 2016), conservation and polymorphisms (Y.-F. Huang, Gulko, and Siepel 2017), or

GWAS summary statistics directly (J. Pickrell 2013; Kichaev et al. 2014; Y. Li and Kellis 2016). Apart from training datasets, new methods will also need to maximize the use of different annotations, which in general are not uniformly available across cell types; this is particularly the case for TF binding. How to appropriately weight the relative importance of different mechanisms, as well as the quality and informativeness of datasets with different coverage across cell types, is a question that has barely been addressed.

A final challenge in developing predictive models of gene regulation is the lack of a gold-standard dataset of causal variants influencing complex traits. Such a dataset would enable accurate evaluation of the performance of different regulatory scores and fine-mapping methods. EQTL datasets identify associated regions but not causal variants. The human gene mutation database (Stenson et al. 2014) contains thousands of protein-coding variants and hundreds of non-coding variants associated with Mendelian phenotypes, but these are not representative of the common variation that influences complex traits. A growing number of variants, originally discovered by GWAS to have trait associations, have been experimentally shown to have molecular effects on plausible candidate genes. The number of such variants is likely to grow more rapidly with the increasing application of CRISPR-Cas9. No database has yet been set up to collect these examples, but the time to do so is ripe.

5.4 Concluding remarks

GWAS have discovered thousands of associations with human complex traits, yet deciphering the causal variants and molecular mechanisms for these associations is challenging. The burgeoning of relevant genomic data has made it an exciting time to investigate the effects of non-coding genetic variants. A number of important questions have answers within reach, at least in part, in the next decade. How many causal variants are there at individual loci and genome-wide for different traits? What fraction of transcription factor binding sites are functional? How frequently do variants affect multiple genes rather than a single gene? What is the prevalence of different molecular mechanisms influencing complex traits? And more broadly, how can the data from large-scale sequencing and genomic assays be used to understand differences in traits between people, and to develop therapies for common diseases? I look forward to new approaches, both experimental and computational, that will shed light on these questions.