Using molecular QTLs to identify cell types and causal variants for complex traits

Jeremy Schwartzentruber

Wellcome Trust Sanger Institute

Clare College

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

October 2017

Declaration of Originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the beginning of each chapter. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution. This dissertation does not exceed the word limit set by the Degree Committee for the Faculty of Biology.

Signature:

Date:

Jeremy Schwartzentruber
October 2017

Abstract

Genetic associations have been discovered for many human complex traits, and yet for most associated loci the causal variants and molecular mechanisms remain unknown. Studies mapping quantitative trait loci (QTLs) for molecular phenotypes, such as gene expression, RNA splicing, and chromatin accessibility, provide rich data that can link variant effects in specific cell types with complex traits. These genetic effects can also now be modeled *in vitro* by differentiating human induced pluripotent stem cells (iPSCs) into specific cell types, including inaccessible cell types such as those of the brain. In this thesis, I explore a range of approaches for using QTLs to identify causal variants and to link these with molecular functions and complex traits.

In Chapter 2, I describe QTL mapping in 123 sensory neuronal cell lines differentiated from human iPSCs. I observed that gene expression was highly variable across iPSC-derived neuronal cultures in specific gene categories, and that a portion of this variability was explained by commonly used iPSC culture conditions, which influenced differentiation efficiency. A number of QTLs overlapped with common disease associations; however, using simulations I showed that identifying causal regulatory variants with a recall-by-genotype approach in iPSC-derived neurons is likely to require large sample sizes, even for variants with moderately large effect sizes.

In Chapter 3, I developed a computational model that uses publicly available gene expression QTL data, along with molecular annotations, to generate cell type-specific probability of regulatory function (PRF) scores for each variant. I found that predictive power was improved when the model was modified to use the quantitative value of annotations. PRF scores outperformed other genome-wide scores, including CADD and GWAVA, in identifying likely causal eQTL variants.

In Chapter 4, I used PRF scores to identify relevant cell types and to fine map potential causal variants using summary association statistics in six complex traits. By examining individual loci in detail, I showed how the enrichments contributing to a high PRF score are transparent, which can help to distinguish plausible causal variant predictions from model misspecification.

Acknowledgements

I have been fortunate to spend the past four years at the Wellcome Genome Campus, a place with many inspired scientists whose door is always open. Firstly, I would like to thank my supervisor Dan Gaffney, for allowing me to explore different project and ideas; for being involved every step of the way; for his commitment to think carefully about each claim we make; and for his focus on clearly explaining *why*. I am also grateful to members of our group - Kaur, Natsuhiko, Angela - who so readily shared their thoughts and expertise with me, as well as their code. I benefited greatly from the collaborative and supportive environment at the Sanger Institute, including discussions with Jeff Barrett, Leo Parts, Carl Anderson, Annabel Smith, and Christina Hedberg-Delouka, as well as Chris Wallace in Cambridge. A big thank you Alex Gutteridge, for his cheerfulness, clear code and helpfulness in many analyses.

Thanks also to the other PhD students who always included me and played so well with Maia (junior PhD13 member who grew up from 0 to 4 years old on campus). I wish you all the best Katie, Nicola, Sumana, John, Masha, Li Meng, Veli, Liliana, Alice!

I am grateful to Dr. Jenny Clapham, who was one of the few doctors that seemed to take the chronic pain that I suddenly developed seriously. It was a dark ~1.5 years of my life, and a doctor who believes what you say and trusts you can make such a difference, even when there are no clear answers.

Finally, thank you to my two little girls for the many delightful moments. To Neeltje for supporting me and our girls through both delightful and difficult moments, and for shouldering the lack of sleep! To Kees and Michalien for being so quick to help, so steadfast, and calm. And to my parents, for always listening to me and supporting me no matter what.

Abbreviations and key terms

ATAC	Assay for transposase-accessible chromatin
AUC	Area under the curve (used for ROC or other classifier metrics)
BAM	Binary sequence alignment file format
BF	Bayes factor
CADD	Method predicting deleteriousness of coding and non-coding variants
CAGE	Cap analysis of gene expression
caQTL	Chromatin accessibility QTL
CD	Crohn's disease
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CNS	Central nervous system
CQN	Conditional quantile normalization for gene expression
CV	Coefficient of variation (standard deviation / mean)
DNase I	Enzyme that preferentially cuts DNA at open chromatin
DRG	Dorsal root ganglion
E8	Essential 8 iPSC culture medium
eGene	eQTL gene
eQTL	Gene expression QTL
ESC	Embryonic stem cell
FANTOM	Dataset of TSS usage based on CAGE
FDR	False discovery rate
fgwas	A fine-mapping method incorporating functional genomic annotations
FPKM	Fragments per kilobase of exon per million mapped reads
GERP	Genomic evolutionary rate profiling, a sequence conservation metric
GTEx	Genome-tissue expression project
GWAS	Genome-wide association study
GWAVA	Method predicting deleteriousness of non-coding variants
HDL	High density lipoprotein
HGMD	Human gene mutation database
HIPSCI	Human induced pluripotent stem cell initiative
IBD	Inflammatory bowel disease
iPSC	Induced pluripotent stem cell
IPSDSN	iPSC-derived sensory neuron
LCL	Lymphoblastoid cell line
LD	Linkage disequilibrium
LDL	Low density lipoprotein
LLK	Log-likelihood (of a model, given certain parameters)
MAF	Minor allele frequency

1 Introduction

A decade ago, large-scale genome-wide association studies (GWAS) conducted as part of the Wellcome Trust Case Control Consortium led to the discovery of 24 genomic loci associated with common human diseases (Wellcome Trust Case Control Consortium 2007). Prior to this, linkage mapping in human pedigrees had been successful in identifying genes and genetic variants leading to Mendelian diseases, but had largely failed for complex traits. Although the amount of phenotypic variation explained by these GWAS associations was small, they provided the first unbiased, genome-wide view of the genetic architecture of complex traits. Since then, GWAS have been done with increasing sample sizes for many human traits, leading to thousands of genomic loci associated with hundreds of traits. However, at the vast majority of these loci the causal variants and molecular mechanisms are uncertain. At present, most GWAS associations only represent leads into a wealth of underlying biology that will require new data and new methods to unravel. If we can do so, there is the promise that they will lead to a new understanding of complex traits, and new treatments for common diseases that together affect a large fraction of the population.

In this chapter, I outline the reasons that determining the causal variants and mechanisms behind GWAS associations is so challenging. I discuss how studies of molecular traits can provide insight into the functionality of different genomic regions, and introduce the reference datasets that many of my analyses are based on. Some human cell types are difficult to access, but differentiating specific cell types from induced pluripotent stem cells (iPSCs) can enable studying molecular traits in these cells *in vitro*. I provide background to the use of iPSC-derived cells as model systems, as well as the challenges to their use. Finally, I review existing methods that use functional genomic data to predict the functionality of genetic variants, and to fine-map causal variants at GWAS loci.

1.1 The challenge of determining mechanisms underlying complex trait genetic associations

1.1.1 Common variants with small effects

A complex trait is one that is not determined by a single locus with a large effect, and can include anthropometric traits such as height, molecular traits such as metabolite levels, or risk for common diseases like cancer or type 2 diabetes. An early observation from GWAS was that across many human traits, the effect sizes of the loci discovered were small. Moreover, even for the most highly powered studies, the fraction of trait heritability explained

by all genome-wide significant loci together was also small, typically below 10%. This came to be termed the problem of “missing heritability” (Manolio et al. 2009). Based on this, some criticized the principle behind GWAS, suggesting that rare variants may explain more heritability than the common variants which GWAS is well-powered to discover (McClellan and King 2010). However, recent work has shown that when all assayed variants are accounted for, more than 30% of heritability can be explained by common variants for many complex traits (Speed et al. 2017), and for one of the most highly powered GWAS, human height, more than 50% is explained by all common variants at current sample sizes (Yang et al. 2015). This is supported more directly by large-scale sequencing, which for type 2 diabetes has shown that low-frequency and rare variants appear to play only a minor role in disease risk (Fuchsberger et al. 2016). It appears that, to understand the genetic contribution to complex traits, unraveling the biology behind common variant GWAS associations is essential.

It is common to refer to “causal variants” for complex traits, but it is not always explicit what this means. For Mendelian diseases the picture is clearer: most such diseases have high penetrance, and a single mutation either occurs *de novo* or segregates within a family along with a clear phenotype. The vast majority of Mendelian diseases with known genetic causes have been explained by mutations in protein-coding genes (Chong et al. 2015). In contrast, GWAS now routinely discover dozens of loci associated with individual complex traits. At each locus, there are usually many variants statistically associated with the trait, and it is assumed that only one or a small number of these variants causally influence the trait. Here, causal means that some molecular mechanism links a particular variant to the trait, and that having a different allele of that variant would alter the quantitative trait or the risk for disease. Because the effects of these loci are small, a given causal variant has only a minor influence on the value of a quantitative trait. Similarly, for common diseases a causal risk variant is neither necessary nor sufficient to cause disease.

Pathway and tissue-specific enrichments of genes at common variant associations have led to new insights into the aetiopathogenesis of many disorders, including ankylosing spondylitis (Evans et al. 2011), schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014), and obesity (Claussnitzer et al. 2015). However, GWAS identify associated variants rather than genes, and these enrichments are only possible by looking broadly at the genes in a window around each association. In other words, both the causal variants and the relevant genes at individual GWAS loci are usually unknown.

Identifying the causal variants for GWAS associations is important because it facilitates experiments to investigate disease mechanisms, which typically must be done in specific cell types (or model organisms) and with specific perturbations. This is increasingly done using genome editing with the CRISPR/Cas9 nuclease to engineer deletions or alter alleles of specific variants, followed by evaluating the molecular effects of the changes. Because these experiments are limited in throughput, it is essential to have a clear hypothesis as to which variant is causal. Genome editing experiments can then elucidate which genes are affected by causal variants, and this in turn can inform therapeutic hypotheses. However, a number of challenges make it difficult to identify causal variants at GWAS loci.

1.1.2 Linkage disequilibrium and genotype imputation

GWAS are based on the principle of "tagging" the majority of common genetic variation in the genome by assaying only a subset of variants, specifically, single nucleotide polymorphisms (SNPs). This cost-effective approach enables a large sample size, which maximizes the power to detect associated loci. Tagging is possible because nearby regions of a chromosome tend to be transmitted together to offspring, which leads to correlation between alleles that is referred to as linkage disequilibrium (LD). The level of LD between two variants is commonly measured by the R-squared (r^2) of their correlation across chromosomes in a population. Over time, recombination of chromosomes between two alleles reduces their pairwise LD. Recombination is not evenly distributed across chromosomes, but tends to occur in hotspots. As a result, the human genome has segments of variable length, typically 10 - 100 kb but sometimes longer, with many alleles in high LD.

Because only tag SNPs are measured, the causal variant for a GWAS association is often not among those tested. This has been changing with the availability of reference panels of genetic variation discovered using whole genome sequencing (WGS), such as the 1000 genomes project (1000 Genomes Project Consortium et al. 2015). Using WGS, the full range of genetic variation in an individual can be discovered, including SNPs, but also insertions or deletions (indels) of various lengths, copy number variants (CNVs), and more complex structural variants. GWAS can leverage these reference data to impute genotypes at variants that are not directly assayed in the study. This enables association testing for all 5 - 10 million common variants present in human populations. However, the LD which makes imputation possible also means that many variants at a locus have similar association statistics.

Imputation brings its own problems: while common variants can generally be well imputed, accuracy is reduced for low-frequency variants, and greatly reduced for rare variants (Howie et al. 2012), due to their lower LD with tag SNPs. Most GWAS to date have been imputed using the 1000 genomes project reference panel, which was based on low-depth (7x) whole genome sequencing of individuals from multiple global populations. Imputation can only be as accurate as the reference panel itself. While SNPs are accurately recovered in low-depth sequencing, sensitivity for detection of short indels is lower, and for different classes of structural variants ranges from 32% - 88% (Sudmant et al. 2015). Reference panels for imputation are improving — by combining together many low-depth WGS studies, the majority from European cohorts, the Haplotype Reference Consortium has created a panel of 64,976 haplotypes at ~39 million SNPs (S. McCarthy et al. 2016). Despite this, it is important to realise that association statistics are influenced by the quality of the genotypes for particular variants. As statistical power increases, the effects of genotyping inaccuracies on association statistics are amplified.

In the future, high-depth WGS may become feasible for large-scale association studies. This will enable better discovery of rare variant associations and will overcome some of the challenges of genotype imputation. However, the difficulty of resolving causal common variants will remain due to broad LD at many associated loci.

1.1.3 Most associated variants are non-coding

At the majority of loci discovered by GWAS for complex traits, no variants in protein-coding genes are compelling candidates for explaining the association (Hindorff et al. 2009; H. Huang et al. 2017; Farh et al. 2015). This contrasts sharply with Mendelian diseases, where most causal variants alter protein-coding sequence, and are either *de novo* mutations or are extremely rare. A common hypothesis is that many complex trait associations are driven by changes to gene expression. A key mechanism whereby genetic variants influence gene expression is by altering DNA sequence motifs for transcription factors (TFs) at their binding sites (TFBS), which are commonly found at enhancer, repressor, and promoter elements. One of the pioneering studies demonstrating this showed that a variant 35 kb distal from *SORT1* creates a C/EBP binding site and alters expression of *SORT1* in liver hepatocytes (Musunuru et al. 2010). This in turn alters plasma LDL-C levels and provides a plausible explanation for the GWAS association of this locus with myocardial infarction. Examples are accumulating of complex trait-associated variants which disrupt or create TFBS, and thereby affect gene expression (Praetorius et al. 2013; Guenther et al. 2014; Soldner et al. 2016).

If we could predict which non-coding genetic variants are functional, then we might be able to unravel non-coding GWAS associations. Yet, despite a large body of work investigating the genetic basis of gene regulation, our ability to predict these effects remains poor. Unlike protein-coding variation, where the genetic code maps precisely from nucleotide sequence to amino acids, the rules governing noncoding sequence are more probabilistic and complex. Understanding this “regulatory code” is a key goal in genomics.

Researchers therefore face clear challenges to interpreting non-coding GWAS associations. Many mechanisms for non-coding variants to influence complex traits are possible, including by altering gene splicing (Gregory et al. 2012), the action of noncoding RNAs (Ling et al. 2013), or altering expression of microRNAs or their binding sites in the untranslated regions (UTRs) of genes (Ghanbari et al. 2016). Indeed, multiple noncoding variants can act independently or in concert to affect gene expression (Glubb et al. 2015; Bojesen et al. 2013). In addition, regulatory variants can influence distal genes, and so at each locus many genes are candidates to mediate the association. Finally, because gene regulation can be cell type- and context-specific, it is difficult to know which context is the most relevant for investigating a given trait association.

1.1.4 Non-coding associations may span long distances

GWAS loci are also enriched near genes, but because the mechanisms for these associations are not generally known, it is unclear how often the nearest gene to a GWAS hit is the one mediating the association. In some fraction of cases, the top GWAS variants appear to be in or very near the causal genes. For example, there are cases where a GWAS association occurs in a gene that is the known target of a drug for the same disease, as with cholesterol-lowering statin drugs and the LDL association at *HMGCR* (Kathiresan et al. 2008). Further, enrichment of genes at GWAS loci for known pathways and biological mechanisms has been shown for a number of traits, such as pancreatic islet cell function in type 2 diabetes (Pasquali et al. 2014), and inflammatory signalling pathways in a number of autoimmune disorders (Parkes et al. 2013).

GWAS variants can also regulate distal genes. An early GWAS success was the discovery of a strong association between obesity and variants in introns 1-2 of *FTO*. This gene was initially seen as a strong candidate for regulating body mass, and indeed studies in mice showed that *FTO* knockout led to growth retardation and reduced adipose tissue (Fischer et al. 2009), while *FTO* overexpression increased body and fat mass (Church et al. 2010). However, these studies either removed or duplicated the *FTO* obesity risk region along with

FTO itself, and so did not preclude that the effect was mediated by another gene. Subsequently, it was found that the *FTO* intronic variants form long-range connections with *IRX3*, more than 500 kb distal (Smemo et al. 2014). It was then convincingly shown that the causal obesity risk variant acts through *IRX3* rather than *FTO* (Claussnitzer et al. 2015).

The number of examples of complex trait associations acting via distal genes is growing: the SNP that causes blond hair in Europeans likely acts via a reduction in *KITLG* expression, some 350 kb away (Guenther et al. 2014); a vascular disease association acts through *EDN1*, 600 kb away (Gupta et al. 2017); and a prostate cancer risk variant acts through *SOX9*, 1 Mb away (Zhang et al. 2012). Newly developed methods that integrate gene expression with summary association statistics from GWAS have estimated that around two-thirds of GWAS associations are not mediated by the nearest gene (Zhu et al. 2016; Gusev et al. 2016). These examples illustrate that, even when a plausible gene overlaps a GWAS association, it is not safe to assume that it is causal for the association.

1.1.5 Gene regulation can be cell type- and context-specific

Genetic variants act via molecular pathways to alter higher-level phenotypes. It is intuitive that the effects of such variants will be specific to certain cell types relevant to the phenotype. A primary way to study these effects is by measuring gene expression in specific tissues across multiple individuals, which enables the discovery of loci that influence gene expression, termed expression quantitative trait loci (eQTLs). An early eQTL study examined primary fibroblasts, T cells, and lymphoblastoid cell lines (LCLs), and suggested that 68-70% of regulatory variants were cell type-specific (Dimas et al. 2009). However, the degree of overlap is highly dependent on power, and subsequent studies have demonstrated that the majority of eQTLs are shared (Ding et al. 2010; Nica et al. 2011). The pilot analysis of the genotype-tissue expression project (GTEx), examining eQTLs across 100 - 150 samples in nine tissues, found that more than 50% of eQTLs were shared across all nine tissues, and only 10-30% of eQTLs were tissue-specific (GTEx Consortium 2015). As GTEx increases its sample size and the number of tissues measured, the degree of tissue-specificity is likely to drop even lower. These results suggest that to integrate eQTLs with GWAS results, it may not be necessary to have a perfect match between the eQTL tissue and the “causal” tissue for the trait association.

Despite the widespread sharing of gene regulatory effects across tissues, it would seem odd to investigate trait associations using a cell type with no apparent connection to the phenotype, not least because the results would be hard to interpret. Even if the genetic

effect of a trait-associated variant were detectable in such a cell type, the results might not reflect the downstream molecular mechanisms relevant to the trait. The high estimates of sharing from tissue studies may also reflect the broad mix of cell types present in whole tissues, and so underestimate eQTL specificity at the level of cell types. In the *FTO* and obesity example above, regulatory variants were found to influence *IRX3* expression in pre-adipocytes but not in whole adipose tissue (Claussnitzer et al. 2015), indicating that specificity is possible even across closely related cell types. For this reason, investigators are performing gene expression studies using increasingly specific cell types, such as sorted regulatory T cells implicated in autoimmune disease (Ferraro et al. 2014), specific brain regions associated with psychiatric and neurodegenerative disorders (Ramasamy et al. 2014), and multiple regions of the colon associated with inflammatory bowel disease (Singh et al. 2015).

In addition, some gene regulatory effects responsible for disease associations may only be detectable under specific conditions, such as in response to immune stimulus. Fairfax et al. exposed primary CD14+ human monocytes from 432 individuals to interferon-γ or two durations (2 and 24 hrs) of bacterial lipopolysaccharide (LPS), and found hundreds of context-specific eQTLs dependent on the type or duration of stimulus (Fairfax et al. 2014). Similarly, Lee et al. derived dendritic cells from human peripheral blood monocytes of 534 individuals, and exposed these to either LPS, influenza virus, or the cytokine interferon-β (IFN-β), followed by measuring expression of 415 genes (M. N. Lee et al. 2014). Among the eQTLs they discovered were a number which overlapped with common disease associations and which were only discovered in stimulated cells.

1.2 Genomics of molecular traits

Deeper understanding of molecular traits holds great promise for revealing the mechanisms behind many complex trait associations. A large number of molecular traits are potentially informative, including gene expression and splicing, protein expression, chromatin accessibility, chromosomal conformation, histone modifications, and transcription factor binding. Whereas GWAS for complex traits generally only began to discover loci at genome-wide significance with samples sizes of thousands of individuals, genetic studies of molecular traits routinely discover replicable effects with fewer than one hundred samples. A likely reason for this is that molecular traits are more directly downstream of DNA sequence in the cascade of events influenced by genetics. Also, the technologies and analysis methods differ between GWAS and molecular traits. For many molecular traits, we know where to look in the genome — it is common to statistically test only variants near the gene

or feature, since these are more likely to have an effect. To make appropriate inferences, it is important to understand the opportunities and limitations of these different types of data.

1.2.1 Gene expression

While proteins are key actors in carrying out cellular functions, it is technically difficult to measure protein levels in a precise and high-throughput manner, although some progress is being made towards these ends (Melzer et al. 2008; Stark et al. 2014; Battle et al. 2015). Due to the lower cost and more accessible technology, most studies of gene regulation have quantified steady-state mRNA levels. Early studies measured total expression using arrays of probes specific to one or more exons of each of the approximately 20,000 protein-coding genes in the human genome. Subsequently, the availability RNA sequencing made it possible to measure not just total expression levels, but also alternative splicing, which revealed previously unknown exons and alternative UTRs (J. K. Pickrell et al. 2010).

When gene expression is measured genome-wide across many samples, mapping eQTLs is similar to doing a GWAS for each gene. However, a key difference is that it is common to only test variants within 1 Mb of each gene for association with the gene's expression, since very few variants beyond this distance influence expression. For example, in data from a large-scale study of LCLs (Lappalainen et al. 2013), which is used in Chapter 3 of this thesis, only around 25% of lead eQTL variants are more than 50 kb from the genes they regulate. Since most eQTLs are local to the regulated gene, they are presumed to act in *cis*, that is, the alleles of a variant lead to differential expression of a target gene nearby on the same chromosome. When an individual is heterozygous for a variant within a gene transcript, allele-specific expression (ASE) can be detected, which can confirm that an eQTL acts in *cis*. While eQTLs can also act in *trans*, few *trans*-eQTLs have been discovered.

Based on highly powered eQTL studies in blood, it has become clear that most genes in the genome have a detectable *cis*-eQTL (Battle et al. 2014; Westra et al. 2013). Furthermore, most eQTLs appear to propagate their effects to protein levels, and to be associated with changes to chromatin accessibility at nearby regulatory elements (Y. I. Li et al. 2016). This leads to the concept of a regulatory cascade: genetic variants alter TFBS at distal regulatory regions or promoters; this leads to changes in chromatin accessibility and histone modifications, followed by changes to gene expression, and finally translation and protein expression levels. Not all gene regulation occurs via this model, however. With RNA-seq data, QTLs can also be mapped for the rate of splicing of gene introns (sQTLs), and a recent study estimated at least as high an enrichment of GWAS hits for sQTLs as for eQTLs (Y. I.

Li et al. 2016). This suggests that, to date, changes to gene splicing may have been an underappreciated mechanism linking genetic variation to complex traits.

Because eQTLs and sQTLs are linked to specific genes, a powerful way to interpret GWAS associations is to look for overlap with a QTL, in which case the regulated gene is a good candidate for mediating effects on the complex trait. Yet, despite growing eQTL datasets, only a few GWAS loci have been clearly demonstrated to act via this mechanism. In addition, estimates of the fraction of autoimmune GWAS loci that share causal genetic variants with eQTLs put the number at just 25% (Chun et al. 2017). This is puzzling, since in the absence of coding variant associations, effects on gene expression would seem to be the primary alternative explanation. A few reasons may explain the failure so far to link a large number of GWAS associations with eQTLs. First, eQTLs suffer the same problem as GWAS, in that LD makes it difficult to identify causal variants. Second, it is almost certain that we have not yet discovered all of the QTLs, across all cell types and contexts. Third, determining overlap between QTLs and GWAS associations is non-trivial, since the prevalence of QTLs across the genome means that chance overlaps are common (Lappalainen et al. 2013). Rigorous statistical methods, such as coloc (Giambartolomei et al. 2014), are essential to evaluate whether a given overlap is more consistent with shared or distinct causal variants. However, the sensitivity of these methods to detect true overlaps, particularly in the case of multiple causal variants, is unknown.

There remains hope that studies of gene expression will ultimately help to elucidate molecular mechanisms at a large fraction of GWAS loci. The GTEx project has to date released eQTL analyses for only about half of its target sample size. In contrast with previous reports, the latest GTEx analysis found that 52% of GWAS associations across 21 traits were colocalized with an eQTL in at least one tissue (GTEx Consortium et al. 2017). As the GTEx sample size grows, this fraction is likely to increase. In addition, a number of ongoing eQTL studies are being performed in specific cell populations not profiled by GTEx, and under conditions of cellular stress or immune challenge. It is noteworthy, however, that half of the GWAS loci colocalized with a GTEx eQTL actually colocalized with more than one eGene. This implies that identifying the causal gene will still require further mechanistic evaluation of any colocalized eGenes.

1.2.2 Transcription factor binding

There are an estimated 1,000 - 2,000 human genes encoding transcription factors (TFs) (Vaquerizas et al. 2009). Each TF binds to DNA having specific sequence features, which is

primarily determined by the nucleotide sequence itself. The preferred sequences to which a TF binds can be captured as a short sequence motif (< 20 nucleotides). As a cell differentiates from pluripotent or progenitor cell states to more specific cell types, numerous genes are activated and others are repressed. A subset of TFs known as pioneer TFs are efficient at opening repressed chromatin, and these are often critical “master regulators” of specific cell lineages, such as FoxA for liver (Iwafuchi-Doi et al. 2016), GATA4 for heart (P. Zhou, He, and Pu 2012), and PU.1 and C/EBPα for macrophages (Ruffell et al. 2009; Heinz et al. 2010). Where pioneer TFs have opened the DNA, other TFs are then able to bind, creating loops between regulatory regions and gene promoters to determine expression levels of specific genes. This co-binding of multiple TFs is a common feature of gene regulation and lineage specification (Chronis et al. 2017).

The ENCODE project has used chromatin immunoprecipitation followed by sequencing (ChIP-seq) to profile the binding of dozens of TFs across many human cell types (ENCODE Project Consortium 2012). Since it is thought that many gene regulatory variants act by altering TF binding, a knowledge of the locations of TFBS should be highly informative for locating causal regulatory variants. There are a number of reasons why this has not yet been fully realized. First, even the more than 2,000 TF binding assays performed by ENCODE are still a sparse sampling of the full matrix of TFs and cell types. Second, high quality antibodies are only available for some TFs. Third, it is unknown what fraction of TBFS are functional; because TF motifs are short, they are highly numerous across the genome. Not all occurrences of a motif are bound by an expressed TF with preference for that motif, and conversely, not all TFBS have a distinguishable motif present. Lastly, TF binding can be influenced by factors outside of core motifs, such as DNA shape (Mathelier et al. 2016), DNA accessibility, and the co-binding of other TFs to nearby sites.

1.2.3 Histone modifications

Histones are proteins conserved across all living organisms which bind as octamers to DNA, composed of pairs of subunits H1, H2, H3, and H4. These histone octamers, called nucleosomes, each have around 150 base pairs of DNA wrapped around them, and are bound across most of the DNA in a cell. Histones are essential to compacting the billions of base pairs of DNA sequence in a eukaryotic cell into the nucleus. They also are key factors in determining which regions of DNA are active or repressed, which differs between cell types. The N-terminal tails of histones H3 and H4 protrude from the nucleosome, and specific amino acid residues can have covalent modifications added to them. These post-translational modifications are highly correlated with different aspects of the DNA sequence,

such as transcribed and regulatory regions. For example, the gene bodies of actively transcribed genes tend to have nucleosomes with high levels of trimethylation at lysine 36 of histone H3, abbreviated H3K36me3. Both promoters and active transcriptional enhancers are often marked with acetylation at lysine 27 of histone H3 (H3K27ac). Antibodies are available for many specific histone modifications, so that ChIP-seq can be used to measure these genome-wide. There are more than a hundred sites at which histones can be modified (Tan et al. 2011), but only a few of these have been studied in depth, summarised in Table 1.

Histone modification	**Association**
H3K4me1	activation: broad peaks at enhancers
H3K4me3	activation: sharp peaks at promoters of poised and active genes
H3K9ac	activation: active promoters, release of paused RNA Pol II (Gates et al. 2017)
H3K9me3	silencing: broad regions of heterochromatin
H3K27ac	activation: sharp peaks at active enhancers and promoters (Creyghton et al. 2010)
H3K27me3	silencing: broad regions of Polycomb repression, poised/bivalent gene promoters
H3K36me3	activation: active transcription, transcriptional elongation
H3K79me1/me2	activation: active and silent promoters

Table 1: Properties of widely studied histone modifications. Most are described in (Barski et al. 2007). Abbreviations are as follows: me1: mono-methylation; me2: di-methylation; me3: tri-methylation; ac: acetylation.

Histone modifications have contributed greatly to the annotation of regulatory DNA. They are one of the widely-used inputs for genome segmentation, in which a model attempts to integrate multiple annotations to assign distinct states, such as enhancer, promoter, or transcriptional elongation, to each DNA segment across the genome (Ernst and Kellis 2012; Hoffman et al. 2012). Genome segmentation simplifies interpretation of the complex patterns of co-occurring histone modifications along the genome, which are often highly correlated amongst each other. However, because histones are present in nucleosomes with a periodicity along the DNA of ~150 - 200 bp, the resolution of regulatory information contained in histone modifications is also naturally limited to around 200 bp.

1.2.4 Chromatin accessibility

Regulatory regions of the genome have a DNA conformation that is open and accessible to transcription factor binding, whereas most other genomic regions have lower accessibility. A key advantage of measuring chromatin accessibility is that it is agnostic to the particular TFs that bind and open DNA at regulatory regions and promoters. Therefore, these regions can be identified in any cell type, without a need to assay each TF independently. In addition, using TF sequence motifs it is sometimes possible to predict the TFs that are bound at a regulatory region without directly measuring them. A method based on this idea, called centisnp, suggested that 97% of genetic variants in inferred TF binding footprints have no effect on chromatin accessibility (Moyerbrailean et al. 2016). Predictions from this method are used as one of the inputs the model that I describe in Chapter 3.

Until recently, chromatin accessibility was primarily measured by digesting native DNA with DNase I followed by sequencing, which found open chromatin covering about 1% of the genome. More recently, an alternative measure of chromatin accessibility based on integration of Tn5 transposase into native chromatin, known as ATAC-seq, has come into widespread use due to the simpler protocol and the ability to perform it with fewer cells (Buenrostro et al. 2013). Genetic variants influencing chromatin accessibility (caQTLs) can be identified with very modest sample sizes (< 100 individuals). While caQTLs do not directly indicate a target gene, the majority of caQTLs appear to be due to variants within the chromatin accessibility peak that they regulate (Degner et al. 2012). This suggests that overlap between eQTLs and caQTLs can be a powerful method to localise causal eQTL variants.

1.2.5 Chromosomal conformation

Techniques such as Hi-C, which capture information about chromosomal interactions genome-wide, have shown that the genome is organized into topologically associating domains (TADs) of around 100 kb - 1 Mb, with the regulation of genes in different TADs insulated from each other (Dixon et al. 2012; Rao et al. 2014). One of the key factors determining chromosomal conformation is CTCF, a transcription factor with a particularly clear binding motif, which is found at TAD boundaries. Whereas TAD boundaries are largely conserved across cell types and across species, DNA contacts within TADs, such as enhancer-promoter loops, are more dynamic and vary between cell types. Hi-C therefore has the potential to help in identifying causal regulatory variants by indicating the regulatory regions in contact with specific gene promoters.

Because Hi-C attempts to assay all pairwise interactions between DNA segments, achieving resolution of better than 25 kb requires a very large number of cells and deep sequencing (Rao et al. 2014). These requirements are reduced in promoter capture Hi-C, where an oligonucleotide pulldown enriches the Hi-C library for promoter-interacting regions before sequencing. Promoter capture Hi-C in 17 blood cell types by the BLUEPRINT consortium was used to prioritize 2,604 candidate genes for association with 31 GWAS traits (Javierre et al. 2016), although at each locus there were often a number of genes prioritized. Even when high resolution (<= 5 kb) is possible, it is difficult to detect significant interactions between DNA regions less than 25 kb apart, because nearby regions have a high rate of random collisions that are captured by the cross-linking used for Hi-C. However, it is clear from eQTL studies that the majority of causal regulatory variants are nearer than 25 kb from gene TSSes. This limitation reduces the utility of Hi-C data in localising causal regulatory variants.

1.3 iPSC-derived cellular models

While molecular traits can be measured in many cell types, technical limitations often make this difficult. Most assays require millions of cells, which are not always available from primary tissues, and this is particularly limiting for rare cell types. As a result, many of these assays have been performed on LCLs or cancer cell lines, even though these immortalized cells may not be good models for the relevant cell type. The use of induced pluripotent stem cells (iPSCs) provides a potential solution to both limiting cell numbers and poor cell type models. iPSCs can be expanded in vitro to the required number of cells, and then differentiated in to specific cell types. Because iPSC technology is still quite new, it is important to understand the current state of the art.

1.3.1 Reprogramming somatic cells to pluripotency

In 2006, Takahashi and Yamanaka reported that somatic cells can be reprogrammed to pluripotency by the ectopic expression of just four transcription factors (Takahashi and Yamanaka 2006). This led to great excitement about the potential uses of these iPSCs for regenerative medicine, and in particular their advantages over embryonic stem cells (ESCs). IPSCs derived from a patient's own cells could provide an unlimited supply of stem cells, which could be differentiated into desired cell types for cell-replacement therapies, and would be unlikely to face immune rejection. An early demonstration of the potential for this was provided by the Jaenisch research group, who derived dopaminergic neurons from reprogrammed mouse fibroblasts, and showed that implanting these into the brains of rat models of Parkinson's disease led to functional integration and improved disease symptoms (Wernig et al. 2008). The development of such therapies for humans, however, depends

upon a thorough assessment of the safety and reliability of iPSC-based cells for specific applications, as well as the development of efficient protocols to derive specific cell types.

ESCs can differentiate into any cell type in the body; to be considered pluripotent, the same should be true of iPSCs. Although a number of reprogramming methods have now been developed, reprogramming is inefficient, and typically only up to 1% of somatic cells appear to attain pluripotency (Robinton and Daley 2012). As a result, iPS cell lines are grown from single-cell clones, and pluripotency is usually assessed by looking for molecular hallmarks, such as expression of genes *OCT4*, *SOX2*, and *NANOG* at levels comparable to ESCs. More robust validation of pluripotency can be obtained by showing that the cells can differentiate into the three embryonic germ layers in vitro. Even though iPSCs seem to be capable of differentiating into any cell type, a number of groups reported that individual cell lines showed more efficient differentiation to specific lineages (Kim et al. 2010). In particular, a concern is that iPSCs retain an epigenetic memory of the cell type they originated from, implying that reprogramming to pluripotency is generally incomplete (Bar-Nur et al. 2011; Polo et al. 2010). A problem with these comparisons was that the cell lines used were derived from different donors, and thereby had different genetic backgrounds. When Bock and colleagues used a quantitative differentiation assay as well as measuring DNA methylation in 20 ESC and 12 iPSC lines, they found substantial variation among ESCs as well as iPSCs (Bock et al. 2011). In addition, global differences in gene expression are more significant in earlier passages of iPSCs, suggesting that pluripotency is gradually established over time (Polo et al. 2010).

More recently, cell banks of hundreds of human iPSC lines have been generated in a consistent manner by the NextGen consortium (Warren, Jaquish, and Cowan 2017) and the HIPSCI initiative (Kilpinen et al. 2017). These have revealed that donor genetic background contributes substantially to molecular variation in iPSCs. As well, improved protocols and characterisation of cell lines may help to overcome challenges related to the pluripotency and heterogeneity of iPSCs (D'Antonio et al. 2017; Panopoulos et al. 2017). For example, although it was was previously necessary to culture iPSCs on a "feeder" layer of mouse embryonic fibroblasts, new media with specific growth factors have enabled maintaining iPSCs without feeders, simplifying cell culture protocols.

A growing use of iPSCs is to differentiate them into specific cell types to model disease phenotypes *in vitro*. This is particularly valuable for rare and inaccessible cell types, which otherwise would be difficult to study. iPSC-derived cells can be used to discover cell type-specific molecular QTLs, to screen drugs for effects on cellular phenotypes, and to identify

causal variants via gene editing with CRISPR-Cas9. These capabilities provide the motivation for the experiments described in Chapter 2. We differentiated iPSC lines from different donors in the HIPSCI project to pain-sensing sensory neurons, a cell type that would be difficult to study *in vivo*. By collecting multiple molecular phenotypes across a large panel of cell lines, the aim was to link common genetic variation with both molecular traits and electrophysiological traits of sensory neurons. In parallel with this experiment, a GWAS was conducted comparing more than 18,000 individuals with chronic pain to controls. The hope was that molecular QTLs from our study would inform interpretation of any GWAS loci associated with pain. Unfortunately, no genome-wide significant loci for pain were found. Despite this limitation, we found a number of QTL-GWAS overlaps that likely reflect neuronal functions more broadly. In addition, during the course of our work we came across a challenge that has been noted in previous iPSC work, but which was particularly acute given the large number of differentiations in our study.

1.3.2 Heterogeneity and limited maturity of iPSC-derived cells

Differentiating iPSCs into defined cell types is a process that generally involves the addition of combinations of specific growth factors and media over a period of weeks, attempting to mimic endogenous developmental signals. A key challenge in using these as models is that although the resulting cells display characteristics of the desired cell type, they usually appear immature; this immaturity has been reported for multiple cell types, including neurons (Handel et al. 2016), hepatocytes (Dianat et al. 2013), cardiomyocytes (Veerman et al. 2015), and hematopoietic cell types (Smith et al. 2013). This may reflect in part the trade-off between experimental throughput and the time allowed for differentiation; however, it could also indicate that full maturation of cells requires a multicellular tissue environment that is absent in most culture systems (Passier, Orlova, and Mummery 2016).

A related but distinct challenge for iPSC-derived cell models is that differentiation tends to produce a mixture of cells, only a fraction of which express the expected marker genes. These differentiation outcomes can be highly variable between cell lines, and even across cultures of the same cell line. The nature of these “contaminating” cells is not generally known, but single-cell characterisation of cultures at multiple time points during differentiation is beginning to shed light on factors that lead to this variability. Reconstructing the differentiation course of cells undergoing MyoD-mediated reprogramming to contractile myotubes suggested that cells can take alternative branches, with those that select incorrect branches ending in aberrant cell states (Cacchiarelli et al. 2017). For some cell types and applications, it is possible to sort differentiated cells to enrich for the desired outcome.

Another approach was described recently for neurons, where single cells were measured with patch clamp electrophysiology, followed by sequencing of the same single cells (Bardy et al. 2016). This approach enabled the researchers to link molecular cell states directly with functional features of the same cells.

Despite heterogeneity in iPSC differentiation, it has been possible to observe disease-relevant phenotypes for Mendelian diseases in iPSC-derived cells. For example, iPSC-derived sensory neurons from patients with inherited erythromelalgia, which is due to mutations in the sodium channel Nav1.7, showed greater spontaneous firing than those from control individuals, and this was reverted to normal with the addition of a Nav1.7-blocking drug (Cao et al. 2016). Similarly, iPSC-derived cardiomyocytes from individuals with long-QT syndrome showed prolonged action potentials, and this could be modulated by existing drugs used for long-QT syndrome (Itzhaki et al. 2011).

Heterogeneity is likely to be a greater problem when attempting to model the effects of complex trait-associated variants *in vitro*, due to their smaller effect sizes. Studies reported as part of the NextGen consortium have taken the first steps in this effort. Warren et al. recruited individuals homozygous for the major or minor genotypes (17 each) at the LDL-C associated variant rs12740374. They differentiated 68 iPSC lines from these individuals into hepatocytes and adipocytes, demonstrating that rs12740374 influenced *SORT1* expression primarily in hepatocytes (Warren et al. 2017). Pashos et al. differentiated iPSCs from 91 healthy donors to hepatocyte-like cells, and used RNA-sequencing to map eQTLs. For four eQTLs that colocalized with GWAS associations for lipid traits, they used a massively parallel reporter assay to identify putative causal SNPs, followed by CRISPR/Cas9 genome editing to validate the candidate SNPs (Pashos et al. 2017). These impressive studies showed that using iPSC-derived cells to model complex trait associations is possible, but also revealed that doing so requires large samples sizes and very considerable effort.

1.4 Predicting variant functionality

Identifying causal variants for Mendelian or complex traits requires experimental validation. However, investigating the molecular effects of individual genetic variants is a laborious process that can usually only be done for a handful of variants at most. Computational approaches to prioritize variants to investigate are thus essential. The enormous growth in number and types of genomic data has fueled a growth in methods using these data to predict variant functionality and to fine-map GWAS associations. Predicting the functional effects of genetic variants is also of more general interest, since it may shed light on the

basic regulatory grammar that determines genome function. It is important to first clearly define what is meant by “functional”.

A widely held assumption, supported by evolutionary theory, is that most genetic variants are neutral, meaning that they have little effect on organismal fitness. This is consistent with the observation that a relatively small fraction of the genome (4-7%) has significant sequence conservation across mammals (Siepel et al. 2005; Davydov et al. 2010). Moreover, since selection depletes deleterious variants, the fraction of common variants with fitness effects is likely to be even lower. However, conservation differs from function. First, transcription factor binding sites generally have low conservation across species, despite having clear functional effects if disrupted (Doniger and Fay 2007). Second, variants may have “functional” molecular effects without having organismal effects. Third, a functional variant may have a deleterious effect only late in life, and therefore have little effect on fitness and sequence conservation, even though it affects a trait. In this thesis, a functional variant is one with a molecular effect, regardless of whether that effect propagates to any other phenotype. Still, since neutral variants are unlikely to have effects on complex traits, functional variants are more likely to be causally related to complex traits.

1.4.1 Variant annotation

One way to stratify variants into classes that are more or less likely to have functional effects is to annotate them with available genomic features. A researcher can then manually assess the evidence for a given variant’s function using prior knowledge. Early annotation tools focused on identifying the effects of variants on protein-coding genes (Cingolani et al. 2012; McLaren et al. 2016). This is more technically challenging than it appears at first glance, and different tools often produce discordant annotations (D. J. McCarthy et al. 2014). Reasons for this include differences in gene annotations across reference databases, as well as the difficulty in determining the effects of variants on splicing. While these are essential tools, other annotations are important for predicting non-coding variant function.

The tool HaploReg (Ward and Kellis 2012) integrates a large number of genomic datasets, and reports a variant’s overlap with genes, enhancer and promoter marks, DNase hypersensitive sites, protein binding sites, known eQTLs, and TF binding motifs. A useful feature is that it uses LD information to also annotate variants with $r^2 > 0.8$ with the query variant. Interpreting the large number of overlaps is challenging, however. For GWAS associations it is typical to have dozens of variants in LD, a majority of which overlap some potentially relevant feature. RegulomeDB provides a heuristic solution to interpreting

overlapping annotations (Boyle et al. 2012): a variant is assigned to one of 14 handmade categories based on prior beliefs about the informativeness of each annotation. For example, a variant would receive a very high score if it alters a TF motif in a known TFBS, within a DNase hypersensitivity peak, and is also a known eQTL. A variant with fewer overlaps would receive a lower score depending on which annotations were absent.

1.4.2 Integrative approaches to score variant functionality

A general problem in predicting the functions of genetic variants is that informative annotations are correlated with each other. For example, while histone modifications H3K4me3 and H3K27ac are both enriched for eQTL variants, they frequently colocalise at gene promoters and transcribed enhancers, and so combining their independent enrichments would overly prioritise variants where these annotations overlap. Methods have been developed which address this problem to predict variant functionality more rigorously.

Polyphen (Adzhubei et al. 2010) and SIFT (Kumar, Henikoff, and Ng 2009) are widely-used methods to predict the likelihood that a nonsynonymous protein-coding variant is deleterious. Both methods rely on the frequency of amino acid substitutions in homologous proteins, and Polyphen also considers protein structural features such as transmembrane domains and ligand-interacting regions. These methods are sometimes used as input to “meta-prediction” tools, which evaluate both protein-coding and non-coding variants genome-wide.

CADD applied machine learning to integrate many annotation sources, assigning a “deleteriousness” score to coding and non-coding variants (Kircher et al. 2014). Its inputs include Polyphen and SIFT, as well sequence conservation, and annotations from ENCODE such as DNase hypersensitivity, TFBS, and genome segmentations. CADD’s score is based on a support vector machine trained to distinguish common variants and human derived alleles from simulated variants, which are not present in the genome and so are presumed to have been depleted by selection. CADD has been widely used to prioritize variants in studies of Mendelian disease. A particular strength is that it scores both coding and non-coding variants on the same scale, which enables evaluating both of these types of variation in relation to disease. However, CADD’s prediction performance is likely to be different for coding and non-coding variants, and CADD has been criticized for performing poorly on identifying functional variants in eQTL datasets (Gulko et al. 2015).

A number of methods focus exclusively on non-coding variation, such as GWAVA (Ritchie et al. 2014), LINSIGHT (Y.-F. Huang, Gulko, and Siepel 2017), DeepSEA (J. Zhou and

Troyanskaya 2015), and Basset (Kelley, Snoek, and Rinn 2016), and these are discussed in more detail in Chapter 3. Both the models implemented by these methods and the annotations used as input differ, which makes it difficult to disentangle the factors influencing their relative performance. The general lack of a gold standard set of functional non-coding variants means that prediction performance can only be assessed relative to proxies, such as results from reporter assays. Although these scores of variant deleteriousness or functionality are clearly useful in some contexts, it is not clear how a variant's score relates to its probability of influencing a particular trait. For complex traits, there is a need for methods that can be applied in a rigorous way to help in identifying causal variants.

1.5 Fine-mapping GWAS associations

Identifying causal variants at a GWAS locus is a key step towards deciphering the molecular mechanism behind the association. The first step towards this is fine-mapping - reducing the set of candidate variants from all those at the locus to a smaller set that is highly likely to contain the causal variant. Although most GWAS have used sparse genotyping to tag causal variants, a key assumption of fine-mapping is that the causal variant is among the variants considered. Therefore, samples within the study must either have whole genome sequencing (the ideal case) or must have genotypes imputed using a reference panel from a genetically similar population of individuals. The set of candidate causal variants is often referred to as a credible set, which can be defined to have a specified probability of containing the causal variant(s), given that particular assumptions are met. Commonly, 95% or 99% credible sets are reported. Approaches to fine-mapping can be roughly divided into those which are purely statistical, those which leverage additional data such as epigenomics and gene annotations, and experimental approaches.

1.5.1 Experimental evaluation of variants

The gold standard evidence to indicate that a specific variant has a causal effect is to experimentally replace the allele in a native cellular context. Following allelic replacement, cellular phenotypes such as gene expression can be assayed to determine whether the alleles differ in their activity on the same genetic background, with the assumption that alleles showing a molecular effect are likely to also causally influence the complex trait. Only recently have "genome editing" molecular tools such as CRISPR/Cas9 made this feasible, and the number of GWAS loci validated in this way remains small. Performing even a single such "knock-in" currently takes months at a minimum, and is difficult to perform in some cell types. A slightly lower standard of evidence is to use CRISPR/Cas9 to create small deletions overlapping a variant, which can show that the region covering the variant is

functional. When applying these approaches, it is critical to be highly confident that the causal variant is among the small number that can feasibly be tested.

A higher-throughput experimental alternative to genome editing is to use a reporter assay. Here, putative regulatory sequences, such as sequences surrounding highly ranked variants in an association study, are synthesized as oligonucleotides and inserted upstream of a reporter gene (e.g. green fluorescent protein (GFP) or luciferase) and transfected into cells. Sequences that regulate expression of the reporter gene will alter the measured level of GFP or luciferase. Reporter assays can also be done at scale. Tewhey and Sabetti used a massively parallel reporter assay (MPRA) to test 32,373 variants from 3,642 cis-eQTL loci for differential effect between alleles (Tewhey et al. 2016). Although this study focused on eQTLs rather than GWAS, the same approach could be used to test credible set variants from GWAS in cases where altered gene expression is the most likely mechanism. Even among eQTLs, MRPA only detected an expression-modulating effect of a genetic variant for ~9% of eQTLs tested, some of which will be false positives (Tewhey et al. 2016). It should be kept in mind that MPRA will only detect an effect for variants that alter gene transcription levels, i.e. enhancer or promoter variants, and not mechanisms that alter splicing or post-transcriptional regulation. Also, the effect that a variant has in a native cellular context may be unobservable or have a different direction of effect when tested in a reporter construct (Inoue et al. 2016). As a result, MPRA can be a useful complement to other fine-mapping approaches but does not obviate them.

1.5.2 Statistical fine-mapping

Early approaches to statistical fine-mapping, such as that used in the Wellcome Trust Case Control Consortium (Wellcome Trust Case Control Consortium et al. 2012), assumed that a single causal variant was present at an associated locus. A concern with this assumption is that when multiple causal variants do exist, the most strongly associated variants may be non-causal, due to being in LD with more than one causal variant. Leading methods developed more recently, such as CaviarBF (Chen et al. 2015), GUESSFM (Wallace et al. 2015), and FINEMAP (Benner et al. 2016), account for the possibility that multiple causal variants may explain the association signal. These approaches require testing different combinations of putatively causal variants, a task which quickly becomes computationally infeasible as more causal variants are allowed. GUESSFM and FINEMAP search a subspace of potential causal variant configurations that approximates the results obtained from examining all configurations. CaviarBF and FINEMAP require only summary statistics from the association study, which extends their utility to the many cases where sample

genotypes are not available. To accomplish this, they depend upon an external reference panel of pairwise LD between variants, such as the 1000 genomes project or the Haplotype Reference Consortium. Importantly, the reference panel population's ethnicity must be well-matched with that of the GWAS, otherwise the statistical fine-mapping may give spurious results. An additional drawback of purely statistical approaches is that when there are multiple variants in very high LD, they are nearly statistically indistinguishable.

1.5.3 Functional fine-mapping

Another set of methods incorporates functional genomic information to prioritize variants that have similar association statistics. This approach is supported by simulations showing that the size of the credible set can be reduced while retaining an equal probability of containing the causal variant (van de Bunt et al. 2015). Relevant annotations include overlap with gene bodies and proximity to gene TSSes, as well as the epigenetic traits discussed previously, such as chromatin accessibility, histone modifications, DNA methylation, and genome segmentation. Large-scale international consortia, such as ENCODE, Roadmap Epigenomics (Roadmap Epigenomics Consortium et al. 2015), and FANTOM (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014), have collected epigenomic data across many cell lines and tissues.

The vast scale of epigenomic datasets makes their use for variant interpretation especially challenging because most variants overlap some molecular annotation. Furthermore, in the context of GWAS it is unclear how much weight to place on the statistical association for a given variant versus the annotations the variant overlaps. There has been rapid progress in computational methods that attempt to solve these two problems simultaneously. Fgwas (J. Pickrell 2013) learns each annotation's overall enrichment for statistically associated variants directly from GWAS summary statistics, controlling for LD between variants. In a Bayesian framework, each variant's prior probability of causality then depends on its annotations, and this is combined with its association statistic to determine a posterior probability that the variant causes the association in a region. Fgwas assumes that a single variant in each region causes the association. PAINTOR (Kichaev et al. 2014) uses a similar Bayesian approach, but allows for multiple causal variants, at the cost of running time that increases exponentially with the number of potential causal variants allowed. Because these approaches rely on the signal from GWAS summary statistics, they are only likely to work well for highly powered GWAS where many associations contribute to the estimated annotation enrichments. RiVIERA-MT (Y. Li and Kellis 2016) and fastPAINTOR (Kichaev et

al. 2017) improve upon this by analyzing many GWAS jointly, while also allowing multiple causal variants at a locus.

In general, these approaches are flexible in that the user can select any relevant annotations to use in the model. However, this imposes the considerable burden of identifying the most relevant annotations. Prior knowledge of a trait can often be used to select a relevant cell type, but the number of available cell type-specific annotations is still large. Including too many annotations in fine-mapping raises the risk of overfitting, which can worsen performance. This therefore leads to the related problem of identifying the most relevant cell types for a given trait.

1.5.4 Identifying causal genes

In many respects, it is more important to identify causal genes than causal variants, since the proteins encoded by genes are the targets for drug development. Identifying the causal variant for a GWAS association can sometimes implicate a particular gene as causal, such as when it is located at a promoter or in a gene's transcript or splice sites. For other regulatory variants, because long-distance gene regulation is prevalent, the location of a causal variant is only weak evidence that nearby genes are involved. Hi-C and promoter-capture Hi-C can suggest causal genes that are distal from a GWAS association peak, although most datasets are limited in resolution to blocks ~25 kb or larger. These data also have not yet been generated across the broad array of cell types for which other regulatory annotations are available.

A unique approach to GWAS can discover associations directly tied to specific genes, and has been developed recently by several groups (Gamazon et al. 2015; Gusev et al. 2016; Zhu et al. 2016; Barbeira et al. 2017). These methods use reference eQTL data to predict gene expression levels in a GWAS cohort, and then test for association between the predicted expression and a trait. The latest of these methods can operate directly from GWAS summary statistics. In addition to defining the associated gene, they indicate the direction of effect between gene expression and the trait. Statistical power is dependent on the quality of eQTL reference dataset, as well as the match in LD patterns between the eQTL and GWAS cohorts. Just as many GWAS associations do not overlap with known eQTLs, not all associations discovered by traditional GWAS will be identified by GWAS for predicted gene expression.

1.6 Outline of the thesis

In this thesis, I used two approaches to leverage molecular QTL data to understand genetic associations with human phenotypes.

In Chapter 2, I describe mapping QTLs for gene expression and chromatin accessibility in 123 cell lines differentiated from iPSCs to sensory neurons. A key observation in this effort was that the differentiated cells contained a mixture of neurons and contaminating fibroblast-like cells that was highly variable from one differentiation to the next. Using single-cell RNA-seq from one cell line, I generated reference gene expression profiles for the neurons and contaminating cells, and used these to estimate purity for each of the bulk samples. I found that sensory neuronal purity was influenced by whether the iPSCs they were derived from had been cultured on feeder cells or in feeder-free medium. Although this contributed to increased gene expression variability in the sensory neurons, by leveraging additional information from allele specific expression I found QTLs which were in high LD with with GWAS catalog associations, providing links to putative causal genes.

In Chapters 3 and 4, I used public eQTL data to develop a model that uses functional genomic data to predict non-coding variant functionality, and subsequently applied this to fine-map GWAS associations from summary statistics. In Chapter 3, I evaluated a number of hypotheses on how genomic annotations could optimally be used to generate a model predicting the locations of causal eQTL variants. The resulting model enables the computation of genome-wide "PRF" scores, which reflect the cell type-specific **p**robability of **r**egulatory **f**unction, in any of 119 cell types profiled by the Roadmap Epigenomics project. In Chapter 4, I applied PRF scores to address two problems in post-GWAS analysis: (i) identifying relevant cell types, for individual loci and genome-wide, and (ii) fine-mapping to identify candidate causal variants.

The work in this thesis illustrates how both iPSC-based cellular models and large-scale data integration can link human genetic variation to complex trait phenotypes.

2 Molecular and Functional Variation in iPSC-Derived Sensory Neurons

Collaboration note

The work described in this chapter has been accepted (pending revisions) for publication by Nature Genetics (Schwartzentruber et al., 2017). Supplementary Tables are available at: https://github.com/js29/ipsdsn. Daniel Gaffney and Alex Gutteridge conceived and directed the project. Stefanie Foskolou performed all differentiations from iPSCs to sensory neurons. I performed nearly all data analyses described herein. Kaur Alasoo provided some of the code used for expression quantification and QTL calling. Helena Kilpinen determined which eQTLs were tissue specific. Alex Gutteridge performed quality control and analysis of neuronal calcium imaging and electrophysiology. Anna Wilbrey prepared samples for single-cell sequencing, and Alex Gutteridge conducted preliminary analyses of the scRNA-seq data.

2.1 Introduction

Cellular disease models are critical for understanding the molecular mechanisms of disease and for the development of novel therapeutics. In principle, induced pluripotent stem cell (iPSC) technology enables the development of these models in any human cell type. Initial uses of iPSCs for disease modelling have focused mostly on highly penetrant, rare coding variants with large phenotypic effects (Itzhaki et al. 2011; G.-H. Liu et al. 2011; Wainger et al. 2014; G. Lee et al. 2009; Cao et al. 2016). However, there is growing interest in using iPSCs to model the effects of the common genetic variants of modest effect size that drive complex disease (Warren, Jaquish, and Cowan 2017). A key question is to what extent variability in directed differentiation is a barrier to studying the effects of common disease-associated variants in iPSC-derived cells. In addition, because cultured cells are imperfect models of primary tissues, not all common disease-associated genetic variants also alter cell phenotypes in iPSC-derived systems.

Here, we present the first large-scale study of common genetic effects in a neuronal cell type differentiated from human stem cells, iPSC-derived sensory neurons (IPSDSNs). Peripheral sensory nerve fibres innervate the skin and other organs and are brought together at the dorsal root ganglia (DRG) before synapsing with the spinal cord around the dorsal horn. The development of efficient protocols to differentiate iPSCs into nociceptive (pain-sensing) neurons (Young et al. 2014) provides the opportunity to model common genetic effects on

human sensory neuron function, which may underlie individual differences in pain sensitivity and chronic pain. We investigate how power to detect common genetic effects is affected by the variability introduced by differentiation and demonstrate how initial iPSC growing conditions influence cell phenotypes in IPSDSNs. We identify quantitative trait loci (QTLs) for gene expression, RNA splicing, and chromatin accessibility and identify a number of overlaps between molecular QTLs and common disease associations. In generating this gene regulatory map we establish effective techniques for using IPSDSN cells to model molecular phenotypes relevant to common diseases.

2.2 Results

2.2.1 Sensory neuron differentiation and characterisation

We obtained 107 IPS cell lines derived from unrelated apparently healthy individuals by the HIPSCI resource (Kilpinen, Goncalves, et al. 2016), and followed an established small molecule protocol (Young et al. 2014) to differentiate these into sensory neurons of a nociceptor phenotype (Figure 1a). We performed a total of 123 differentiations; 13 of these were done with an early version of the protocol (P1) which was subsequently refined (P2) to reduce the number of differentiation failures and to yield a higher proportion of neuronal cells in the final cultures. One RNA-seq sample failed sequencing, and four others were outliers based on principal components analysis and were excluded. This left a set of 119 differentiations with gene expression data from 100 unique iPSC donors; all subsequent analyses focused on the 106 P2 protocol samples, except for QTL calling, where we used all samples to maximize discovery power.

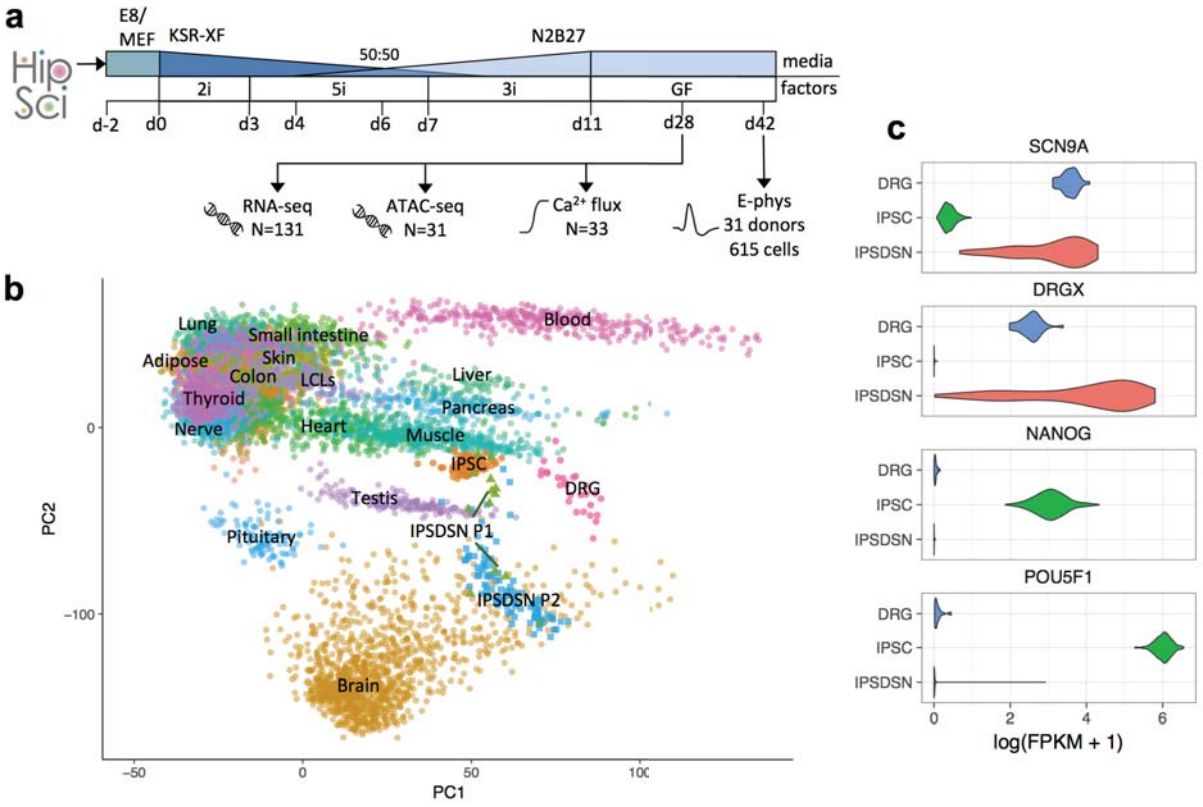


Figure 1 Characterization of molecular phenotypes in iPSC-derived sensory neurons. (**a**) Schematic of IPSDSN differentiation and assays. iPSCs were received in Essential 8 (E8) medium (N=82) or on mouse embryonic fibroblasts (MEFs, N=49), and transferred to KSR-XF medium. Over 11 days, different inhibitor combinations were added (2i, 5i, 3i, see Methods), and N2B27 medium phased in, followed by transfer to growth factor medium at day 11 for neuronal maturation. (**b**) PCA plot projecting IPSDSN, iPSC, and DRG samples onto the first two principal components defined based on RNA-seq FPKMs in GTEx tissues. Some GTEx tissues are unlabeled due to overlapping labels. (**c**) Expression of sensory neuronal marker genes (*SCN9A*, *DRGX*) and key iPSC genes (*NANOG*, *POU5F1*).

We clustered our gene expression data with 239 iPSC samples from the many of same donors, as well as 28 post-mortem DRG tissue samples from 10 different donors, and 44 primary tissues from the GTEx project (Mele et al. 2015) (Figure 1b). Globally, IPSDSN samples showed greatest similarity to iPSCs (gene expression correlation Spearman ρ=0.89), followed by DRG (ρ=0.84), and then brain samples from GTEx. However, because different gene expression quantitation methods were used in GTEx, we cannot be certain of relative similarities between GTEx tissues and the samples we uniformly processed (DRG, IPSDSNs, iPSCs). The similarity to iPSCs may reflect lack of maturity in IPSDSNs, which is a well-recognized problem with iPSC-derived cells (Warren et al. 2017; Sala, Bellin, and Mummery 2016; Soldner et al. 2016; Pashos et al. 2017). We also note that because the

same iPSCs were differentiated to IPSDSNs, both donor genetic background and cell culture effects may contribute to the observed similarity. Despite this, key sensory neuronal marker genes were highly expressed in IPSDSNs, while pluripotency genes were not (Figure 1c).

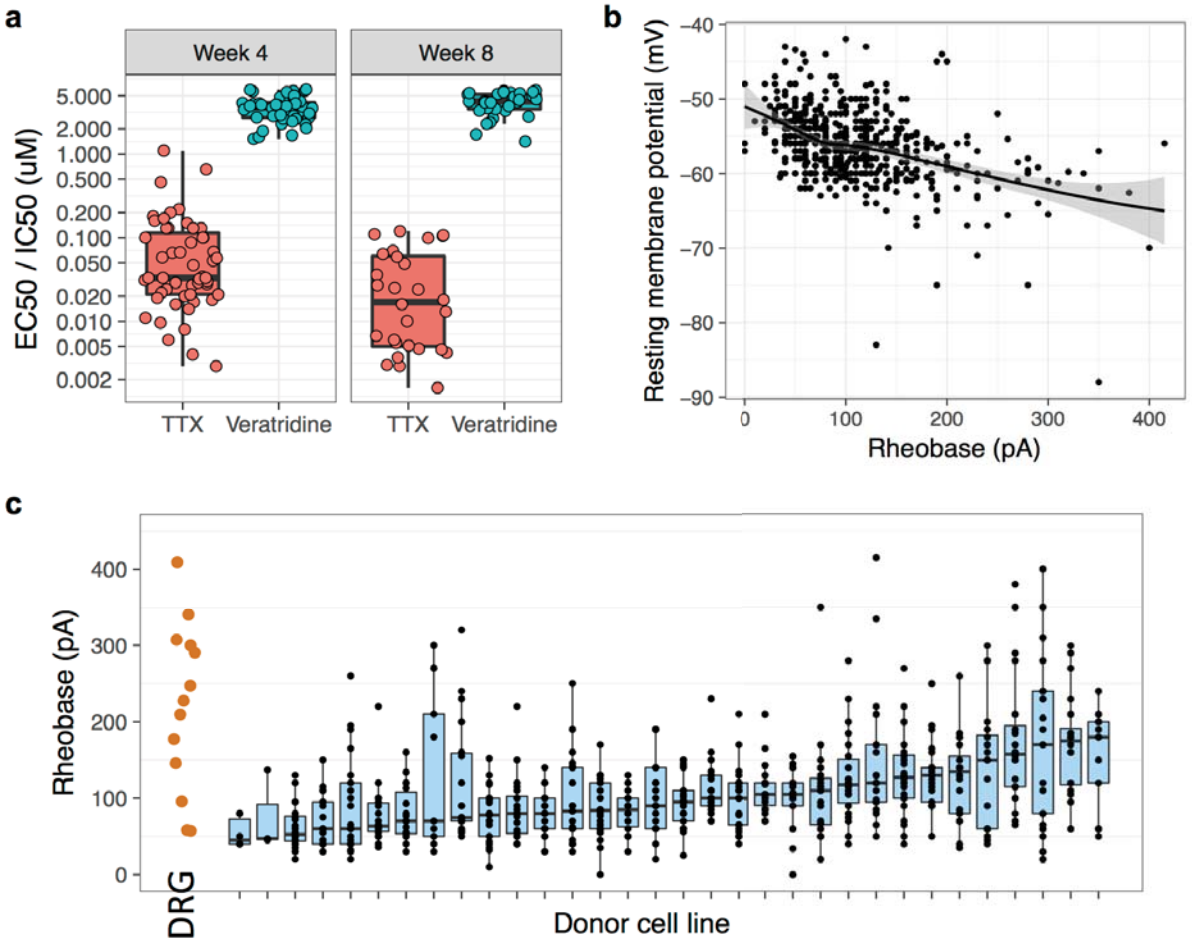


Figure 2: Functional characterization of IPSDSNs. (**a**) Ca2+ flux measurements on IPSDSN cultures (n=31) shows that neuronal firing is enhanced by veratridine (EC50/IC50 > 1) and is reduced by tetrodotoxin (EC50/IC50 < 1). (**b**) Rheobase is inversely related with resting membrane potential in individual neurons (n=616). (**c**) The distribution of rheobase values for the 31 samples with electrophysiology recordings, as well as literature values for DRG (leftmost bar).

Using Ca^{2+} flux measurements on a subset of differentiated cultures (n=31) we confirmed that the cells consistently responded to veratridine (a sodium ion channel agonist) and tetrodotoxin (a selective sodium ion channel antagonist), as expected (Figure 2a). We also performed patch-clamp electrophysiology recordings for 616 individual neurons from 31 donors, with a median of 21 cells measured per line. The rheobase is the minimum current input that will cause an individual neuron to fire an action potential, and we used this as a measure of the overall membrane excitability. Rheobase showed the expected inverse relationship with resting membrane potential (Figure 2b). The distribution of rheobases was

comparable to those obtained from primary DRG cells, but showed significant variation between donors (Figure 2c).

We next investigated whether variation in excitability was reflected in differences in gene expression of cells derived from the same donor. We examined the correlation between expression of individual genes and mean rheobase, which were measured in sister cultures from the same donor and differentiation batch. After correcting for multiple testing, no individual genes were significantly correlated with rheobase at FDR < 0.1. Similarly, none of the first five gene expression principal components were correlated with mean rheobase.

2.2.2 Quantifying differentiation variability using single-cell RNA-seq

Our samples appeared to differ in the fraction of cells with a neuronal morphology in microscopy images. A previous study using the same differentiation protocol showed that not all individual cells express neuronal marker genes after differentiation (Young et al. 2014). To further characterize this heterogeneity, we sequenced 177 IPSDSN cells differentiated in three batches from one iPSC line, and clustered them based on expression profiles using SC3 (Kiselev et al. 2017). The data were best explained by two clusters, with 63% of cells forming a tight cluster expressing sensory-neuronal genes (e.g. *SCN9A*, *CHRNB2*), and the remaining 37% of cells forming a looser cluster expressing genes typical of a fibroblastic cell

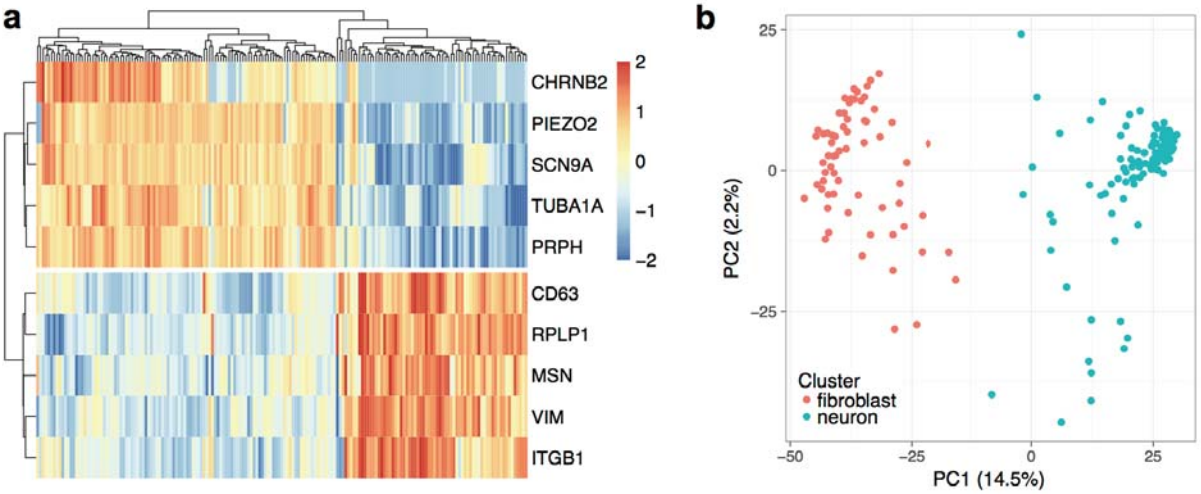


Figure 3: Single-cell sequencing of 177 IPSDSN cells. (**a**) Heatmap of RNA-seq data for ten marker genes of the two cell clusters identified by SC3. Color scale denotes normalized, Z-scaled gene expression counts for each gene. (**b**) PCA plot of PC1 vs. PC2 for 177 single cells, based on quantile normalized expression across all genes, with colour indicating SC3 cluster label.

type (e.g. MSN, VIM) (Figure 3a). The two cell types also separated cleanly in a principal components plot (Figure 3b), indicating that the cells do not fall on a smooth gradient from more neuronal to less, but rather have differentiated to distinct cell states. Comparing gene expression from each cluster to other tissues showed that the neuronal cluster was most similar to DRG (Spearman's ρ=0.654), followed by iPSCs (ρ=0.609) and GTEx brain (mean ρ=0.599) (Figure 4) while the fibroblast-like cluster was most similar to GTEx transformed fibroblasts (ρ=0.683), DRG (ρ=0.662), and iPSCs (ρ=0.653). The similarity of these cells to GTEx fibroblasts could suggest a general similarity of adherent cultured cells, although the neuronal cluster had lower similarity to GTEx fibroblasts (ρ=0.579) than many other tissues.

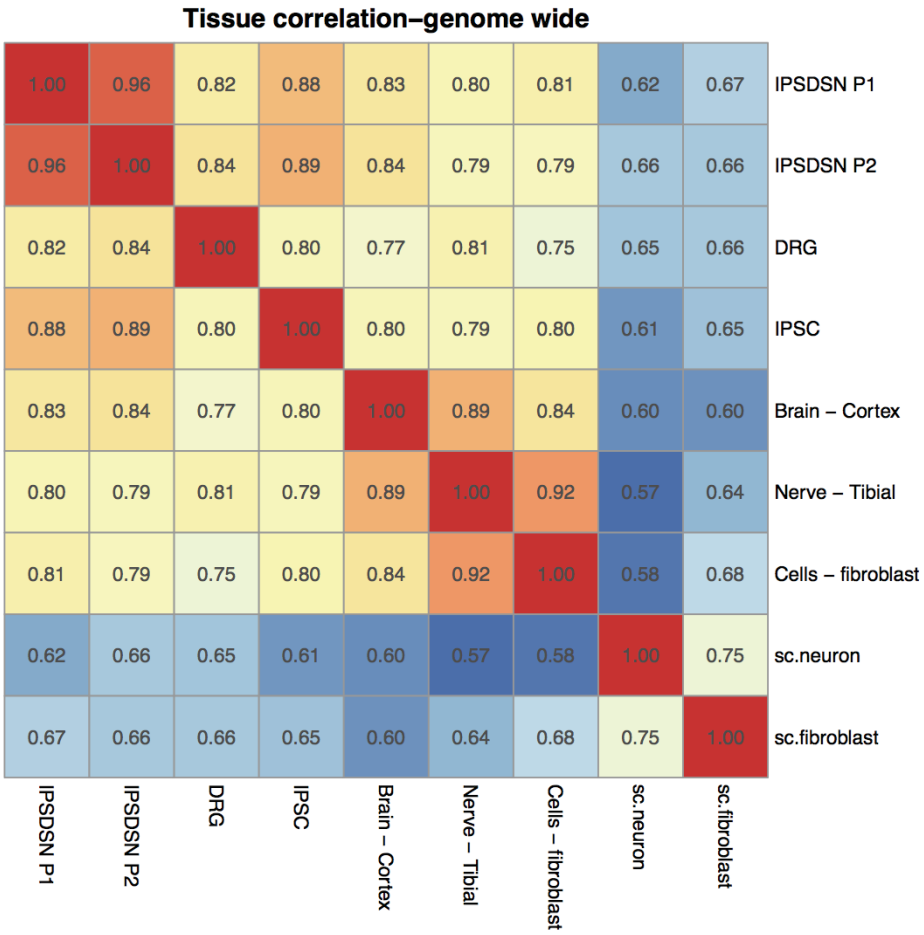


Figure 4: Correlation of genome-wide gene expression in different tissues and cell cultures. Gene expression for IPSDSN single cells was averaged within a cluster (sc.neuron, sc.fibroblast). For each gene the mean expression (FPKM) in the group of samples was computed, and these values were correlated genome-wide across groups. Spearman correlation values are shown in each square. P1 and P2 protocol samples are highly similar to each other, and compare similarly to other tissues. Both single cell neurons and single cell fibroblast-like cells have similarity to DRG, although single fibroblast-like cells have greater similarity with GTEx fibroblasts.

Next, we used CIBERSORT (Newman et al. 2015) to estimate the fraction of RNA from neuronal cells in our bulk RNA-seq samples, using the single cell gene expression counts with their cluster labels from SC3 as signatures of neuronal or fibroblast-like expression. The estimated neuronal content was strongly correlated with the first principal component of gene expression (R^2 = 0.75, Figure 5), and this corresponded well with a visual assessment of neuronal content from microscopy images (Figure 6).

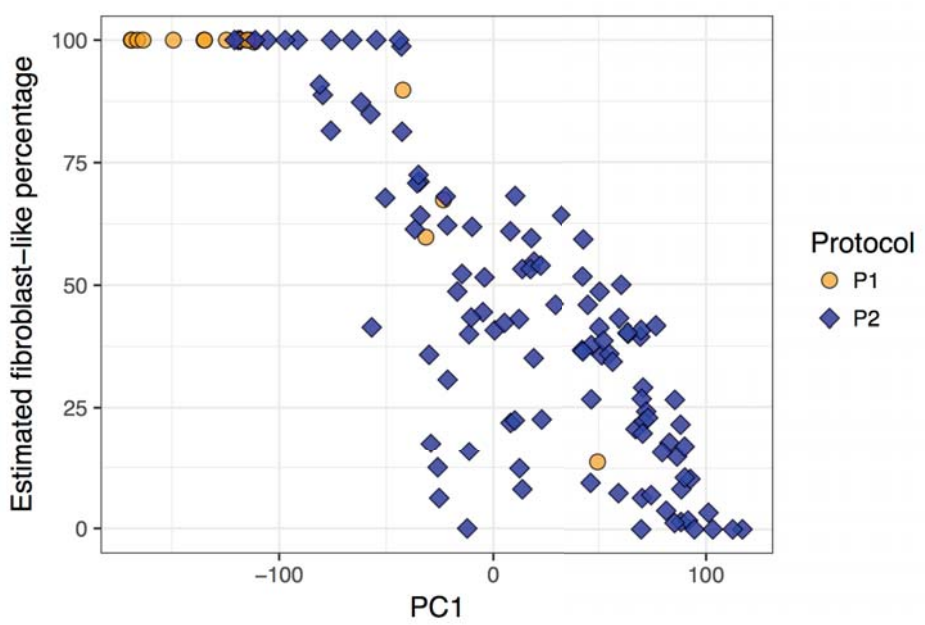


Figure 5: Plot of estimated fibroblast percentage for bulk RNA-seq samples versus gene expression principal component 1, after excluding 5 outlier samples. Although many samples have estimated fibroblast percentage close to 100%, these samples contained a significant fraction of neuronal cells in microscopy images.

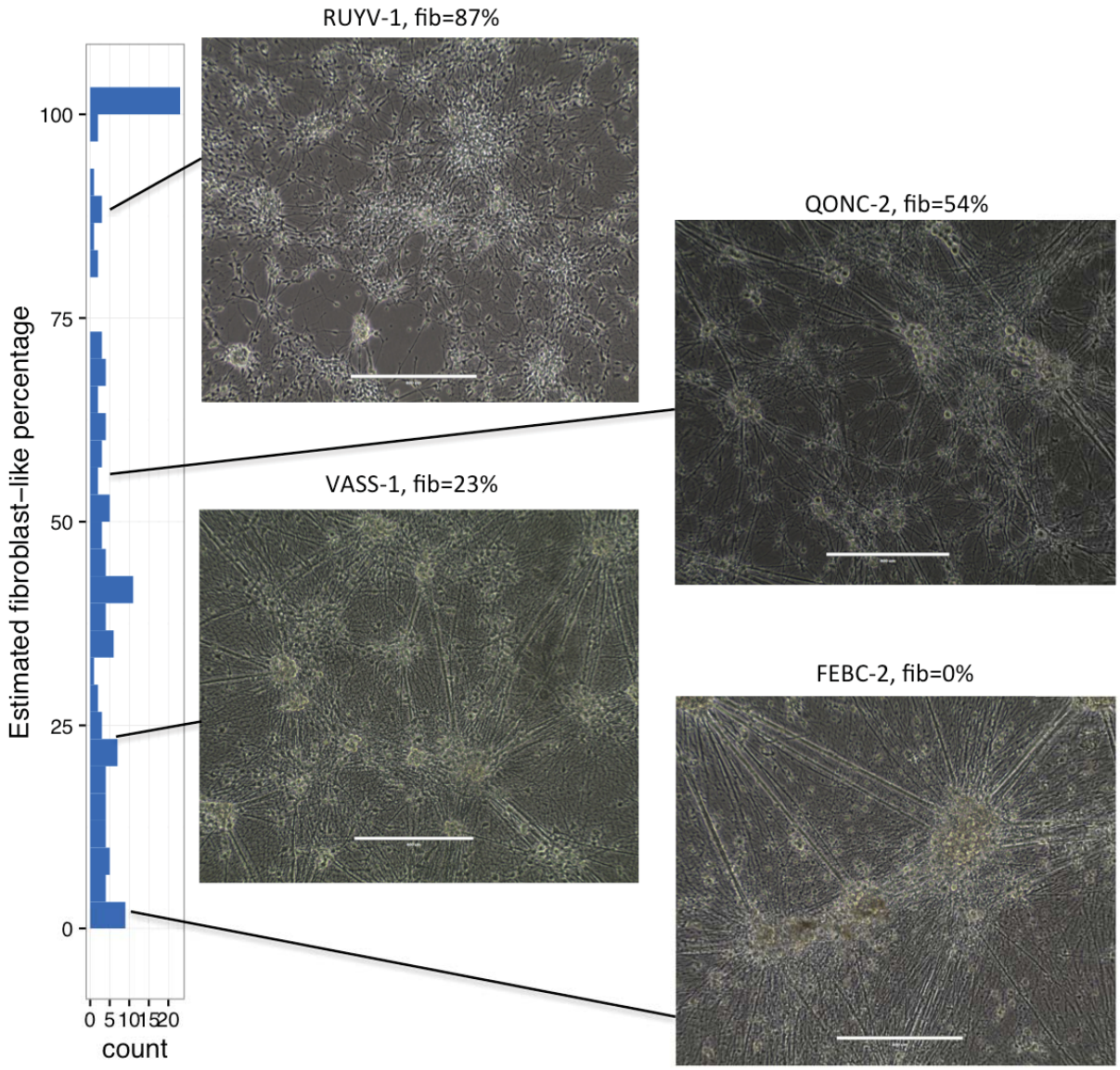


Figure 6: Images of four cell lines, taken 4 weeks post induction of differentiation, with their estimated fibroblast-like content shown.

Although a majority of samples appeared by microscopy to have high neuronal content, CIBERSORT estimated relatively high fibroblast-like content for many samples (mean 49%). A factor contributing to this may be the greater RNA content of fibroblast-like cells, which had 2.3-fold more reads in the single-cell RNA-seq data. Indeed when the single cell counts were pooled, CIBERSORT estimated the fibroblast content of this "sample" as 60%, considerably higher than the 37% of single cells in the fibroblast-like cluster. A second consideration is that our scRNA-seq sample was matured for 8 weeks, whereas our bulk RNA-seq samples were matured for 4 weeks. Although gene expression changes are minor after 4 weeks maturation (Young et al. 2014), this difference in maturity means that our single cell reference profiles do not perfectly represent cells in our bulk samples. Despite

this, IPSDSN samples estimated to have high fibroblast content still showed greater similarity in genome-wide gene expression with DRG than with any GTEx tissue, including fibroblast cell lines (Figure 7). Although these similarities are reassuring, we note that technical factors could contribute to the greater similarity with DRG, as different gene expression quantification tools were used for GTEx (RNASeQC) and for our iPSC, DRG, and IPSDSN samples (featureCounts).

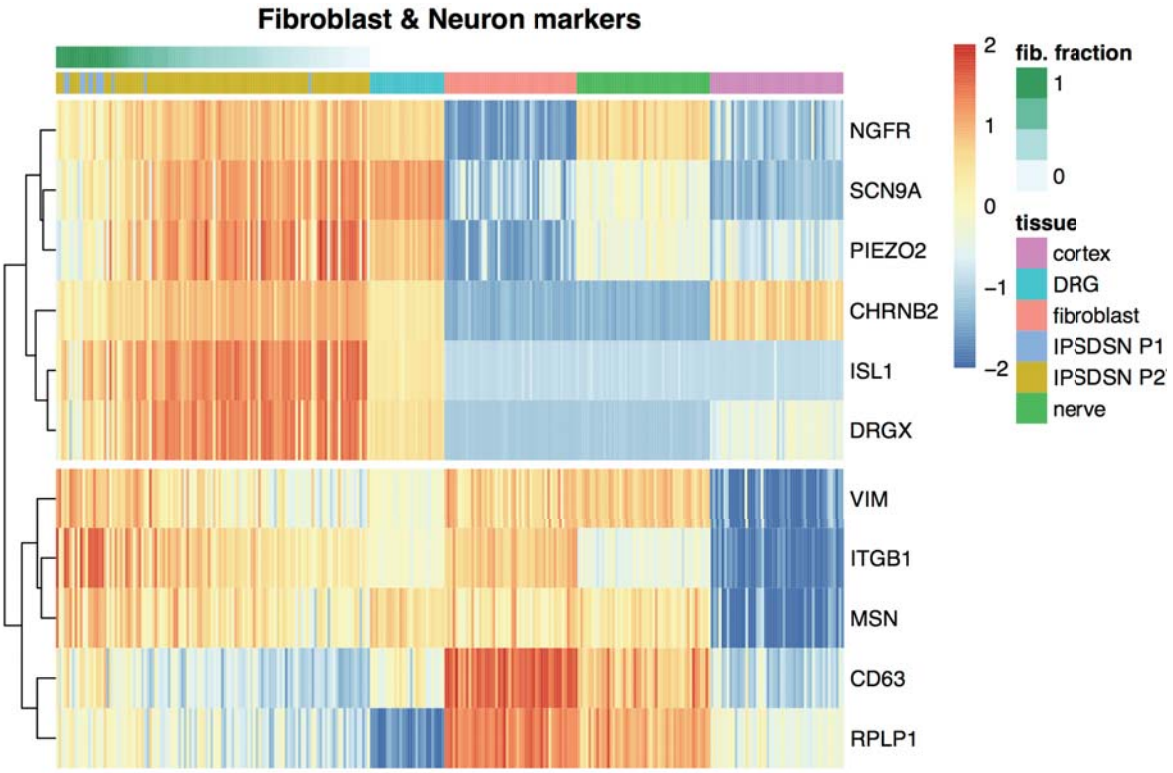


Figure 7: Expression of sensory neuronal and fibroblast marker genes across IPSDSN samples, in comparison with DRG (N=28) and selected GTEx tissues (N=50 each). The overall similarity with DRG is high for neuronal markers, but less so for fibroblast markers. Gene expression was determined as log10(FPKM+0.1), and then was mean-centered and Z-score normalized across samples for each gene.

2.2.3 Heterogeneity in IPSDSN gene expression

A central issue for genetic studies in iPSC-derived cells is heterogeneity of cellular phenotypes. This heterogeneity could arise from donor genetic background, effects of clonal selection, and effects of the cell culture environment during reprogramming and differentiation. Genome-wide gene expression was highly correlated within lines differentiated multiple times (median Spearman ρ=0.96) and reduced slightly between IPSDSNs from different donors (median ρ=0.93) (Figure 8a). However, differentiation

replicates within donor cell lines did not consistently cluster together (Figure 8b), suggesting that variability due to differentiation was at least as large as that due to donor genetic background and iPSC reprogramming together.

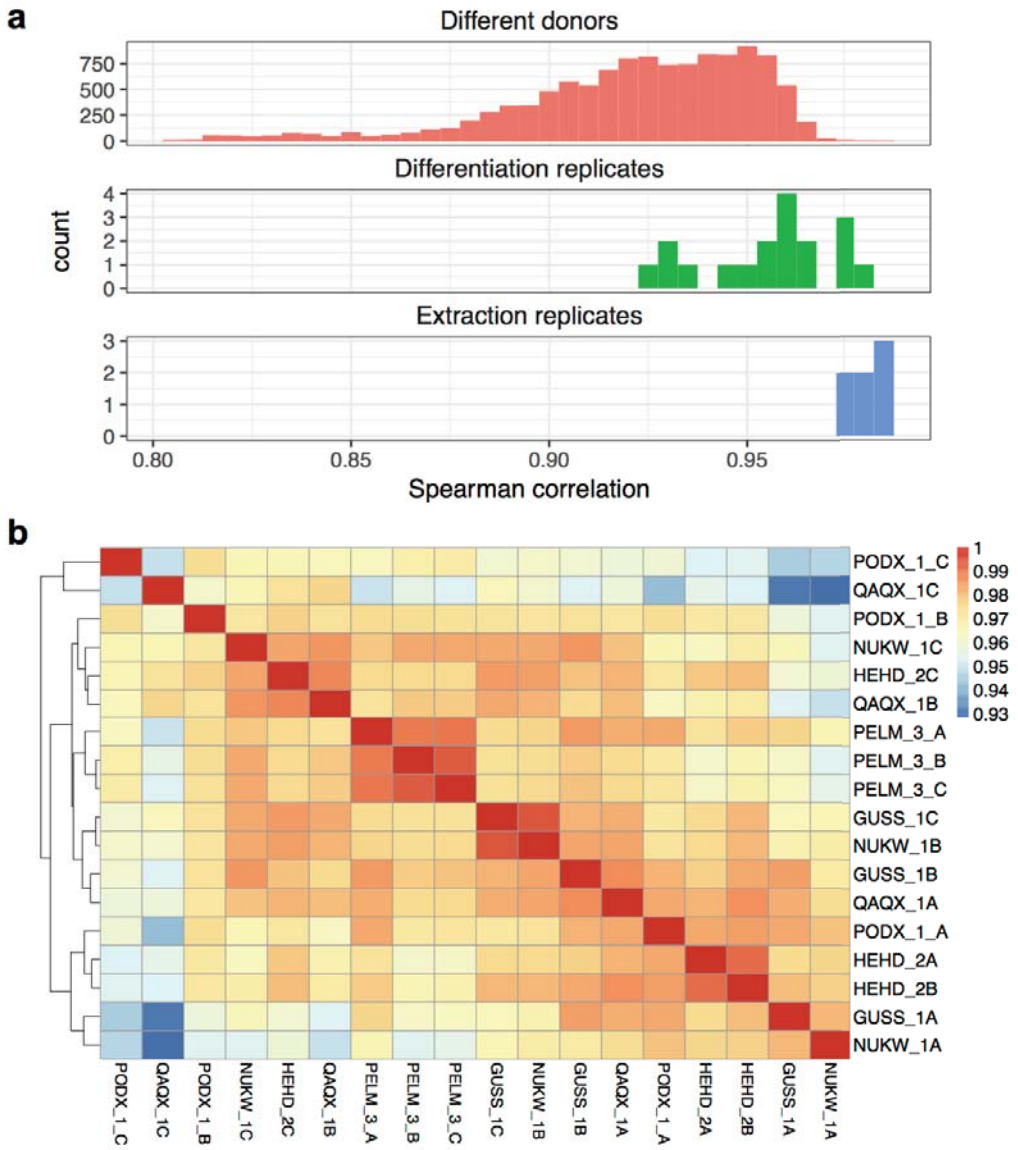


Figure 8: Global gene expression correlations across donors and differentiation replicates. (**a**) Histogram of pairwise spearman correlation coefficients of gene expression among RNA extraction replicates (N=7), differentiation replicates from the same donor (n=6 donors, 3 replicates each), or across donors (n=94). RNA extraction replicates were highly repeatable (spearman ρ of 0.97 - 0.98). Differentiation replicates within a donor cell line were less highly correlated (median ρ=0.96, range 0.93 - 0.98), but had higher correlation than differentiations across donors (median ρ=0.93, range 0.80 - 0.98). (**b**) Clustered heatmap of gene expression correlations between samples having differentiation replicates. Replicates for a given donor do not consistently cluster together.

Although marker genes specific to sensory neurons and nociceptors were expressed (FPKM > 1) in nearly all samples, we observed a high degree of heterogeneity in the level of expression of some genes compared with DRG and other tissues (Figure 9a), despite the fact that a cell culture system is theoretically more pure in cell type composition than a complex tissue. These observations were independent of sample size, and were robust when comparing with DRG samples from unique donors only, rather than all 28 DRG samples. Next, we examined how between-sample variability in global gene expression of IPSDSNs compared with other somatic tissues and cell lines. The distribution of coefficient of variation (CV) of gene expression in IPSDSNs fell within the range of most GTEx tissues (Figure 9b). However, the median CV of gene expression in IPSDSNs (0.37) was considerably higher than in DRG (0.23), indicating that IPSDSNs have greater between-sample variability in expression than the primary tissue they are intended to model.

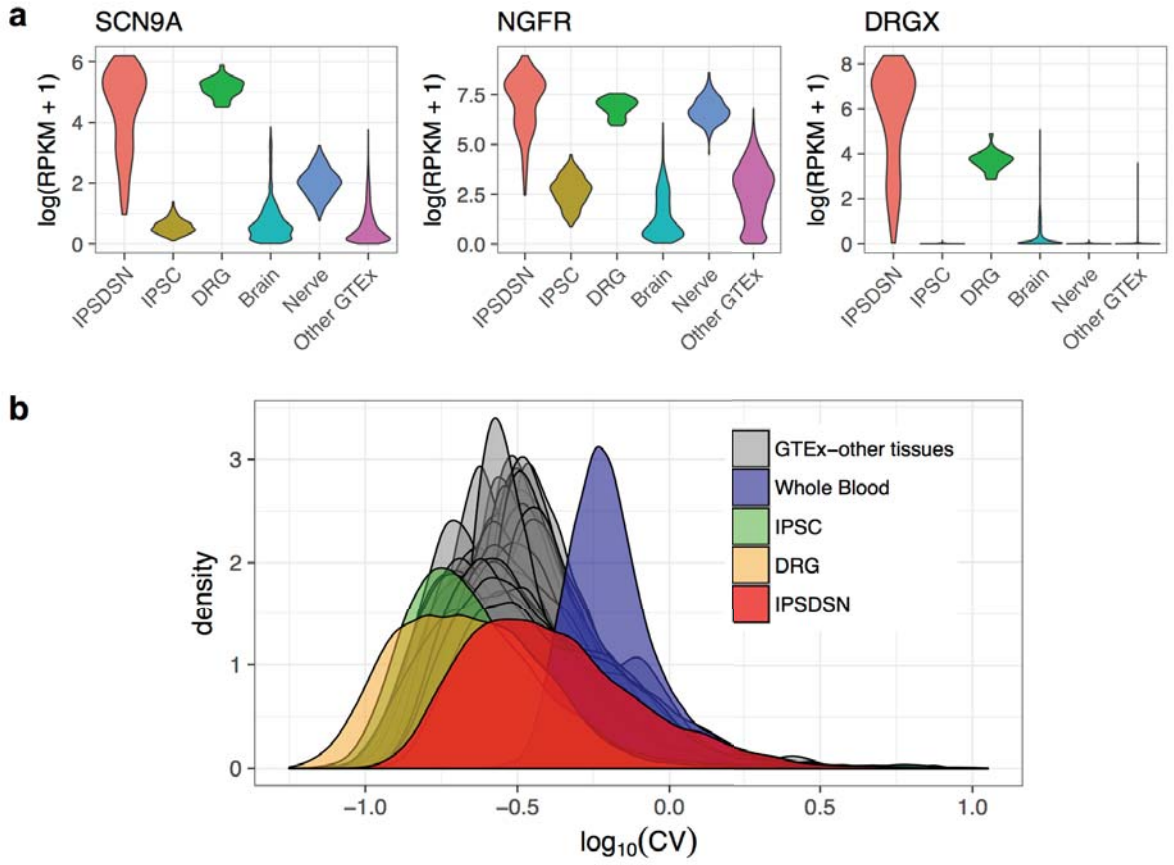


Figure 9: Gene expression is highly variable across IPSDSN cell lines. (**a**) Distribution of expression levels for selected sensory neuronal marker genes in IPSDSN, DRG, GTEx tibial nerve, GTEx brain, and all other GTEx tissues. (**b**) Density plot of the coefficient of variation of genes across samples, separately for each GTEx tissue, IPSDSN samples (n=106, P2 protocol only), iPSC (n=200), and DRG (n=28).

Highly variable genes in IPSDSNs were enriched for function in neuronal differentiation and development (Supplementary Table 4). Genes that were significantly upregulated between iPSCs and IPSDSNs, which will include those essential for sensory neuronal function, were also more variable than remaining genes (Figure 10). Importantly, we did not observe similar levels of expression variability of neuronal or developmental gene groups in DRG, iPSCs, or GTEx nervous tissues (Figure 11).

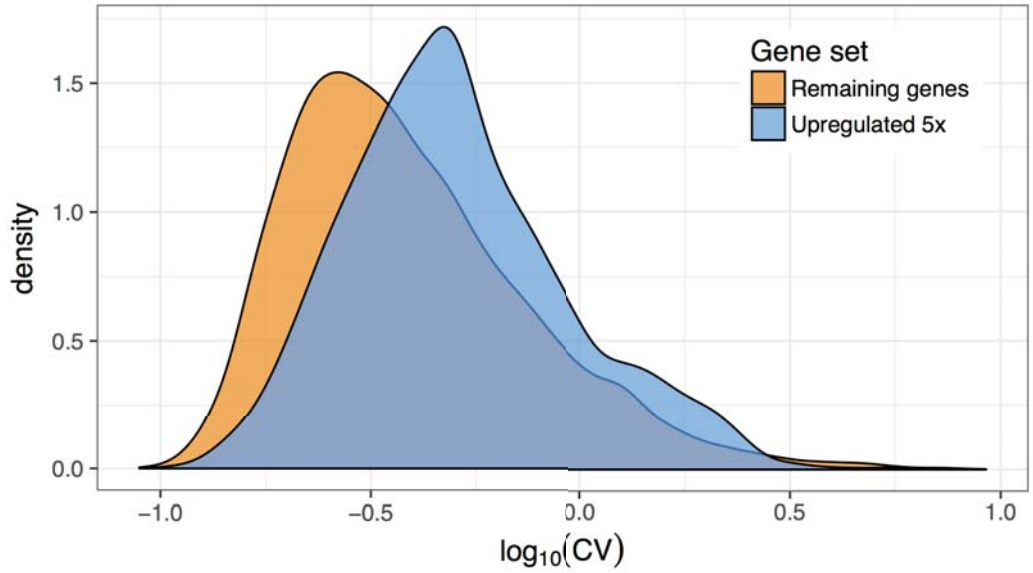


Figure 10: Genes upregulated at least 5-fold upon differentiation from iPSC to IPSDSN (FDR < 1%, N=4,246 genes) have increased variability relative to the remaining genes, despite similar levels of expression (median/mean FPKM of upregulated genes: 4.1 / 15.6; remaining genes: 4.6 / 11.8).

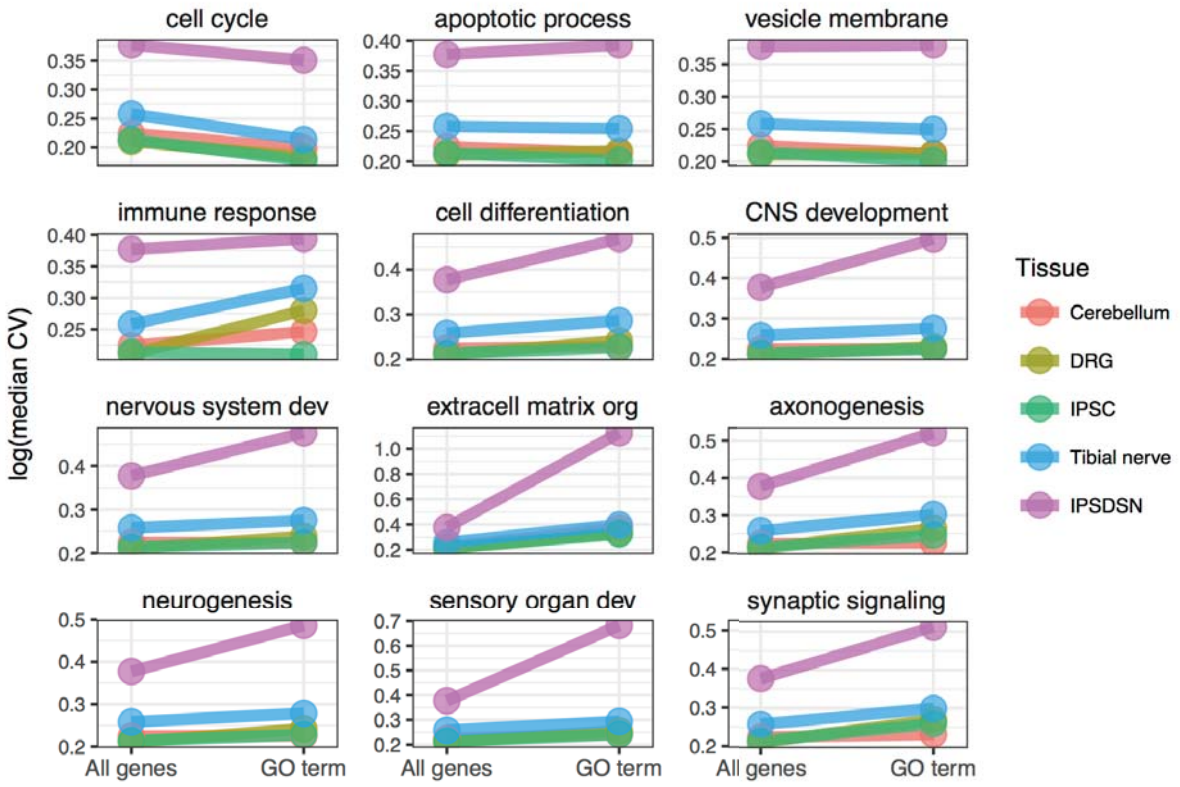


Figure 11: Median variability of genes in specific GO terms is compared to median variability for all genes, separately for IPSDSNs, iPSCs, DRG, and GTEx cerebellum and tibial nerve. Genes categories related to neuronal function and differentiation have increased median variability in IPSDSNs relative to all genes, but this is much less the case for iPSCs or nervous tissue samples, and for other gene categories such as immune response or cell cycle genes.

These results highlight that expression of neuronal genes varies substantially more in IPSDSNs than in somatic nervous tissue, probably as a result of variability in differentiation. Consistent with this, variance components analysis (Figure 12) showed that as much or more variation was explained by differentiation batch (median 24.7%) as donor/iPSC line of origin (median 23.3%), which would include both donor and reprogramming effects.

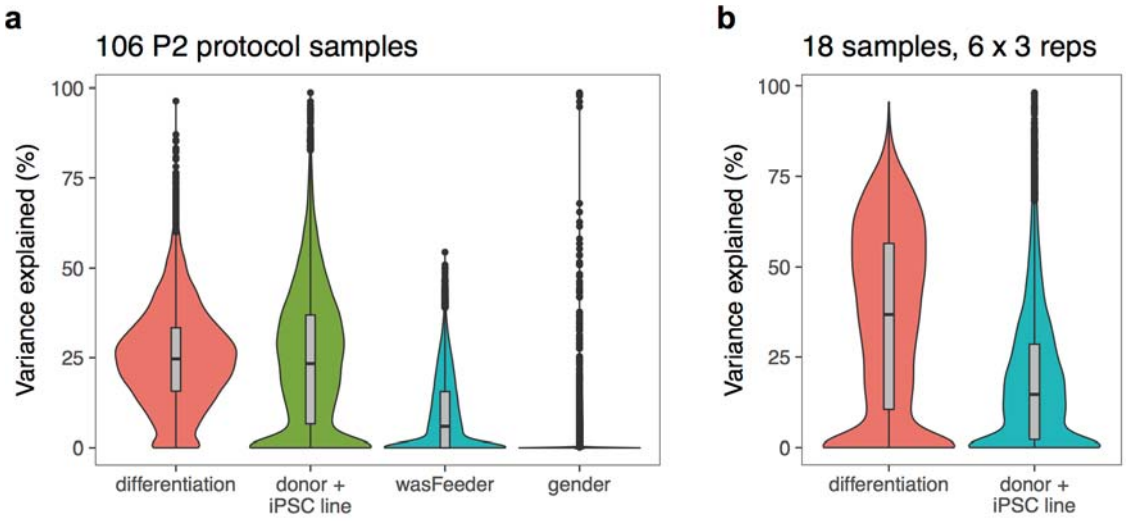


Figure 12: Variance components analyses of IPSDSN gene expression. (**a**) Variance partitioning for 119 samples (13 P1 protocol, 106 P2 protocol). (**b**) Variance partitioning for 18 samples, from 6 iPSC lines differentiated 3 times each. All 18 samples were from E8-medium iPSC lines derived from females, and were differentiated with the P2 protocol.

2.2.4 iPSC culture conditions influence cell fate

Intriguingly our variance components analysis suggested that, although the cell lines for this analysis were differentiated using an identical protocol, starting iPSC cell culture conditions influenced gene expression patterns in the IPSDSNs produced four weeks later (Figure 13a). Of the 106 successful P2 protocol differentiations, 27 were from iPSCs maintained on mouse embryonic fibroblast (MEF) feeder cells (feeder-iPSCs), while the remaining 79 were grown in Essential 8 medium (E8-iPSCs). The first principal component (PC) of iPSC gene expression clearly differentiated feeder- and E8-iPSCs (Figure 13a), indicating that culture conditions are among the largest global effects on transcription. Similarly, PC1 of gene expression in IPSDSNs distinguished samples originating from feeder- and E8-iPSCs; moreover, IPSDSNs from E8-iPSCs had higher neuronal content (Figure 13b, 28% higher for E8-iPSCs, t-test $p=1.84 \times 10^{-5}$). A possible technical explanation for these results is that protocol implementation and batch effects changed subtly over the course of the project. However, the difference in neuronal content between IPSDSNs derived from E8 or feeder-iPSCs remained when sample derivation date was included as an explanatory covariate (linear regression $p=6.5 \times 10^{-4}$, 36% higher for E8-iPSCs, Figure 13c).

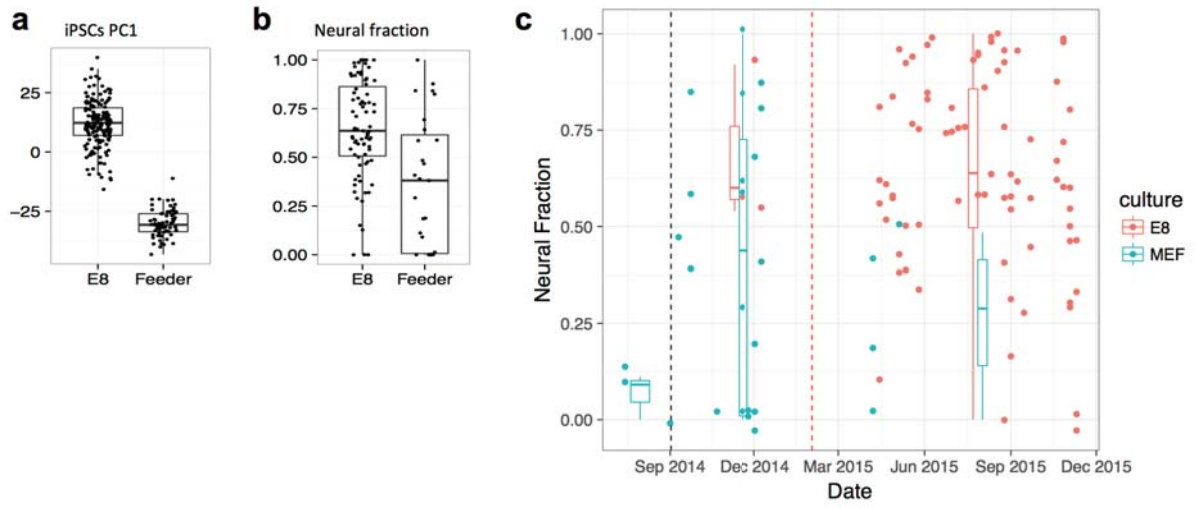


Figure 13: Neural fraction and global gene expression in IPSDSNs is influenced by iPSC culture conditions. (**a**) Global gene expression differences between feeder- and E8-iPSCs are captured in PC1. (**b**) Estimated neural fraction of samples differs in IPSDSNs derived from feeder- and E8-iPSCs. (**c**) Neural fraction in IPSDSN samples varies with date and iPSC culture condition. Only P2 protocol samples are shown, and each dot represents the estimated neural fraction for one sample. We used linear models in R to estimate the effects of culture conditions or differentiation date on neural fraction. We either used date as a continuous variable, or split differentiation date into 2 bins (red line) or 3 bins (black and red lines). In all models including both factors, culture conditions were more strongly associated with neural fraction than was date, and the association of culture conditions remained significant. In contrast, the association of date with was not significant with date as a continuous variable or split into 2 bins ($p>0.3$), and was marginally significant with date split into 3 bins ($p=0.038$).

Next, we determined genes that were differentially expressed between E8- and feeder-iPSCs and IPSDSNs (Figure 14). Genes more highly expressed in feeder-iPSCs were strongly enriched for mesenchyme development, stem cell differentiation, and Wnt and TGF-β signalling, while genes more highly expressed in E8-iPSCs showed less clear enrichment (Supplementary Tables 5-7). Notably, inhibition of TGF-β/SMAD signalling is a key step in sensory neuronal differentiation. Top differentially expressed genes include early developmental regulators such as *EMX1* (15-fold higher in E8-iPSCs), important for specific neuronal cell fates, and *BMP2* (13-fold higher in feeders), which has been shown to suppress differentiation to sensory cell fates by antagonizing Wnt/beta-catenin (Kléber et al. 2005) (Figure 14b). In addition, *SCN9A* and *TAC1*, key markers of sensory neurons, were expressed at low levels in iPSCs, with 2.2-fold and 2.9-fold higher expression in E8-iPSCs. We also considered genes differentially expressed between IPSDSNs derived from E8- and feeder-iPSCs (Figure 14c). Genes more highly expressed in IPSDSN samples from feeder-iPSCs were overrepresented in extracellular matrix components, pattern specification, organ

morphogenesis, and Wnt signalling (Supplementary Tables 8-10), and include *FGFR2*, *BMP7*, and *WNT5A* (Figure 14d). Genes more highly expressed in IPSDSN samples from E8-iPSCs were overrepresented in ion channel complexes, peripheral nervous system development, and synapse organisation, and include *SCN9A*, *DRGX*, and *CACNA1A*. These differences likely reflect the increased neuronal content of samples from E8-iPSCs. Together these results suggest that iPSCs are primed towards different cell fates depending on the iPSC culture medium.

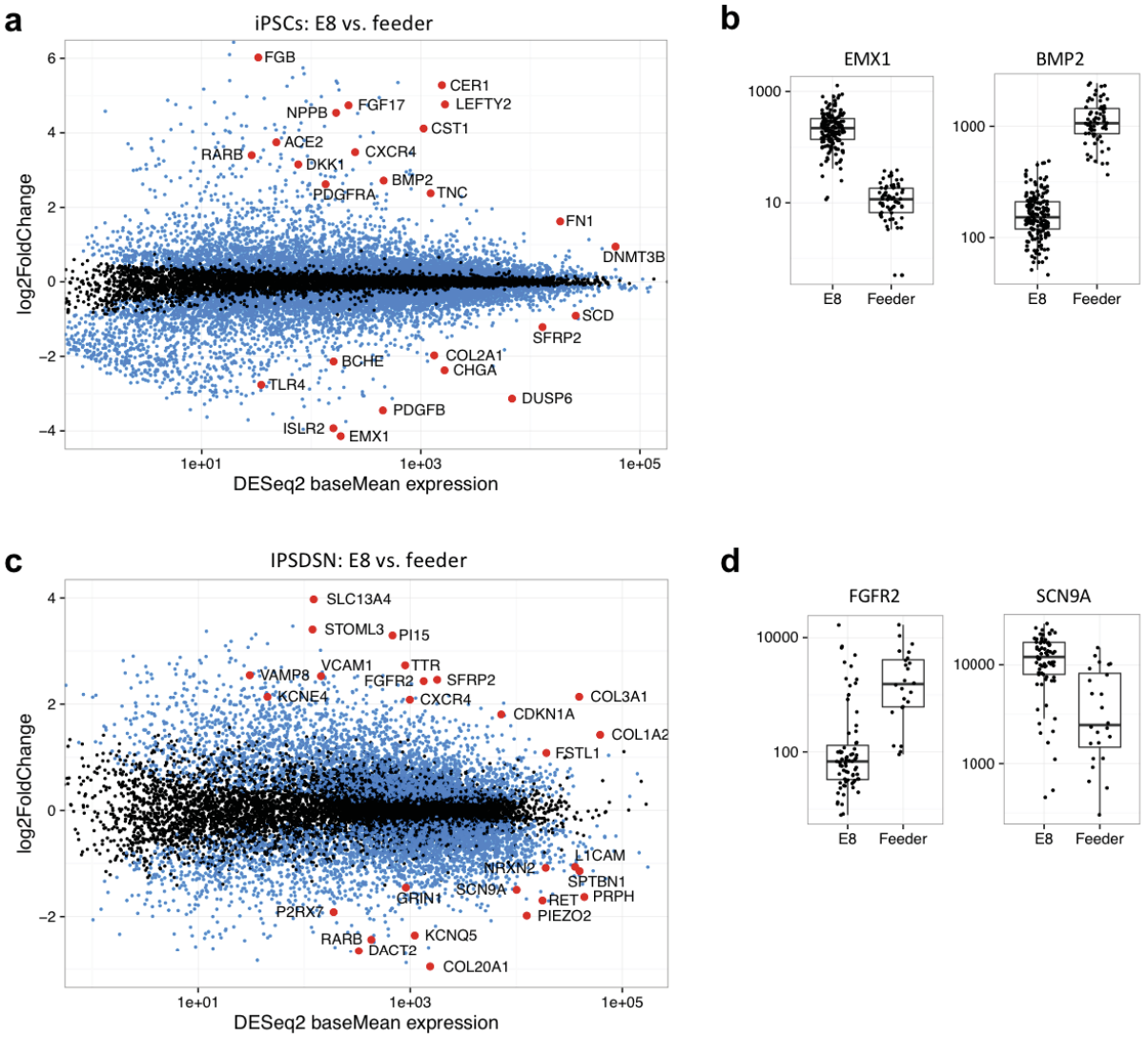


Figure 14: (**a**) Differentially expressed genes (FDR 1%, blue and red points) between iPSC samples grown on feeders (n=68) vs. E8 medium (n=171). (**b**) Barplots of selected genes differentially expressed between feeder- and E8-iPSCs. (**c**) Differentially expressed genes (FDR 1%) between IPSDSNs from feeder- (n=27) and E8-iPSCs (n=79). Neuronal differentiation genes, such as *RET* and *L1CAM*, are more highly expressed in samples from E8-iPSCs. (**d**) Barplots of selected genes differentially expressed between IPSDSNs derived from feeder- and E8-iPSCs.

Since iPSC culture conditions influenced differentiation outcomes, we examined gene expression variability within subsets of IPSDSN samples. IPSDSNs differentiated from feeder-iPSCs had somewhat higher global gene expression variability, yet those from E8-iPSCs were still highly variable relative to DRG and iPSCs (Figure 15), with neuronal and developmental gene sets enriched for highly variable genes (Supplementary Table 11). Among the 79 IPSDSNs from E8-iPSCs, samples with high fibroblast content had somewhat higher variability, but those with low fibroblast content still showed high variability relative to DRG and iPSCs.

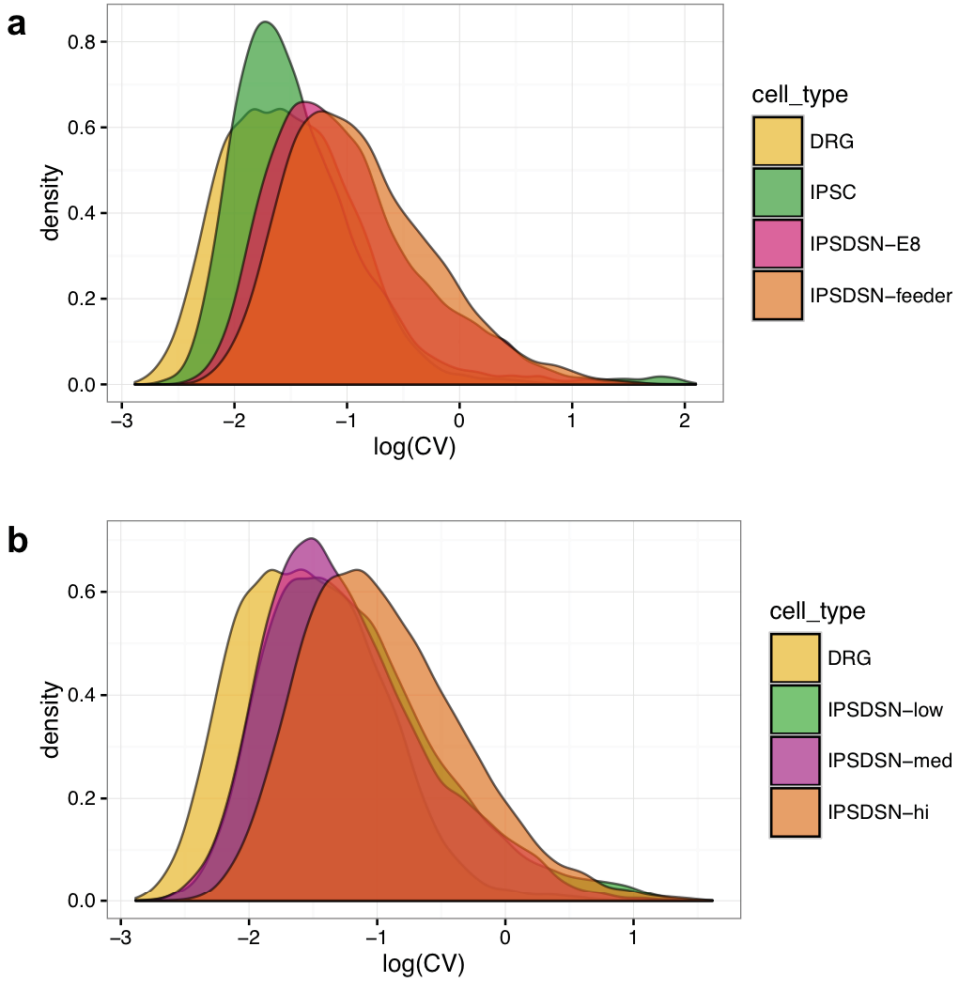


Figure 15: The natural log of the coefficient of variation across samples is plotted for different subgroups of samples. **(a)** Samples from feeder-iPSCs (N=27) have slightly higher variability than those from E8-iPSCs (N=79), but samples from E8-iPSCs still have much higher variability than iPSC or DRG. **(b)** Comparing DRG and P2 protocol IPSDSN samples from E8-iPSCs only, separated based on fibroblast-like content: low (estimated < 20%, N=24), medium (20-50%, N=31), and high (> 50%, N=24). High fibroblast-like samples have slightly higher CV across all genes, but this accounts for only a fraction of the increased variability seen relative to primary DRG.

2.2.5 Genetic variants influence gene expression, splicing and chromatin accessibility in sensory neurons

Using the linear model FastQTL (Ongen et al. 2016), we mapped 1,403 expression quantitative trait loci (eQTLs) at FDR 10%, of which 746 were expressed at a moderate level (FPKM > 1). We noted that we discovered many fewer eQTLs than in GTEx tissues of comparable sample size (Figure 16a). This suggested that power for eQTL discovery was lower in IPSDSNs than somatic tissues, possibly due to additional variability introduced by differentiation. Using the allele-specific method RASQUAL (Kumasaka, Knights, and Gaffney 2016) we detected 3,778 genes with expression-modifying genetic variants, termed eGenes, at FDR 10% (Supplementary Table 12), with 2,607 of these expressed at FPKM > 1. Notably, it was only using the additional information from allele specific signals that we achieved approximately similar statistical power to GTEx tissues with equivalent sample sizes.

To identify eQTLs that were not already reported in GTEx (v6), we used a protocol described previously for the HIPSCI project (Kilpinen, Goncalves, et al. 2016). Of all 3,778 eGenes, 954 had tissue-specific associations (Supplementary Table 15), including genes with known involvement in pain or neuropathies, such as *SCN9A*, *GRIN3A*, *P2RX7, CACNA1H*/Cav3.2, and *NTRK2*. Because these novel eQTLs were not seen in any GTEx tissue, this suggests that these regulatory variants may have IPSDSN-specific function.

We investigated whether the improvement in power from using allele-specific information was related to gene expression variability. Splitting genes into quartiles of expression variability revealed that power improvement was greatest among genes with high variability across samples (Figure 16b,c). The fraction of novel eQTLs increased for both fastQTL and RASQUAL as gene variability increased (Figure 16d), with RASQUAL overall finding a higher fraction of novel eQTLs. One explanation would be a higher false positive rate for RASQUAL; however, various properties of the novel eQTLs did not differ significantly from known eQTLs, including expression levels, and eQTL variant allele frequency, hardy-weinberg equilibrium, and mapping bias. In addition, the rate of novel eQTLs increased only moderately from the least to the most highly variable genes, and the trend was similar for FastQTL and for RASQUAL. It is possible that relative to GTEx, which used a linear model, RASQUAL finds true eQTLs that are more difficult to discover without examining allele-specific expression.

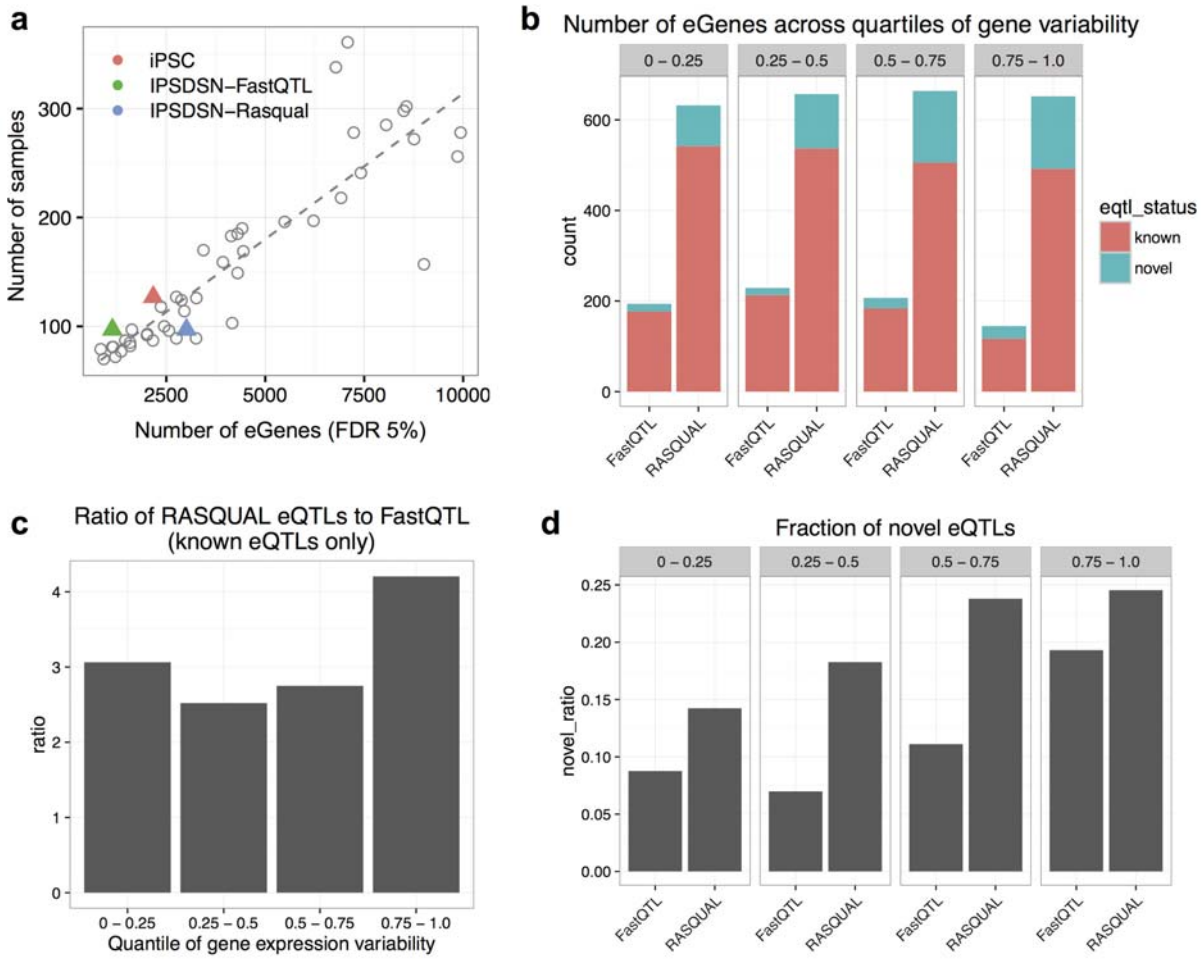


Figure 16: (**a**) Number of eGenes discovered for IPSDSNs using either FastQTL (IPSDSN-FastQTL) or RASQUAL (IPSDSN-Rasqual), in relation to sample size, compared with GTEx tissues. (**b**) Number of eGenes discovered by RASQUAL and FastQTL across quartiles of gene expression variability. The bottom 25% of eGenes ranked by sample-to-sample expression variability are at the left, while the top 25% are at the right. Bar colors show the number of eQTLs that overlap with a GTEx eQTL ("known") or where no GTEx tissue has p < 0.01 for our lead eQTL SNP ("novel"). (**c**) Restricting to known eQTLs discovered by either method, the ratio of the number of eGenes discovered by RASQUAL to FastQTL is highest in the highest quartile of gene expression variability, although power gains from RASQUAL are high across the board. **(d)** The fraction of novel eQTLs, separately for fastQTL and RASQUAL, across quartiles of gene expression variability.

Variants affecting gene splicing (sQTLs) often change either protein structure or context-dependent gene regulation, and may be more enriched for complex trait loci than are eQTLs (Y. I. Li et al. 2016). To detect sQTLs we used the annotation-free method LeafCutter (Y. I. Li, Knowles, and Pritchard 2016) to define 30,591 clusters of alternatively spliced introns. Using FastQTL (Ongen et al. 2016) we discovered QTLs for 2,079 alternative splicing clusters at FDR 10% (Supplementary Table 13). Notably, only 538 (26%) of the lead variants for these splicing associations were in linkage disequilibrium (LD) r^2 >= 0.5 with a lead eQTL

variant in our dataset, indicating that the sQTLs extend our catalog of expression-altering variants and are not merely proxies for gene-level eQTLs (or vice versa).

	Number	GWAS overlap
eQTLs	3778	156
sQTLs	2079	129
ATAC QTLs	6318	172
Joint ATAC/eQTLs	177	14

Table 1: QTL associations. Columns show the number of associations, and the number of unique overlaps ($r^2 > 0.8$) between lead QTL SNPs and GWAS catalog SNPs after removing duplicates for each GWAS trait.

We collected ATAC-seq data for 31 samples (Buenrostro et al. 2013) and used this to identify active regulatory regions in IPSDSNs and to map 6,318 caQTLs chromatin accessibility QTLs (caQTLs) at FDR 10% (Supplementary Table 14). To identify transcription factors in IPSDSNs whose binding is altered by regulatory variants, we used the LOLA Bioconductor package (Sheffield and Christoph 2015) to test for enrichment of our lead QTL SNPs, relative to GTEx lead SNPs, in ENCODE ChIP-seq peaks and JASPAR transcription factor motifs (Supplementary Tables 16,17). Tissue-specific eQTLs were highly enriched within SMARCB1 and SMARCC2 peaks (odds ratios 5.8 and 14.1; $p < 5x10^{-5}$), which are both members of the neuron-specific chromatin remodeling (nBAF) complex (Lessard et al. 2007). Also enriched were REST/NRSF (OR=5.7, $p=1.1x10^{-4}$) and SIN3A (OR=3.9, $p=1.0x10^{-4}$), which bind neuron-restrictive silencer elements during development, but have suggested roles in the development and maintenance of neuropathic pain (Willis et al. 2016). Considering all IPSDSN eQTLs, we found enrichments for ELK1 and ELK4, as well as c-Fos, a target of ELK1 and ELK4 which is widely expressed but is known to have specific functions in sensory neurons (Hunt, Pini, and Evan 1987; Kohno et al. 2003). Notably, DNA sequence motifs for REST, ELK1 and ELK4 are also among the most highly enriched motifs in our ATAC-seq peaks (Supplementary Table 18).

2.2.6 Sensory neuron eQTLs and sQTLs overlap with complex trait loci

While we were interested in comparing our set of QTLs with GWAS for pain, the largest GWAS for pain to date included just 1,308 samples and found no associations at genome-wide significance (Peters et al. 2013). We therefore considered all GWAS catalog associations with $p < 5x10^{-8}$ that were in high LD ($r^2 > 0.8$) with a QTL in our dataset, with

two purposes in mind: to determine whether any GWAS traits are enriched overall for overlap with sensory neuron QTLs, and to find individual cases where a QTL is a strong candidate as a causal association for the GWAS trait. Overall, IPSDSN eQTLs were significantly enriched for overlap with GWAS catalog SNPs ($p < 0.001$) relative to 1000 random sets of SNPs matched for minor allele frequency (MAF), distance to nearest gene, gene density, and LD (Pers, Timshel, and Hirschhorn 2015), and the overlap was consistent with that seen for eQTL studies in other tissues (Figure 17). Although nociceptive neurons are specialized for sensing and relaying pain signals, they share characteristics with other neurons; thus, we might expect enrichment for traits known to involve the nervous system more generally. However, among the 41 traits with at least 40 GWAS catalog associations, we could not detect any trait with significantly greater overlap with our QTL catalog than other traits after correcting for multiple testing (Supplementary Table 19).

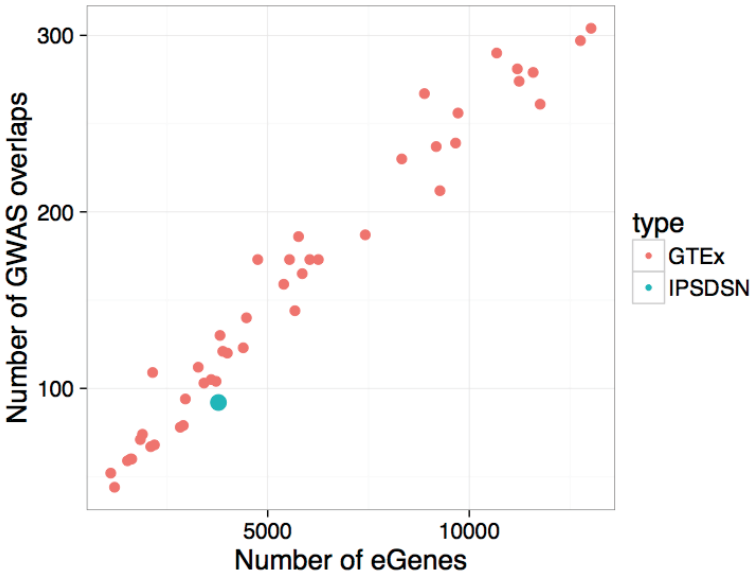


Figure 17: The number of overlaps at LD $R^2 > 0.8$ between eQTLs in IPSDSNs and GWAS catalog SNPs is within the range seen for similarly powered tissues in GTEx. The number of eQTL-GWAS overlaps is heavily dependent on the number of eQTLs discovered, which is reflected in the tight linear relationship with the number of eGenes.

Across all traits, we found 156 genes with an eQTL overlapping at least one GWAS association, and similarly 129 sQTLs and 172 caQTLs with GWAS overlap (full catalog in Supplementary Tables 20-22). We examined individual associations, in conjunction with ATAC-seq peaks and LD information, to identify candidate causal variants influencing both a molecular phenotype and a complex trait. For most of these associations we do not expect that sensory neurons are the most relevant cell type; rather the overlaps may reflect either

general neuronal mechanisms or non-cell-type-specific functions. We thus focused on traits where neurons are likely to be a relevant cell type.

Among overlapping associations we found a number that relate to neuronal diseases, such as Parkinson's disease, multiple sclerosis, and Alzheimer's disease. One striking overlap is between an eQTL for *SNCA*, encoding alpha synuclein, and Parkinson's disease, for which a likely causal variant has recently been identified (Soldner et al. 2016). The lead GWAS SNP and our lead eQTL are both in perfect LD with rs356168 (1000 genomes MAF 0.39), which lies in an ATAC-seq peak in an intron of *SNCA*. Soldner et al. used CRISPR/Cas9 genome editing in iPSC-derived neurons to show that rs356168 alters both *SNCA* expression and binding of brain-specific transcription factors (Soldner et al. 2016). In IPSDSN cells we find that the G allele of rs356168 increases *SNCA* expression 1.14-fold, in line with Soldner et al. who reported 1.06- to 1.18-fold increases in neurons and neural precursors. However, despite residing in a visible ATAC-seq peak in our data, rs356168 is not detected as a caQTL (SNP p value = 0.22). eQTLs for SNCA have recently been reported in the latest GTEx release (v6p), but none of the tissue lead SNPs are in LD ($r^2 > 0.2$) with rs356168, suggesting that the effect of this SNP can be more readily detected in specific cell and tissue types, including IPSDSNs and the frontal cortex tissue and iPSC derived neurons studied by Soldner et al.

We also find multiple compelling overlaps between splice QTLs and GWAS associations (Figure 18). One known example is a strong sQTL for *TNFRSF1A* ($p=9.9x10^{-29}$) with the same lead SNP (rs1800693, MAF 0.30) as a multiple sclerosis association. This likely causal SNP is located 10 base pairs from the donor splice site downstream of exon 6, and has been experimentally shown to cause skipping of exon 6, which results in a truncated, soluble form of TNFR1 that appears to reduce TNF signalling (Gregory et al. 2012). TNFRSF1A is highly expressed (>15 FPKM) in both IPSDSNs and in DRG. We do not see an effect of this variant

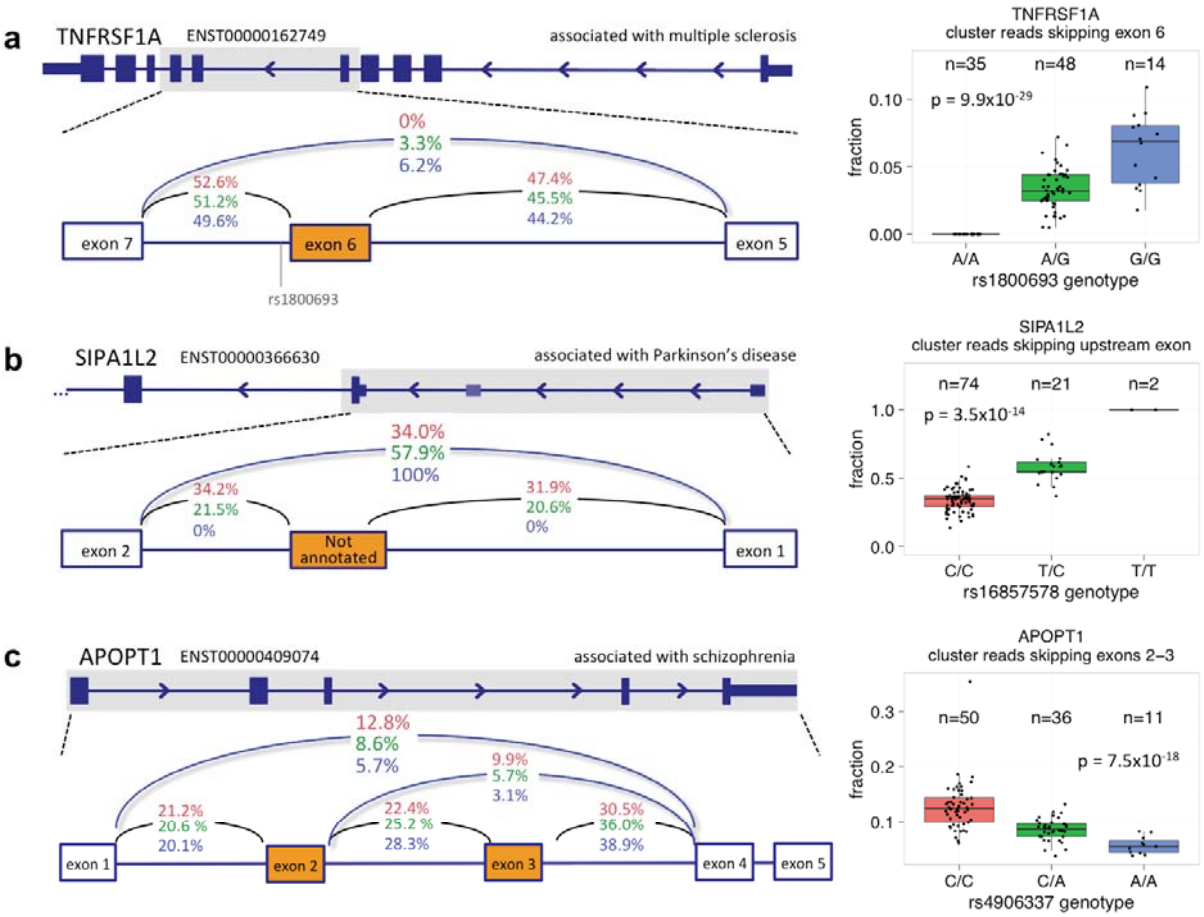


Figure 18: Splicing QTLs overlapping GWAS. **(a)** An sQTL for *TNFRSF1A* leads to skipping of exon 6, and overlaps with a multiple sclerosis association. **(b)** An sQTL for *SIPA1L2* leads to increased skipping of an unannotated exon between alternative promoters, and overlaps with a Parkinson's disease association. **(c)** An sQTL for *APOPT1* alters skipping of exons 2 and 3, and overlaps with a schizophrenia association. P values are from the beta approximation based on 10,000 permutations as reported by FastQTL.

on total expression levels in our cells ($p > 0.5$), but we observe skipping of exon 6 in about 12% of transcripts from individuals homozygous for rs1800693 (Figure 18a). Since these transcripts undergo nonsense-mediated decay (Gregory et al. 2012), the actual rate of exon skipping is likely to be higher. Given the broad role of TNF in inflammation and immunity, it is interesting that rs1800693 is associated with MS but not with other autoimmune disorders, apart from primary biliary cirrhosis (Gregory et al. 2012). Moreover, whereas TNF inhibitors are effective in many autoimmune disorders, they exacerbate MS, an effect that is mimicked by the reduction in TNF signalling produced by the *TNFRSF1A* splice variant. These observations suggest an interplay between cells of the CNS and immune system involving TNF signalling. TNF signalling has been shown to have both inflammatory and neuroprotective effects in the CNS and, despite a large body of research, the exact

mechanisms and cell types responsible for the genetic risk associated with TNF receptor polymorphisms remain unclear (Probert 2015).

An sQTL for *SIPA1L2* (rs16857578, MAF 0.23) is in LD with associations for both Parkinson's disease (rs10797576, r^2=0.93) and blood pressure (rs11589828, r^2=0.94). An unannotated noncoding exon (chr1:232533490-232533583) between alternative *SIPA1L2* promoters is included in nearly 50% of transcripts in individuals with the reference genotype, but splicing in of the exon is abolished by the variant (Figure 18b). SIPA1L2, also known as SPAR2, is a Rap GTPase-activating protein expressed in the brain and enriched at synaptic spines (Spilker, Christina, and Kreutz 2010). Although its function is not yet clear, expression is seen in many tissues profiled by GTEx, with highest expression in the peripheral tibial nerve. Interestingly, the related protein SIPA1L1 exhibits an alternative protein isoform with an N-terminal extension that is regulated post-translationally to influence neurite outgrowth (Jordan et al. 2005).

A complex sQTL for *APOPT1* (rs4906337, MAF 0.22) is in near-perfect LD with a schizophrenia association (rs12887734). The splicing events involve skipping either of exon 3 only or both exons 2 and 3 (Figure 18c). At least 20 variants are in high LD ($r^2 > 0.9$), including rs4906337 which is 40 bp from the exon 3 acceptor splice site, and rs2403197 which is 63 bp from the exon 4 donor splice site. No sQTL is reported in GTEx, and although eQTLs are reported for *APOPT1*, only the thyroid-specific eQTL (rs35496194) is in LD (r^2 = 0.94) with the schizophrenia-associated SNP rs12887734. APOPT1 is localized to mitochondria and is broadly expressed. Homozygous loss-of-function mutations in this gene lead to Cytochrome c oxidase deficiency and a distinctive brain MRI pattern showing cavitating leukodystrophy in the posterior region of the cerebral hemispheres, with affected individuals having variable motor and cognitive impairments and peripheral neuropathy (Melchionda et al. 2014).

2.2.7 Recall by genotype studies in iPSC-derived cells will require large sample sizes

One attractive future use of iPSCs is to experimentally characterise GWAS loci using a "recall by genotype" approach. Here, iPSC lines with specific genotypes are chosen from a large bank and differentiated into target cell types (for example, see (Warren et al. 2017)). Our observations suggested that, for certain protocols, the additional cellular heterogeneity introduced by differentiation could impact the power of these studies to detect the effects of common genetic variants. Importantly, our large set of differentiations gave us accurate

genome-wide estimates of effect size and expression variability in an iPSC-derived cell type, for use as a benchmark “ground truth”. We investigated the performance of iPSC-based recall by genotype studies by bootstrap resampling from a stringent (FDR 1%) IPSDSN eQTL call set. For each eQTL gene we sampled expression counts from an equal number of major and minor homozygotes for the lead SNP, sampling with replacement to achieve a specific sample size. We then estimated power as the fraction of 100 bootstrap replicates where we found a significant difference ($p < 0.05$, Wilcoxon rank sum test) in expression between the homozygotes.

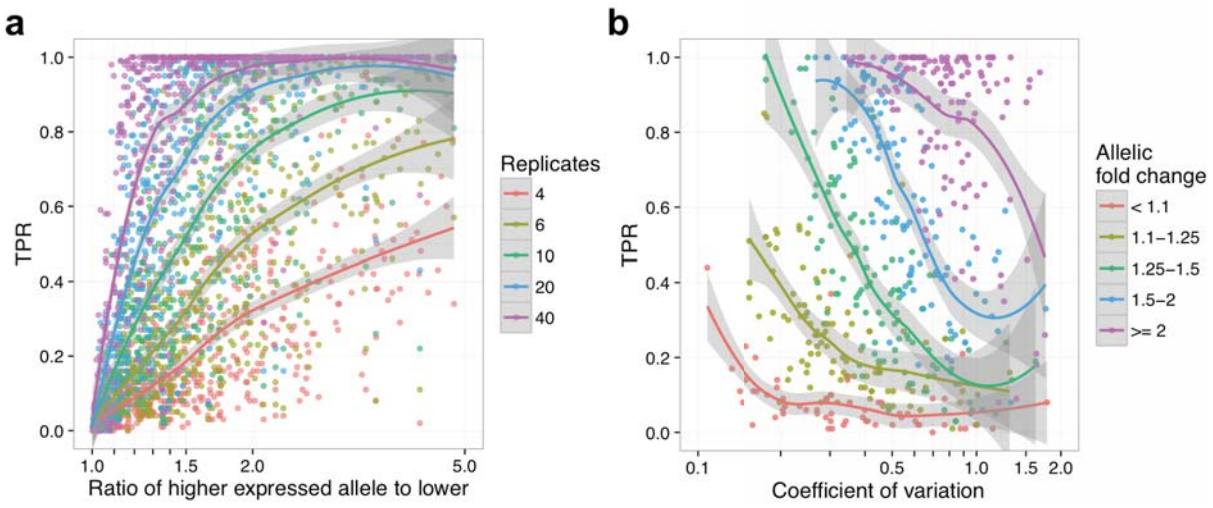


Figure 19: Power to detect a genetic effect in a single-variant single-gene test depends on sample size, allelic effect size, and gene expression variability. (**a**) TPR as a function of allelic fold change for five different numbers of replicates (half the total sample size). (**b**) TPR as a function of CV for five bins of allelic fold change, with 10 samples of each genotype.

Our results illustrate important trends. First, recall by genotype studies in iPSC-derived cells are likely to require relatively large sample sizes, typically 20-80 unrelated individuals, for variants with a 1.5-2-fold effect size (Figure 19a). Second, as expected, highly variable genes are more challenging (Figure 19b) with power below 40% in a sample size of 20 for even moderately variable genes (CV 0.5 - 0.75). While expression noise will not typically be known accurately a priori, an estimate of effect size may be available from previous eQTL studies in specific tissues. This could enable estimating the number of samples needed to achieve a desired power.

Note that these power estimates assume that a single gene is being tested, which is only likely to be the case when there is a very strong prior belief in the causal gene and few genes in the region. Where multiple genes are tested, power will be lower. These results

also suggest that large sample sizes will be required when using genome editing to identify causal GWAS-associated variants: although genetic background can be controlled in such an experiment, differentiation noise will continue to be a major contributor to gene expression variability.

2.3 Discussion

iPSC-derived cells enable the molecular mechanisms of disease to be studied in relevant human cell types, including those which are inaccessible as primary tissue samples. Because the effect sizes of common disease-associated risk alleles tend to be small, observing their effects in cellular models is challenging (Raghavan et al. 2016; Soldner et al. 2016). In an iPSC-based system, this difficulty is compounded by variability between samples in the success of differentiation, as described for hepatocytes (Dianat et al. 2013), hematopoietic progenitors (Smith et al. 2013), and neurons (Hu et al. 2010; Handel et al. 2016).

Our study is the first that we are aware of to perform iPSC differentiation to a neuronal cell type and functionally characterise the resulting cells at scale. Sample-to-sample variability in gene expression in the iPSC-derived cells was greater than in DRGs, with highly variable genes enriched in processes relating to neuronal differentiation and development. This highlights that genes likely to be of particular interest and relevance for the function of these cells are also among the most variable, a challenge which may be broadly true of iPSC-derived cells. Despite the observed variability, we detected thousands of eQTLs, sQTLs, and caQTLs in IPSDSNs, most of which were discovered only with a model that statistically combines both allele-specific and between individual differences in expression to improve power for association mapping. Some of these overlap known expression-modifying variants that are associated with disease, such as an eQTL for *SNCA* associated with Parkinson's disease. However, for most of these disease overlaps the causal variants are not known. This QTL map is thus a starting point for in-depth dissection of individual loci in iPSC-derived neurons where we have shown that a genetic effect is present.

Although our study highlights the potential power of IPSC derived cells as model systems for studying human genetic variation, our results also illustrate the limitations of this approach. First, despite expressing key marker genes and exhibiting neuronal morphology and electrophysiology, it is clear from our data that IPSDSNs are transcriptionally distinct from the other cell types we examined, including DRGs. This reflects a limitation of existing *in vitro* differentiation protocols, which produce cells that are not as functionally or transcriptionally mature as primary tissues. Second, our differentiations did not produce pure

populations of neurons, nor could we measure the purity of the resulting cultures precisely. A portion of the sample-to-sample variability that we observed is likely due to this mixture of cell types, which varied across differentiations. Although mature neurons can be labeled for marker genes, they are not easily sorted by automated systems, which limits the high-throughput options available for purifying neuronal populations. As a result, the eQTLs that we discovered do not represent those of a pure sensory neuronal cell type. For many cell types, sorting is more feasible, and could provide one solution to the variable maturity and heterogeneity of differentiated cell populations.

We used single-cell RNA-seq from three differentiation batches to characterise IPSDSN heterogeneity, which showed that they cluster into neuronal cells and cells with more fibroblast-like gene expression. Using reference profiles from these clusters enabled us to estimate a proxy measure of neuronal cell purity in our bulk RNA-seq samples, and these estimates qualitatively agreed with the neuronal content in images from the cell cultures. Our method is similar to a deconvolution approach described recently using bulk and single-cell sequencing of primary human and mouse pancreas (Baron et al. 2016).

The similarity of the fibroblast-like single cells to DRG raises the important question of whether these cells are immature sensory neurons. Single-cell sequencing at multiple time points during MYOD-mediated myogenic reprogramming has suggested that some individual cells traverse a desired course, while others terminate at incomplete or aberrant reprogramming outcomes (Cacchiarelli et al. 2017). Such an approach in IPSDSNs could reveal determinants of neuronal differentiation trajectories, and may yield useful insights for protocol changes to improve the purity of differentiated neurons, or to specify more precise neuronal subtypes. More generally, replacing bulk RNA-seq with single cell sequencing across many samples could enable *in silico* sorting of cells based on their transcriptome, and better characterisation of the sources of variation within a differentiated population of cells. Further, culturing cells from multiple donors in a pool, along with an scRNA-seq readout, could reduce differentiation-related batch effects while retaining the ability to identify donor-specific genetic effects on gene expression. These advantages suggest to us that a move towards scRNA-seq will be extremely useful in iPSC-derived cell models.

For iPSC models of common disease associated variants to be used effectively, it is critical to know which candidate disease associated variants exhibit a detectable cellular phenotype in an *in vitro* model. We used *in silico* resampling to estimate the sample sizes needed to detect the effects of noncoding regulatory variants in iPSC-derived cells using a recall by genotype design. Power above 80% is only achieved with surprisingly large (40+) samples,

even for alleles with a fold change of 1.5 to 2. Further, the power we report may be overestimated, due to ascertainment bias in defining a set of eQTLs as “true positives”, which fails to include true genetic effects that we did not discover in our samples. Even larger samples will be needed when multiple genes, for example in a single GWAS interval, are to be tested. These observations are consistent with a recent genome-editing experiment that required 136 differentiations in hepatocyte-like cells to discover an effect of rs12740374 on *SORT1* gene expression (Warren et al. 2017). Notably, the modest effect of this variant on expression in hepatocyte-like cells (1.3-fold increase) stands in contrast to the large effect of the variant (4- to 12-fold increase) observed previously in primary liver (Musunuru et al. 2010). Where it is possible to use a coding SNP to assess the allele-specific effect of a genome edit, as done for *SNCA* (Soldner et al. 2016), this may prove a more efficient approach to detecting causal effects of individual regulatory variants.

In summary, we have measured multiple molecular phenotypes in a large panel of iPSC-derived neurons. The catalog of QTLs we provide reveals a large set of common variants and target genes with detectable effects in IPSDSNs. These associations provide promising targets for functional studies to fine-map causal disease-associated alleles, such as by allelic replacement using CRISPR-Cas9, and our study describes the importance of considering differentiation-induced variability when planning these studies in iPSC-derived cells.

Data Availability

Code used for processing and analysing data is available at https://github.com/js29/ipsdsn. RNA-seq and ATAC-seq for open access samples are deposited in the European Nucleotide Archive under accession ERP020576. These data for managed access samples are deposited in the European Genome Archive under accession EGAD00001003145. Summary statistics and gene expression counts are available at https://www.ebi.ac.uk/biostudies/studies/S-BSST16. Sample genotypes and accession numbers are available at http://www.hipsci.org/data.

2.4 Methods

URLs

CIBERSORT, cibersort.stanford.edu.

ENCODE, www.encodeproject.org.

GTEx, www.gtexportal.org.

HIPSCI, www.hipsci.org.

IPS cell lines

A summary of iPSC lines used is available in Supplementary Table 2, and details of processes and assays for these iPSCs generated by the HIPSCI project are available at www.hipsci.org. Briefly, 107 human iPSCs from 103 healthy donors were obtained from HIPSCI[8]. Of these, 38 were initially grown in feeder-dependent medium and the remainder were grown in feeder-free E8 medium. All HIPSCI samples were collected from consented research volunteers recruited from the NIHR Cambridge BioResource, initially under existing ethics for iPSC derivation (REC Ref: 09/H0304/77, V2 04/01/2013), with later samples collected under a revised consent (REC Ref: 09/H0304/77, V3 15/03/2013).

Sensory neuron differentiation

All differentiations in this study were performed by a single individual, and a summary of the IPSDSN cell lines is in Supplementary Table 1. Two protocols were used, named P1 (13 differentiations) and P2 (110 differentiations). P1 protocol samples were included for QTL calling, and other analyses used P2 protocol samples exclusively. The P1 protocol (described in[7]) involved the addition of "2i" inhibitors (LDN193189 and SB-431542) for 5 days, followed by "5i" inhibitors (LDN193189, SB-431542, CHIR99021, DAPT, SU5402) for 6 days. When applying this protocol to a larger number of samples we observed excessive cell death prior to obtaining neural progenitors (days 9-12). We altered the protocol to make it more similar to (Chambers et al. 2012), by:

- using E8 rather than mTeSR1 media when maintaining iPSCs prior to differentiation;
- phasing in neurobasal media beginning at day 4, and gradually increasing this to 100% by day 11;
- beginning addition of inhibitors 5i at day 3 rather than day 5;
- stopping addition of small molecule inhibitors LDN193189 and SB-431542 beginning at day 7 rather than day 11, referred to as "3i" for the 3 inhibitors that continued to be added.

Functional assays (Ca^{2+} flux, response to Veratridine) confirmed that response of the sensory neurons produced by each protocol was equivalent; however, the P2 protocol performed more consistently across cell lines and culture parameters.

P2 protocol details

All reagents were from Life Technologies unless otherwise indicated. Clump-passaged iPSCs were single-cell seeded in E8 media on growth factor-reduced Matrigel (BD Biosciences) 48 hours prior to neural induction (day 0). KSR Media was prepared as 500ml DMEM-KO 130 ml Knockout Serum Replacement Xeno-Free, 1x NEAA, 1x Glutamax, 0.01 mM β-mercaptoethanol (Sigma). KSR media containing small molecule inhibitors

LDN193189 (100 nM) and SB-431542 (10 μM) was added to cells from day 0 to 3 to drive anterior neuroectoderm specification. From day 3, CHIR99021 (3 μM), DAPT (10 μM) and SU5402 (10 μM) were also added to further promote neural crest phenotypes. N2B27 media was progressively phased in every two days from D4. N2B27 Media was prepared as 500 ml Neurobasal medium, 5 ml N2 supplement, 10 ml B27 supplement without vitamin A, 0.01mM β-mercaptoethanol (Sigma) and 1x Glutamax. On day 7, inhibitors LDN193189 and SB-431542 were no longer used, while CHIR99021, DAPT, and SU5402 continued to be added. On day 11 cells were reseeded at 150,000 cells/cm2 in maturation media containing N2B27 media with human-b-NGF, BDNF, NT3 and GDNF (each at 25 ng/ml). Mitomycin-C treatment (1 μg/ml) was used once at day 14 for 2 hrs to reduce the non-neuronal population. Cells were differentiated in T25 flasks for RNA and nuclei isolation, and onto coverslips and 96-well plates for electrophysiology and Ca^{2+}-flux assays.

P1 protocol details

All reagents and concentrations used were identical to the P2 protocol; the difference was timing of addition. Clump-passaged iPSCs were single-cell seeded in mTeSR1 iPSC (StemCell Technologies, Vancouver) media on growth factor-reduced Matrigel (BD Biosciences) 48 hours prior to neural induction (day 0). KSR media containing LDN193189 and SB-431542 was added to cells from day 0 to 5. From day 5, CHIR99021, DAPT and SU5402 were also added. As for the P2 protocol, cells were reseeded on day 11, and treated with Mitomycin-C on day 14.

Single-cell RNA sequencing

Blood-derived iPSCs from a single individual, who was not a HIPSCI donor, were differentiated to IPSDSNs in 3 batches using the P2 protocol, and were matured for 8 weeks. Dissociated cells were loaded onto a Fluidigm C1 for automatic cell separation, reverse transcription and amplification. Libraries were prepared from C1 chambers containing single cells using the Illumina Nextera XT kit. These were quantified with the Qubit dsDNA HS assay (Thermo Fisher) and KAPA Library Quantification Kit (KAPA Biosystems) and size-checked with Agilent Bioanalyser DNA 1000. Libraries were 96-way multiplexed and sequenced by Illumina Nextseq500 (2x75bp). Reads were aligned to GRCh38 and Ensembl 80 transcript annotations using STAR v2.4.0d with default parameters. We excluded 9 cells expressing fewer than 20% of the ~56,000 quantified genes, and then used SC3[14] to cluster the remaining 177 cells based on expression counts. We examined alternative numbers of clusters from k=2 to 5. With two clusters, marker genes clearly identified one cluster (111 cells) as neuronal, whereas the other cluster (66 cells) had high expression of extracellular matrix genes reminiscent of fibroblasts. With 3 and 4 clusters, the sensory-neuronal cell

cluster remained unchanged, and the fibroblast-like cluster became further subdivided. This suggests that a majority of the cells in this sample were terminally differentiated into sensory neurons, whereas the remaining cells were more heterogeneous in their gene expression.

To display marker gene expression (Figure 2a), we used DESeq2's variance stabilizing transformation, and then R's "scale" function to mean-center and normalize expression values across cells, and plotted the result using the pheatmap R package. To compare gene expression between single-cell clusters and bulk RNA-seq samples (Figure 4), we computed the mean FPKM for each gene separately in single neurons and fibroblast-like cells. We subsetted to genes with nonzero expression in at least one GTEx tissue and in at least one of our tissues (iPSC, DRG, IPSDSN bulk, IPSDSN single cells), and computed the Spearman correlation between each pair of tissues.

Genotypes

We obtained imputed genotypes for all of the samples from the HIPSCI project. We used CrossMap (http://crossmap.sourceforge.net/) to convert variant coordinates from GRCh37 to GRCh38, and used bcftools (http://samtools.github.io/bcftools/) to retain only bi-allelic variants (SNPs and indels) with INFO > 0.8 and MAF > 0.05 in the 97 samples used for QTL calling.

RNA sequencing

The 131 RNA samples corresponded with 103 unique HIPSCI donors, as some samples were differentiation or RNA-extraction replicates. One sample failed in sequencing and was excluded. For QTL analyses, reads for each sample were aligned to GRCh38 and Ensembl 79 transcript annotations using STAR v2.4.0j with default parameters. Using VerifyBamID v1.1.2 (Jun et al. 2012) we identified 5 mislabeled RNA samples for which the matching genotypes could be determined, as well as two samples with no matching genotypes and which were thus excluded. For comparisons among tissues, reads were aligned to the 1000 Genomes GRCh37d5 reference with Gencode v19 transcript annotations using STAR 2.5.3a.

Gene expression quantification, quality control and exclusions

Gencode Basic transcript annotations, GRCh38 release 79, were downloaded from www.gencodegenes.org. Gene expression was counted for uniquely mapping reads using featureCounts (v1.5.0) (Liao, Smyth, and Shi 2014) with options (-s 2 -p -C -D 2000 -d 25). A median of 45 million reads were generated per sample, with median 32.8 million reads (72%) uniquely mapping and assigned to genes. After excluding short RNAs and

pseudogenes, we normalised expression counts for 35,033 genes using the R package cqn v5.0.2 (Hansen, Irizarry, and Wu 2012).

We determined pairwise correlation between samples using normalized counts for 14,215 expressed genes (CQN > 1) and the first five principal components of gene expression against each other. We excluded four outlier samples from subsequent analyses, leaving 126 samples from 97 donors. For QTL calling, replicate BAM files from same donor were merged together using samtools.

To assess gene expression replicability, we determined the spearman correlation coefficient of CQN-normalized expression between samples across all genes for (a) extraction replicates, (b) differentiation replicates, and (c) all possible pairs of samples from different donors, and plotted the histogram of correlation coefficients in Figure 8a.

DRG samples and sequencing

Human tissue acquisition and handling was performed at Pfizer Neuroscience and Pain Research Unit in accordance with regulatory guidelines and ethical board approval. Postmortem human DRG were obtained in dissected form from Anabios or as an encapsulated sheath together with sensory/afferent axons from National Disease Research Interchange and were subsequently dissected to isolate the cell-body rich ganglion. The tissue was homogenised in QIAzol Lysis Reagent according to weight and processed according to the manufacturer's instructions for the Qiagen RNeasy Plus lipid-rich kit. RNA was prepared with the Illumina TruSeq Stranded mRNA Library Prep Kit and sequenced (2x100 bp reads) on Illumina HiSeq 2500. Reads were aligned to GRCh37 using STAR and gene counts and FPKMs obtained using featureCounts and Ensembl v75 gene annotations.

ATAC library preparation and sequencing

Nuclei isolation

Media was removed from T25 flasks and washed twice with 10 mL of room temperature D-PBS -/-. The adherent neuronal cultures were lifted by treating with 3 mL of Accutase (Millipore – SCR005) at room temperature for four minutes. The Accutase was quenched by adding 6 mL of 2% foetal bovine serum in D-PBS. The cells were transferred to a 15 mL conical tube and centrifuged at 300 g for 5 minutes at 4°C. The cell pellet was resuspended in 1 mL of ice-cold sucrose buffer (10 mM tris-Cl pH 7.5, 3 mM $CaCl_2$, 2 mM $MgCl_2$ and 320 mM sucrose) and pipetted briefly to break up the large clumps before incubating on ice for 12 minutes. 50 µL of 10% Triton-X 100 was added to the sucrose-treated cells and mixed briefly before incubating on ice for a further 6 minutes. Nuclei were released by performing

30 strokes with a tight dounce homogeniser on ice. Approximately 1 x 10^5 nuclei were transferred to a 1.5 mL microfuge tube and centrifuged at 300 g for 5 minutes at 4 °C. All traces of the lysis buffer were removed from the nuclei pellet.

Tagmentation and sequencing

The tagmentation and PCR methods used were as described in (Kumasaka, Knights, and Gaffney 2016), based on the Nextera tagmentation master mix (Illumina FC-121-1030). To remove excess unincorporated primers, dNTPS and primer dimers we used Agencourt AMPure XP magnetic beads (Beckman Coulter A63880), followed by size selection using 1% agarose TAE gel electrophoresis, selecting library fragments from 120 bp to 1 kb. Gel slices were extracted with the MinElute Gel Extraction kit (Qiagen 28604), eluting in 20 µL of Buffer EB. A total of 31 ATAC-seq libraries each prepared with a unique Nextera i5 and i7 tag combination were pooled. Index tag ratios were assessed by a single MiSeq run and were balanced before being sequenced at two per lane with paired-end reads (2x75) on a HiSeq with V4 chemistry. However, rebalancing did not appear to work correctly, as the number of reads varied from a minimum of 17 million to a maximum of 987 million. However, 22 samples had over 100 million reads, and 30 samples had over 40 million reads. Across samples, a median of 56% of reads mapped to mitochondrial DNA.

Read alignment

We aligned reads to GRCh38 human reference genome using bwa mem v0.7.12. Reads mapping to the mitochondrial genome and alternative contigs were excluded. As for RNA-seq data, we used VerifyBamID v1.1.2 (Jun et al. 2012) to detect sample swaps. This revealed one mislabeled sample, which we then corrected. We used Picard v1.134 MarkDuplicates (https://broadinstitute.github.io/picard/) to mark duplicate fragments.

Peak calling

We used MACS2 v2.1.1 (Zhang et al. 2008) to call ATAC-seq peaks for individual samples with parameters '--nomodel --shift -25 --extsize 50 -q 0.01'. We defined a consensus set of peaks as regions in which peaks overlapped in at least 3 samples. At regions of overlap, the consensus peak was defined as the union of overlapping peaks. This resulted in 381,323 peaks, with 98% of peaks ranging from 82 - 1191 base pairs.

PCA plot clustering samples with GTEx tissues

We downloaded the GTEx v6 gene RPKM file (GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz) as well as sample metadata (GTEx_Data_V6_Annotations_SampleAttributesDS.txt) from the GTEx web portal

(http://www.gtexportal.org/home/datasets). We computed RPKMs for all genes for the 28 DRG samples, the 119 sensory neuron samples after QC exclusions, and 239 HIPSCI IPSC samples. We used genes that were quantified in all of these sample sets, and where at least 50 GTEx samples had RPKM > 0.1. We passed log2(RPKM + 1) for 8,553 GTEx samples to the bigpca R package to compute the first 5 PCs using the SVD method. We determined sample loadings for each PC using the PC weights and log2(RPKM + 1) values for GTEx samples and for our in-house samples, and plotted PC1 vs. PC2 values as Figure 1b.

Highly variable genes in IPSDSNs and GTEx

For each of the 44 GTEx tissues, as well as IPSDSNs, DRG, and HIPSCI iPSCs, we calculated the coefficient of variation (CV) of each gene's RPKM expression among samples of the same tissue (SMTSD in GTEx metadata). In each tissue, we subsetted the genes considered to those expressed at RPKM > 1. We plotted the distribution of CVs across all genes for each tissue as Figure 3a.

We used GeneTrail2 (https://genetrail2.bioinf.uni-sb.de) to do a gene set over-representation analysis for the top 1000 most variable genes in IPSDSNs by CV (Supplementary Table 4). Similarly, gene set over-representation analysis in E8-IPSDSN subsets was done using Genetrail2 and the top 1000 most variable genes with RPKM > 1 (Supplementary Table 11).

Variance components analysis

For Figure 12a, we selected the 119 QC-passed samples, and used DESeq2 to get FPKM values for each gene after size factor normalization. We included all genes with mean FPKM > 1, and input log2-transformed counts per sample into the variancePartition Bioconductor R package. For Figure 12b, we used 18 samples for which we had 3 differentiation replicates from each of 6 donor cell lines; all 6 iPSC lines were from females and had been cultured in E8 medium. We therefore included only donor and differentiation in the design formula.

Estimation of neuronal purity

We used Cibersort (Newman et al. 2015) to estimate the fraction of RNA from neuronal cells in our bulk RNA-seq samples. We used the 14,786 genes with mean CQN expression > 0 in bulk RNA samples, and retrieved raw counts for these genes in our scRNA-seq data. We labeled the single cell counts as "neuron" or "fibroblast-like" based on the SC3 clustering, and used these as reference samples for CIBERSORT to generate custom signature genes. We used raw expression counts for the same genes for our 126 bulk RNA-seq samples as the mixture file for CIBERSORT to use in estimating the relative fractions of neuron and fibroblast-like cell RNA.

Electrophysiological recordings

Six coverslips per line were placed singularly into a 12-well plate and washed 1x with 1 ml DPBS (+/+). The coverslips were then coated with 1 ml of 0.33 mg/ml growth factor reduced matrigel for > 3 hr at room temperature. D14 cells were prepared at 1.6e6/ml in 15 ml media, then diluted in NB media to create a 0.3e6/ml suspension. The coverslips were transferred into a 12-well plate and 1 ml cell suspension was added. Plates were incubated at 37°C (5% CO2) for 24hr, after which the coverslips were transferred to a 12-well plate with 2 ml media. Cells were treated with Mitomycin C (0.001 mg/ml for 2hr hours at 37°C) post-plating on days 4 and 10. Media was changed twice weekly.

Patch-clamp experiments were performed in whole-cell configuration using a patch-clamp amplifier 200B for voltage clamp and Multiclamp 700A or 700B for current clamp controlled by Pclamp 10 software (Molecular Devices). Experiments were performed at 35°C or 40°C controlled by an in-line solution heating system (CL-100, Warner Instruments). Temperature was calibrated at the outlet of the in-line heater daily before the experiments. Patch pipettes had resistances between 1.5 and 2 MΩ. Basic extracellular solution contained (mM) 135 NaCl, 4.7 KCl, 1 $CaCl_2$, 1 $MgCl_2$, 10 HEPES and 10 glucose; pH was adjusted to 7.4 with NaOH. The intracellular (pipette) solution for voltage clamp contained (mM) 100 CsF, 45 CsCl, 10 NaCl, 1 $MgCl_2$, 10 HEPES, and 5 EGTA; pH was adjusted to 7.3 with CsOH. For current clamp the intracellular (pipette) solution contained (mM) 130 KCl, 1 $MgCl_2$, 5 MgATP, 10 HEPES, and 5 EGTA; pH was adjusted to 7.3 with KOH. The osmolarity of solutions was maintained at 320 mOsm/L for extracellular solution and 300 mOsm/L for intracellular solutions. All chemicals were purchased from Sigma. Currents were sampled at 20 kHz and filtered at 5 kHz. Between 80% and 90% of the series resistance was compensated to reduce voltage errors. Rheobase was measured in current clamp mode by injecting increasing 30 milliseconds current steps until a single action potential was evoked. Intersweep intervals were 2 seconds. Current clamp data was analyzed using Spike2 software (Cambridge Electronic Device, UK) and Origin 9.1 software (Originlab).

Correlation of iPSC and IPSDSN gene expression with cell culture conditions

We selected the 106 IPSDSN samples differentiated with the P2 protocol, as well as the 87 iPSC samples these were derived from and for which we had RNA-seq data, and we used DESeq2's variance stabilising transformation on the raw gene expression counts. We computed the first 5 principal components of gene expression separately in iPSC and IPSDSNs, and used corrplot to compute pairwise correlations among these PCs and sample

metadata: gender, iPSC passage number, iPSC culture conditions (wasFeeder), iPSC PluriTest score, IPSDSN fibroblast content, and IPSDSN processing date.

We determined differentially expressed genes between feeder-iPSCs and E8-iPSCs using DESeq2 (Love, Huber, and Anders 2014), using expression counts for genes with median FPKM > 0.1 across iPSC samples (Supplementary Table 5). We removed associations driven by outliers, defined as a maximum Cook's distance >= 5. Similarly, we determined differentially expressed genes in IPSDSNs derived from either feeder-iPSCs or E8-iPSCs (Supplementary Table 8), again for genes with median FPKM > 0.1. We used GeneTrail2 (https://genetrail2.bioinf.uni-sb.de) to do a gene set over-representation analysis for the 717 genes with expression at least 2-fold higher in feeder-iPSCs than E8-iPSCs, and similarly for the 631 genes at least 2-fold higher in E8-iPSCs (Supplementary Tables 6, 7). We did an equivalent gene set over-representation analysis for the 1,159 genes with expression at least 2-fold higher in IPSDSNs differentiated from feeder-iPSCs, and also for the 958 genes at least 2-fold higher in IPSDSNs from E8-iPSCs (Supplementary Tables 9, 10).

To determine genes upregulated on differentiation from iPSCs to IPSDSNs, we first selected the 19,658 genes with expression FPKM > 1 in at least two samples (iPSC or IPSDSN). We used DESeq2 as before, removing genes with maximum Cook's distance > 5, identifying 4,246 differentially expressed genes at FDR 1%.

QTL calling

Expression QTLs

To call cis-eQTLs we used RASQUAL (Kumasaka, Knights, and Gaffney 2016), which leverages allele-specific reads in heterozygous individuals to improve power for QTL discovery, while accounting for reference mapping bias and a number of other potential artifacts. With RASQUAL a feature is defined by a set of start and end coordinates; for calling a gene eQTL these are the start and end coordinates for exons, whereas for an ATAC-seq peak these are the peak coordinates. RASQUAL requires as input the allele-specific read counts at each SNP within a feature. We used the Genome Analysis Toolkit (GATK) program ASEReadCounter (Castel et al. 2015) with options '-U ALLOW_N_CIGAR_READS -dt NONE --minMappingQuality 10 -rf MateSameStrand' to count allele-specific reads at SNPs (and not indels). We then annotated the AS read counts in the INFO field of the VCF used as input for RASQUAL.

We used RASQUAL's makeCovariates.R script to determine principal components (PCs) to use as covariates, which determined 12 PCs as appropriate from the expression count data.

We ran RASQUAL separately for each of 35,033 genes (19,796 protein-coding genes and 15,237 noncoding RNAs), passing in VCF lines for all SNPs and indels (MAF > 0.05, INFO > 0.8) within 500 kb of the gene transcription start site. We used the --no-posterior-update option in RASQUAL, as we found that not doing so led to some genes having miniscule p values in permuted data. To correct for multiple testing we used permutations; however, because RASQUAL is computationally intensive, it would not be possible to run a thousand or more permutations for every gene. Therefore we used an approach to balance power and computational time. To correct for the number of SNPs tested per gene, we used EigenMT (Davis et al. 2016) to estimate the number of independent tests per gene, and then performed Bonferroni correction on a gene-by-gene basis. To estimate the false discovery rate (FDR) across genes, we used the --random-permutation option of RASQUAL and re-ran it once for every gene, saving the minimum p value (after eigenMT correction) of the SNPs tested for each gene. This gave a distribution of minimum p values across genes for the permuted data. To determine the FDR for eQTL discovery at a given gene, we used R to compute (#permuted data min p values < p) / (#real data min p values < p), where p is the minimum p value among SNPs for the gene in question. With this procedure we obtained 3,778 genes with a cis-eQTL at FDR 10% (2,628 at FDR 5%).

For QTL calling with FastQTL, we first computed principal components from the CQN-transformed gene expression matrix (cqn v5.0.2 (Hansen, Irizarry, and Wu 2012)). We ran FastQTL with permutations 31 separate times, in each run including the first N principal components (N=0...30) as covariates. For each run we used a cis-window of 500 kb, and included SNPs and indels with MAF > 0.05, INFO > 0.8, as we did for RASQUAL. We plotted the number of eGenes found in each of these runs, which plateaued and remained relatively stable at ~1,400 eGenes (FDR 10%) when anywhere from 16 to 30 PCs were used. We arbitrarily chose to use the FastQTL run with 20 PCs in downstream analyses.

ATAC QTLs

As we did for gene expression, we used featureCounts v1.5.0 to count fragments overlapping consensus ATAC-seq peaks and ASEReadCounter to count allele-specific reads at SNPs (and not indels) within peaks. We ran RASQUAL separately for each of 381,323 peaks, passing in VCF lines for SNPs and indels (MAF > 0.05, INFO > 0.8) within 1 kb of the center of the peak. Since >99.9% of peaks were less than 2 kb in size, this meant that we tested effectively all SNPs within peaks. As we did when calling eQTLs, we ran RASQUAL with the --random-permutation option for every gene, and determined FDR as described above. Note that in this case we used Bonferroni correction based on the number

of SNPs tested, without using EigenMT, due to the small size of the windows tested. With this procedure we obtained 6,318 ATAC peaks with a cis-QTL at FDR 10%.

Splice QTLs

We downloaded LeafCutter from Github (https://github.com/davidaknowles/leafcutter) on April 17, 2016. We used the LeafCutter bam2junc.sh script to determine junction counts for each sample, followed by leafcutter_cluster.py. This resulted in 254,057 junctions in 59,736 clusters. To focus on splicing events likely to be significant, we applied a number of filters, including: (a) removing junctions accounting for less than 2% of the cluster reads, (b) removing introns used (i.e. having at least 1 supporting read) in fewer than 5 samples, (c) retaining only clusters where at least 10 samples had 20 or more reads in the cluster. This yielded a filtered set of 95,786 junctions in 30,591 clusters. We first determined the read proportions for all junctions within alternatively excised clusters. We then Z-score standardised each junction read proportion across samples, and then quantile-normalised across introns. We used this as our phenotype matrix for input to FastQTL to test for associations between intron usage and variants within 15 kb of the center of each intron. We chose a cis-window size of 30 kb (2 x 15 kb) because >91% of introns are < 30 kb in size, and so this tests variants near exon/intron boundaries for the great majority of introns, while maximising power.

We ran FastQTL in nominal pass mode 31 times specifying the first 0 to 30 principal components as covariates, and examined the number of intron QTLs with minimum SNP p value < 10^{-5}. This showed that the number of QTLs plateaued when 5 PCs were used, and so we used 5 PCs in subsequent runs. We next ran FastQTL with 10,000 permutations to determine empirical p values for each alternatively excised intron. To correct for the number of introns tested per cluster, we used Bonferroni correction on the most significant intron p value per cluster. We then used the Benjamini-Hochberg method to estimate FDR across tested clusters. This yielded 2,079 significant SNP associations for intron usage (sQTLs) at FDR 10%.

For significant sQTLs we used bedtools closest with GRCh38 release 84 to annotate the gene(s) nearest the lead SNP for the association. To ensure we had relevant genes, we filtered the annotation to include only genes where one of the exon boundaries matched the intron boundary for the sQTL.

Identifying tissue-specific eQTLs

We determined the set of tissue-specific eQTLs using the same procedure and code as in the HIPSCI project (Kilpinen, Helena, et al. 2016). Briefly, we considered the full cis eQTL output of sensory neuron eQTLs and 44 tissues analyzed by the GTEx Project (GTEx Consortium 2015). To enable comparison, lead SNP positions for sensory neuron eQTLs were first lifted back from GRCh38 to GRCh37 using Crossmap (Zhao et al. 2014). For each discovery tissue (including sensory neurons), we tested for the replication of all lead eQTL - target eGene pairs reported at FDR 5%. If the lead eQTL variant was not reported in the comparison tissue, then the best high-LD proxy of the lead variant ($r^2 > 0.8$ in the UK10k European reference panel) was used as the query variant. Replication was defined as the query variant having a nominal eQTL $p < 2.2x10^{-4}$ (corresponding to $p = 0.01 / 45$, where 45 refers to the total number of tissues tested) for the same eGene. We then extracted eGenes for which the lead eQTL did not show evidence of replication in any other tissue ($p > 2.2x10^{-4}$) or could not be tested (i.e. was not measured or reported as expressed in any other tissue).

This analysis gave 954 eGenes where the eQTL is specific to sensory neurons (Supplementary Table 15). We note that some of these “tissue-specific” eGenes could be due to the difference in QTL-calling methods used, notably that we used RASQUAL, a method incorporating both allele-specific and population-level expression variation. Therefore, some of the tissue-specific eGenes we report may actually be present more broadly in GTEx tissues but missed by the linear QTL model used in GTEx. Among the 1,403 eGenes called by FastQTL, 208 were tissue-specific to IPSDSNs.

Motif enrichment analyses

We used the R Bioconductor package LOLA (Sheffield and Christoph 2015) to identify enrichments in transcription factor binding sites (TFBS) and motifs. We defined three sets of loci to consider for enrichment: 1) tissue-specific eQTL SNPs with a window of 50 bp (+/- 25) around the SNP position, 2) all eQTL SNPs (50 bp window), and 3) all ATAC-seq peaks. For the QTLs we used all GTEx eQTL lead SNPs as the “universe” set against which we tested TFBS for enrichment. For this loaded GTEx eQTLs in R and used the liftOver function from rtracklayer to convert their coordinates to GRCh38. We tested for enrichment against the LOLA core database considering only ENCODE TFBS enrichments (Supplementary Tables 16 and 17). We also tested ATAC-seq peaks for enrichment relative to DNaseHS for many tissues from (Sheffield et al. 2013), which are available in the LOLA catalog. Motif enrichments in ATAC-seq peaks are reported in Supplementary Table 18.

Power simulations

Gene expression values were normalized to counts per million. We selected the 544 eGenes discovered by RASQUAL at FDR 1% which met the following criteria:

- at least 10 P2-protocol samples homozygous for each allele of the lead eQTL variant,
- mean expression among homozygous carriers was consistent with RASQUAL's reported direction of effect, and
- CV < 2 (this filter removed only 8 eGenes)

For each gene we resampled the normalized expression values, with replacement, from IPSDSN samples to achieve a specified number of samples ($N \in \{4,6,10,20,40\}$) with each homozygous genotype. From 100 such resamplings, we defined the power to discover a given variant's effect as the fraction of cases with $p < 0.05$ from a Wilcoxon rank sum test comparing expression in each genotype category. We determined the allelic fold change between genotypes using RASQUAL's effect size (pi), as:

fold change = max(pi / (1-pi), (1-pi) / pi)

We used ggplot2 with geom_smooth to display the 95% confidence interval around the fitted mean TPR at each parameter combination. As can be seen on the plots, the deviation about this mean for individual genes is larger than the standard error of the mean.

QTL overlap with GWAS catalog

The GWAS catalog was downloaded from https://www.ebi.ac.uk/gwas/ on 2016-5-08. To determine overlap between variants in the GWAS catalog and our lead QTLs, we first extracted all lead variants (both QTLs and GWAS catalog variants) from the full VCF file. We used vcftools v0.1.14 (Danecek et al. 2011) to compute the correlation R^2 between all lead variants within 500 kb of each other among our samples. We determined overlap separately for eQTLs, sQTLs, and ATAC QTLs, and retained only overlaps with $R^2 > 0.8$ between lead variants. Note that a given GWAS variant may be in LD with an eQTL for more than one gene, and vice versa, an eQTL for a single gene may be in LD with more than one GWAS catalog entry.

To determine whether our QTL overlaps were enriched in any specific GWAS catalog traits relative to other traits, we computed overlap with all GWAS catalog SNPs ($p < 5x10^{-8}$) but sought to eliminate redundant overlaps. For traits that were reported with differing names (e.g. "Alzheimer's disease (cognitive decline)" and "Alzheimer's disease in APOE e4-carriers"), we grouped these into a single trait name (e.g. "Alzheimer's disease"). We then sorted overlaps by decreasing LD R^2, and kept the single overlapping QTL with the highest R^2 for each GWAS catalog entry. Similarly, we removed duplicates with the same reported

GWAS catalog SNP and trait, such as when successive GWAS of the same trait report the same SNP association. We counted the number of such unique GWAS-QTL overlaps separately for eQTLs, sQTLs, and caQTLs, and we report these in Table 1. To avoid bias due to correlation between GWAS power and LD patterns, we restricted our analysis to the 41 traits with at least 40 GWAS catalog associations. We then considered the binomial probability of the observed overlap with each trait, with the expected overlap frequency being the proportion of QTL overlaps among all trait associations (6.2%). After correcting for multiple testing, no traits showed significantly greater overlap with our QTL catalog than other traits.

To test for overall enrichment of QTLs overlapping with GWAS catalog SNPs, we used vcftools to identify 1000 Genomes SNPs in LD $R^2 > 0.8$ with a GWAS catalog SNP, and removed duplicate SNPs. We used our IPSDSN eQTL lead SNPs as input to SNPsnap (https://data.broadinstitute.org/mpg/snpsnap/), and computed 1000 random sets of SNPs matched for LD partners, MAF, gene density, and distance to nearest gene. IPSDSN eQTL lead SNPs had more overlaps (92) with GWAS catalog + $R^2>0.8$ SNPs than did any of the matched sets (median: 58, range 37-87).

Acknowledgments

The iPSC lines were generated under the Human Induced Pluripotent Stem Cell Initiative (HIPSCI) funded by a grant from the Wellcome Trust and Medical Research Council, supported by the Wellcome Trust (WT098051) and the NIHR/Wellcome Trust Clinical Research Facility. HIPSCI funding was used for sensory neuron RNA-sequencing. We acknowledge Life Science Technologies Corporation as the provider of Cytotune. Pfizer Neuroscience (Pfizer Ltd.) funded neuronal differentiation, functional assays, single-cell RNA-sequencing, and collection and sequencing of dorsal root ganglion samples. We thank Natsuhiko Kumasaka for help with RASQUAL, and Florian Merkle for comments on the manuscript.

Conflicts of Interest

SF, RF, CB, AW, MB, EI, LC, SL, AJL, PJW and AGu were all employees of Pfizer at the time the experiments were performed.

3 PRF scores: predicting cell type-specific regulatory function of genetic variants

Collaboration note

Natsuhiko Kumasaka provided eQTL summary statistics for Geuvadis samples used in this chapter. All other work described here is my own, with advisory input from Daniel Gaffney.

3.1 Introduction

Prioritizing genetic variants likely to be functional is a general problem that is relevant to both Mendelian disease and complex traits. Because experimentally demonstrating the molecular effects of individual variants is laborious, computational approaches to prioritize variants to investigate can be extremely useful. Early approaches to variant effect prediction, such as PolyPhen (Adzhubei et al. 2010) and SIFT (Kumar, Henikoff, and Ng 2009), focused on the effects of nonsynonymous protein-coding variants, but these comprise fewer than 1% of all common variants. Effective methods to distinguish functional non-coding variants are essential: at least 85% of complex trait associations appear to be non-coding, and it is suspected that non-coding changes may be involved in Mendelian disease cases for which exome sequencing has failed to identify coding variants.

The enormous growth of functional genomic data has led to a corresponding growth in methods using these data to predict non-coding variant functionality. The simplest approaches, such as HaploReg (Ward and Kellis 2012) and RegulomeDB (Boyle et al. 2012), annotate variants based on their overlap with multiple datasets, but leave interpretation up to the user. This interpretation is particularly difficult because a large fraction of genetic variants overlap at least one functional genomic feature, and these features are also correlated amongst each other. More recently, methods have been developed that use statistical learning to integrate these diverse data inputs into a single score for each variant's likelihood of having a functional effect. These can be broadly divided into two categories: i) those which attempt to distinguish benign from deleterious variants (such as CADD (Kircher et al. 2014), GWAVA (Ritchie et al. 2014), FATHMM-MKL (Shihab et al. 2015), and LINSIGHT (Y.-F. Huang, Gulko, and Siepel 2017)), and ii) those which learn DNA sequences that affect cell type-specific molecular phenotypes (such as deltaSVM (D. Lee et al. 2015), DeepSEA (J. Zhou and Troyanskaya 2015), and Basset (Kelley, Snoek, and Rinn 2016)). A key difference is that methods in the first category produce cell type-

agnostic scores, whereas those in the second category are linked to the cell type-specific annotations used.

The interpretation of these scores, and their utility for different purposes, depend upon both the supervised training data and the functional genomic annotations used as input. For example, CADD trained a support vector machine to distinguish common variants and human derived alleles from simulated variants, which are not present in the genome and so are presumed to have been depleted by selection. CADD therefore measures deleteriousness relevant to fitness, and is likely to be biased towards treating common variants as benign, even though common variants can have functional effects. GWAVA and FATHMM-MKL were trained to distinguish pathogenic variants in the human gene mutation database (HGMD) from common variants. However, the known examples of pathogenic non-coding mutations likely have a massive ascertainment bias, as 75% are within 2 kb of an annotated TSS (Ritchie et al. 2014). The performance of these scores in predicting distal functional regulatory sites, as seems to be more common for GWAS associations, is unknown. In addition, all of these methods produce scores that are opaque, and it is difficult to know why one variant scored more highly than another, which limits mechanistic interpretation of the variants' functions.

Methods that predict the effect of variants on molecular phenotypes are in general tied to a particular cell type-specific dataset. Basset and deltaSVM predict the effect of a variant on DNase hypersensitivity from a given assay, but do not incorporate additional informative annotations, such as distance to TSS, TFBS and histone modifications. DeepSEA provides scores across many cell type-specific assays, including TFBS from ChIP-seq, but does not integrate these scores together, making their interpretation difficult.

In this chapter, we describe PRF scores, which integrate a large set of functional genomic annotations to produce scores that reflect the cell type-specific **p**robability of **r**egulatory **f**unction for common, non-coding variants. PRF scores are transparent, as a variant's score can be broken down into the contributions from individual annotations. Our primary annotation sources are the uniform epigenomic annotations in 119 cell types from the Roadmap epigenomics project, along with FANTOM TSS information, conservation, and gene annotations. Our model is trained using eQTL data, which makes our predictions particularly relevant to common regulatory variants, such as those hypothesized to underlie many GWAS associations. Although eQTL maps are being produced in many cell types by the GTEx consortium (GTEx Consortium 2013), these have limited sample size for many tissues, and cannot hope to cover the full range of human cell types and cellular

contexts/conditions. EQTL studies also provide no information at genomic positions where no variants are observed in the population studied, and are not well powered for low-frequency variants or those with small effect sizes. There thus remains a need for genome-wide predictions of variant regulatory effects across a broad range of cell types and conditions.

In developing PRF scores, we explored alternative ways of using specific epigenomic annotations. We found that using the quantitative level of histone modification and DNase hypersensitivity signals can improve prediction performance. We also found that imputed signal tracks from Roadmap Epigenomics are more predictive of eQTLs than the measured data. We show that, compared with CADD and GWAVA, PRF scores are dramatically better at prioritizing likely causal eQTL variants when distance to the regulated gene is included, but only slightly superior when the relevant gene is not known.

Unlike other variant scoring methods, PRF scores can be converted into relative probabilities that each variant regulates gene expression. When applied to fine-map eQTLs from GTEx, PRF scores reduced the size of the set of credibly causal variants for 67% of loci.

3.2 Model development

3.2.1 Overview

The PRF score model uses eQTLs from the Geuvadis RNA-seq study of lymphoblastoid cell lines (LCLs) (Lappalainen et al. 2013) to learn enrichments for multiple annotations considered together. We used the negative binomial model implemented in RASQUAL (Kumasaka, Knights, and Gaffney 2016) to associate gene expression with single nucleotide polymorphisms (SNPs) in a 2 Mb window centered on each gene's transcription start site (TSS) for 343 European donors in Geuvadis. We selected the 6,340 protein-coding genes with eQTL $p < 10^{-6}$ for the lead variant, and passed association statistics for all tested SNPs as input to fgwas (J. Pickrell 2013). Fgwas implements a Bayesian hierarchical model in which the prior probability for a SNP to be causal is a function of the overall enrichment of each annotation it appears in, and is efficient enough to learn enrichments for hundreds of annotations across thousands of eQTLs. A summary of the fgwas model is provided in Appendix A.

Building a predictive model relies upon having informative data as input. We sought to identify genomic annotations that are broadly available and predictive of the cell type-specific effects of genetic variants. The ENCODE Consortium has performed over 9,000 assays on human tissues and cell lines, including measuring histone modifications, DNase-seq, and

transcription factor ChIP-seq (ENCODE Project Consortium 2012). However, these experiments are distributed unevenly across tissues, and there is no core set of assays that is ubiquitous across a large set of tissues. This would make it difficult to develop a model in one cell type that could be easily translated to other cell types. We therefore focused on data from the NIH Roadmap Epigenomics Mapping Consortium, which performed multiple epigenomic assays across 111 body tissues and 16 cell lines (Roadmap Epigenomics Consortium et al. 2015). Five core assays were measured across all samples, namely, ChIP-seq for the histone modifications H3K4me1, H3K4me3, H3K9me3, H3K27me3, and H3K36me3; a large fraction of samples also had assays for H3K27ac, H3K9ac, DNA methylation, and DNase hypersensitivity. Importantly, a sophisticated imputation algorithm was used to fill in missing data for samples lacking specific assays, by leveraging correlations across assays and samples (Ernst and Kellis 2015).

We began by investigating hypotheses about how the predictive value of annotations could be optimized. Since distance to the TSS of a gene is a highly informative feature for gene regulation, we empirically determined an optimal set of distance annotations. We also hypothesized that the quantitative value of annotations would be more informative than binary assignment of variants as in/out of annotation peaks. We extended fgwas to enable this, and compared quantitative versus binary versions of the same annotations. Next, we compared the predictive value of imputed vs. measured annotation data from Roadmap Epigenomics. Throughout these investigations, we used cross-validation likelihood to assess the different models. In cross-validation, a fraction of the data (the training set) is used to estimate model parameters, and the remaining fraction is used to obtain the likelihood of the model given those parameters. This estimated likelihood is thus not influenced by overfitting on the training set. We used ten-fold cross-validation, so that in each of ten iterations a different 10% of the gene eQTLs were used as validation and the remaining 90% were used to train the model.

3.2.2 Optimising gene distance annotations

Both GWAS associations and eQTLs are highly enriched near the TSSes of genes. For eQTLs, SNP distance to TSS is more predictive of association than any other individual annotation, including DNase I hypersensitivity. Despite this, many cases are known of variants regulating genes from considerable distances (Spitz 2016). It is therefore important to effectively model distance to gene to predict regulatory variants.

3.2.2.1 FANTOM TSSes are more predictive than Ensembl TSSes

Ensembl provides annotation of gene transcripts, including the locations of exons, and by implication the location of TSSes. Many genes in Ensembl have multiple transcript isoforms, making it unclear how to assign a specific TSS distance to each SNP. However, most genes express a single dominant transcript across tissues (Gonzàlez-Porta et al. 2013), suggesting that some Ensembl TSSes are less relevant than others. An alternative annotation of TSSes comes from the FANTOM consortium, which used cap analysis of gene expression (CAGE) to generate quantitative maps of TSS usage for many tissues (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014). The FANTOM annotation thus distinguishes highly used TSSes from those which are weakly used or unused.

We used fgwas to compute the enrichment of causal eQTL SNPs in different distance bins for three different TSS distance annotations:

1. distance to the nearest Ensembl TSS
2. distance to the mean position of all Ensembl TSSes for a given gene
3. distance to the nearest of the top 3 FANTOM TSSes in LCLs

Using the minimum distance allows SNPs near a strongly used TSS to receive maximal enrichment, but requires that SNPs near weakly used TSSes also receive high enrichment, which could reduce prediction performance. Using distance to an average TSS position avoids labeling SNPs with a small TSS distance near different weakly used TSSes, but may fail to correctly label SNPs in the nearest bins for the highly used TSS.

For all TSS-proximal distance bins, enrichment was highest when FANTOM TSSes were used (Figure 1), and this was reflected in a much higher cross-validation likelihood. This indicates that FANTOM TSSes are more informative in localising causal eQTL SNPs than are Ensembl TSSes, and so we used this method of TSS annotation in all subsequent models.

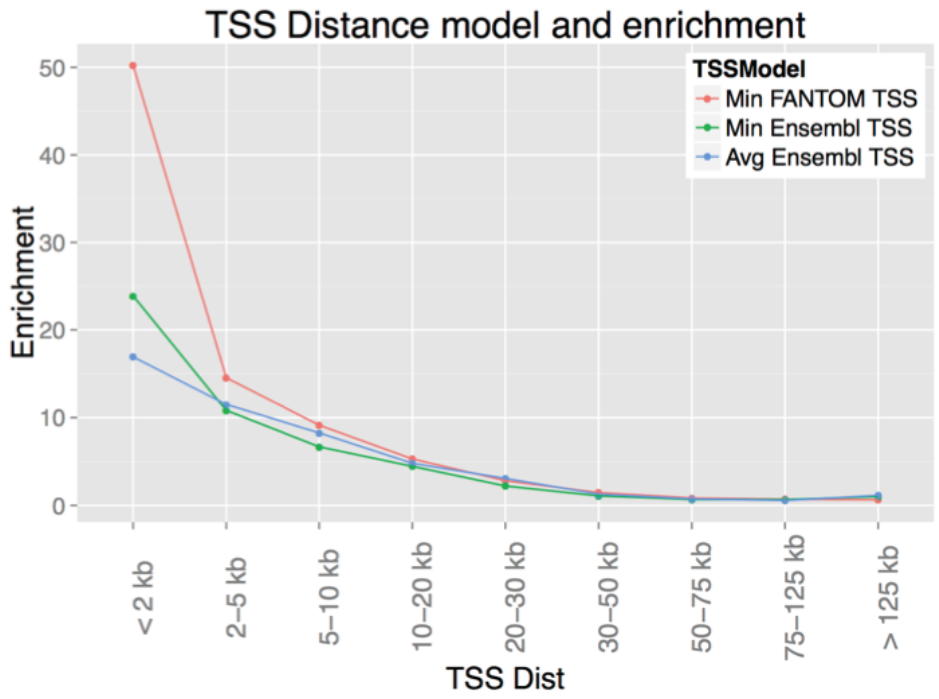
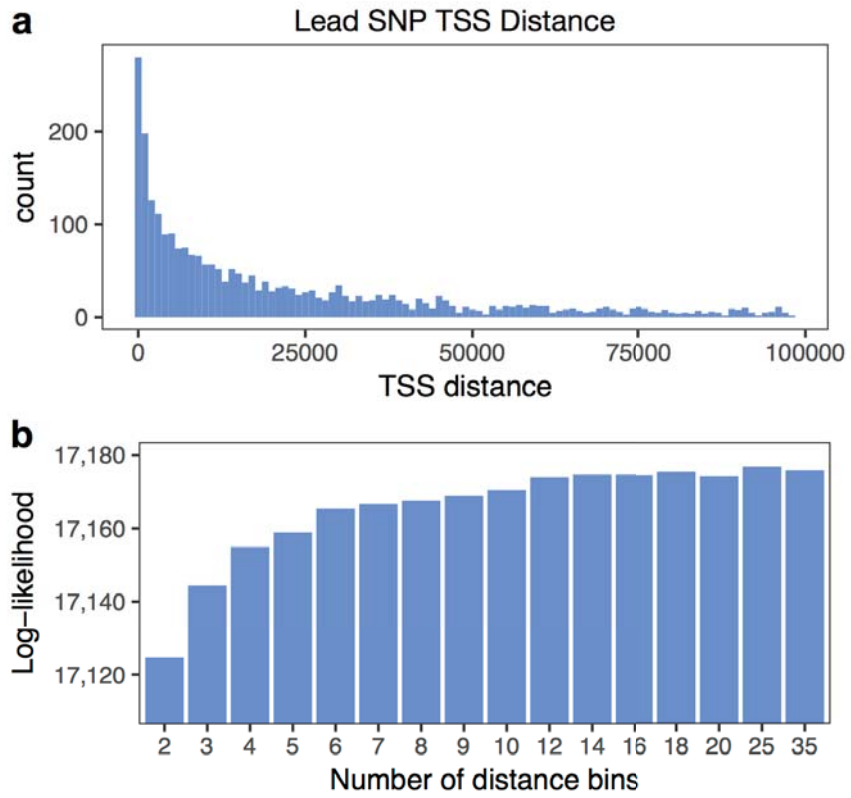


Figure 1: Enrichment of causal SNPs in fixed distance bins for the three TSS distance definitions described in the main text.

3.2.2.2 An optimal spacing of distance bins

When modeling distance to TSS, it is common to define bins at different distances so that each SNP can be assigned to a single bin. A drawback of such a definition is that neighbouring SNPs may fall in different bins and thus receive different enrichments, whereas SNPs many kilobases away but in the same bin receive the same enrichment. Binning is one of many possible smoothing functions, and the fit is less smooth than alternatives such as natural splines. However, splines are difficult to integrate into the iterative approach to model optimisation used in fgwas, since they need to be computed across all (x, y) points simultaneously (here, TSS distance and enrichment). As a binary annotation for each SNP, distance bin enrichments are also rapid to compute, which is essential for optimising the large, multi-annotation models that we evaluate later.

Many previous models have used only coarse bins of TSS distance (e.g. 2 bins (Kindt et al. 2013; J. Pickrell 2013), 3 bins (Ryan et al. 2014), or 4 bins (Schork et al. 2013)). We sought to systematically identify an optimal spacing TSS distance bins. To do this we first determined the distribution of TSS distance for lead eQTL SNPs (Figure 2a), using for each SNP the distance to the nearest of the top 3 FANTOM TSSes for the respective eQTL gene.



c

Bin start	Bin end
0	285
286	1,410
1,411	3,690
3,691	6,420
6,421	10,338
10,339	15,106
15,107	21,855
21,856	33,040
33,041	49,943
49,944	85,174
85,175	164,468
>164,468	

Figure 2: (**a**) TSS distance of lead eQTL SNPs to the gene they regulate. (**b**) Cross-validation data log-likelihood when SNPs are divided into distance bins with different granularity. (**c**) Distance bins used in a 12-bin model which has nearly maximal cross-validation likelihood. Each SNP is assigned to one bin based on its TSS distance.

We created a set of distance annotations with differing numbers of bins, with bin boundaries chosen from the quantiles of the lead SNP TSS distance distribution to contain an approximately equal number of SNPs. When these annotations were used with our eQTL training dataset, the cross-validation likelihood peaked with 25 distance bins (Figure 2b), indicating that distance models more fine-grained than this might be overfit. A model with only 12 distance bins was nearly equivalent and is much faster to fit with fgwas; thus, we chose to use 12 bins going forward. The bin definitions are shown in Figure 2c.

3.2.3 Quantitative annotations improve prediction performance

3.2.3.1 Quantitative annotation model

A standard workflow for using data from a ChIP-seq or DNase-seq experiment begins by calling peaks - that is, genomic regions with read counts that rise above the background observed genome-wide. The boundaries of called peaks can depend on the particular peak calling software used and on the parameters provided. Subsequently, a significance cutoff is used to retain only high-quality peak calls, typically at a specified false discovery rate. A

number of previous works have evaluated the enrichment of genomic annotations for causal eQTL or GWAS variants (Kindt et al. 2013; Gagliano et al. 2014; Schork et al. 2013; Gaffney et al. 2012), while others have incorporated multiple annotations for fine-mapping GWAS loci (Lu et al. 2016; Ryan et al. 2014; J. Pickrell 2013; Kichaev et al. 2014). All of these methods have relied assigning SNPs a binary 1 or 0 for an annotation depending on whether or not they are located within a called peak. Yet, peak calling parameters are often arbitrary, and this includes the threshold below which peaks are considered low quality and are discarded. It is unknown to what extent the quantitative information in the ChIP-seq or DNase-seq signal, such as the height of the peak or the signal value outside of peaks, is useful for identifying causal variants.

To use the quantitative signal value of annotations, we implemented an extension to fgwas (called qfgwas, available at https://github.com/js29/qfgwas) that models enrichment as a logistic function of the annotation's quantitative value at a SNP. The logistic function has two desirable features in this context: first, outliers in the distribution of annotation values will not substantially skew the model fit; second, the function can be most sensitive to input values over a specific range. This second property could be useful, for example, for chromatin accessibility data, where above a certain value the DNA is "open" and larger values contribute no more information. Figure 3 depicts how the two parameters of the logistic function relate the input value to an output.

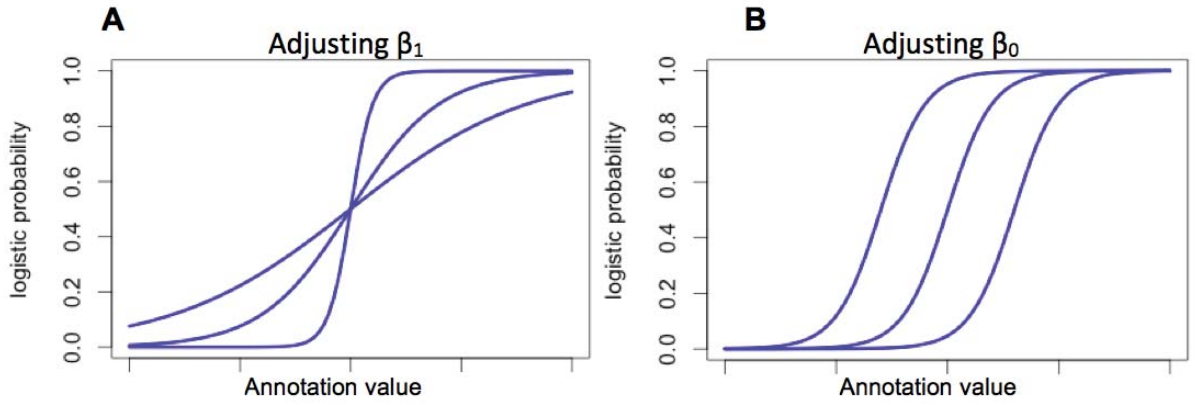


Figure 3: A standard logistic function enables controlling the slope parameter β_1 (**a**) which determines how quickly an annotation becomes informative, as well as the translation parameter β_0 (**b**) which determines at what absolute value the annotation begins to be informative.

In the hierarchical model implemented in fgwas (see Appendix A), the prior probability of a given SNP to be associated, π_{ik}, is allowed to depend on individual annotation enrichments, λ_l, according the following equations:

$$\pi_{ik} = \frac{e^{x_i}}{\sum_{j \in S_k} e^{x_j}} \quad \text{(Equation 2)}$$

$$x_i = \sum_{l=1}^{L_2} \lambda_l I_{il} \quad \text{(Equation 3)}$$

where *i* and *k* denote the i^{th} SNP in the k^{th} locus, S_k is the set of SNPs in locus k, L_2 is the number of annotations in the model, λ_l is the effect of SNP annotation l, and I_{il} is 1 if the SNP falls in annotation l or 0 otherwise. The annotation contribution to the prior probability for a given SNP is thus either λ or zero, depending on whether the SNP falls in the annotation or not. We can interpret λ as the log odds ratio for a causal SNP to appear in the annotation versus outside the annotation. The model is optimized by maximizing the likelihood of the data across all loci, with SNP annotation enrichments shared across loci. The combined enrichment for a given SNP across annotations, x_i, is the quantity that we refer to as a "PRF score", since it reflects the log of the probability for this SNP to be causally associated with gene expression, relative to other SNPs considered.

To exploit quantitative annotations we add to equation 3, replacing the indicator I_{il} with the logistic function that depends on the annotation value, z:

$$I_{il} = \frac{1}{1 + e^{-\beta_1(z-\beta_0)}} \quad \text{(Equation 4)}$$

Each quantitative annotation thus contributes three parameters to the model, λ, β_0 and β_1. Since I_{il} takes on values from 0 to 1 depending on the annotation's quantitative value, a SNP's enrichment relating to a particular annotation, $\lambda_l I_{il}$, varies between zero and λ_l. The

enrichment parameter, λ_l, then has largely the same interpretation as previously - it reflects the enrichment of causal SNPs in sites with the highest quantitative value, relative to the lowest value. β_0 controls the value at which the annotation has half-maximal enrichment, while β_1 influences the slope of the transition from uninformative to informative based on the annotation's quantitative value.

3.2.3.2 Model comparison

We selected three annotations from Roadmap Epigenomics LCLs to use in assessing the usefulness of quantitative annotation values: DNase hypersensitivity, histone H3K27ac ChIP-seq, and histone H3K4me3 ChIP-seq. As input we used imputed annotation values (Ernst and Kellis 2015) and applied a quantile normal transform. For each annotation we compared the cross-validation likelihood of four models (Figure 4):

1. standard fgwas + binary annotation (peak calls)
2. standard fgwas + 3 binary annotation levels (top/mid/bot third of values *within peaks only*)
3. standard fgwas + 3 binary annotation levels (top/mid/bot third of *all annotation values*)
4. quantitative fgwas + quantitative annotation

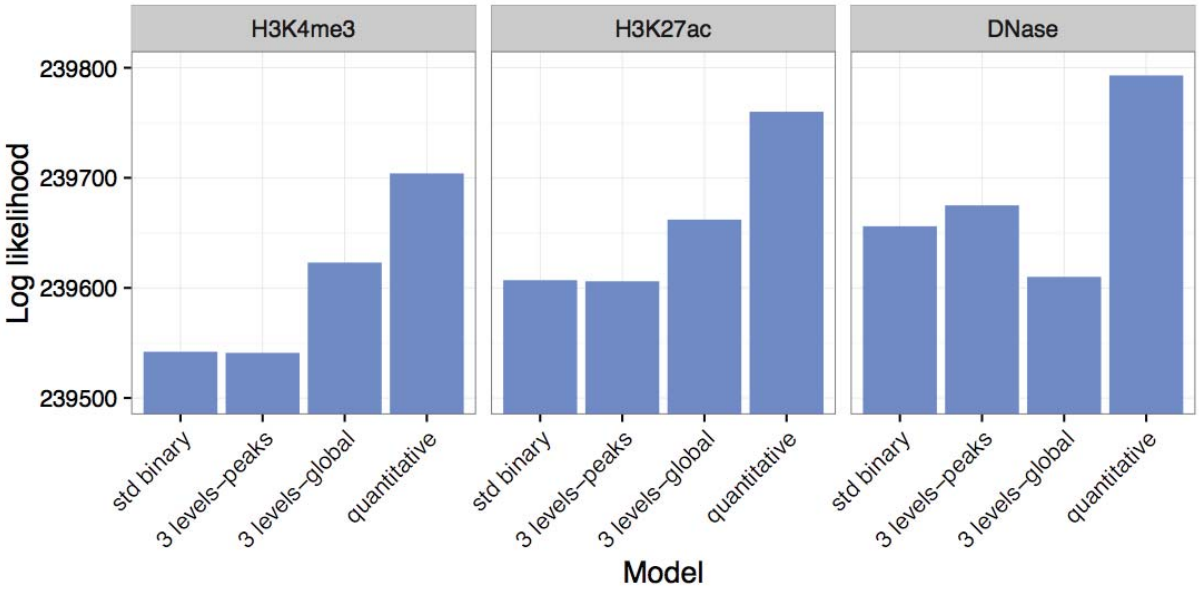


Figure 4: Cross-validation likelihood of quantitative and binary annotation models applied to three different annotations. 12-bin distance annotations were included in all models.

Models 2-4 each have three parameters, whereas model 1 has a single parameter. Model 2 is most similar to model 1 as it assigns an enrichment to SNPs within peaks only; however, SNPs in the top, middle, or bottom tertile of annotation values within peaks can receive different enrichments. Model 3 assigns an enrichment to every SNP based on its presence in the top, middle, or bottom tertile of all annotation values, regardless of peak calls. Thus, model 3 can indicate whether annotation values outside of peaks are informative. The logistic function in model 4 can be seen as a smoothed intermediate between models 2 and 3. By comparing model 4 with the other three-parameter models, we can assess whether its performance justifies the added complexity.

For all annotations tested, the quantitative model (4) was superior. Interestingly, for the two ChIP-seq annotations, the global 3-level binary annotation model (3) was better than the peaks-only 3-level model (2); however, for DNase hypersensitivity, the peaks-only model was better than the global model. By examining the parameters of the quantitative model we can get a hint as to why this might be.

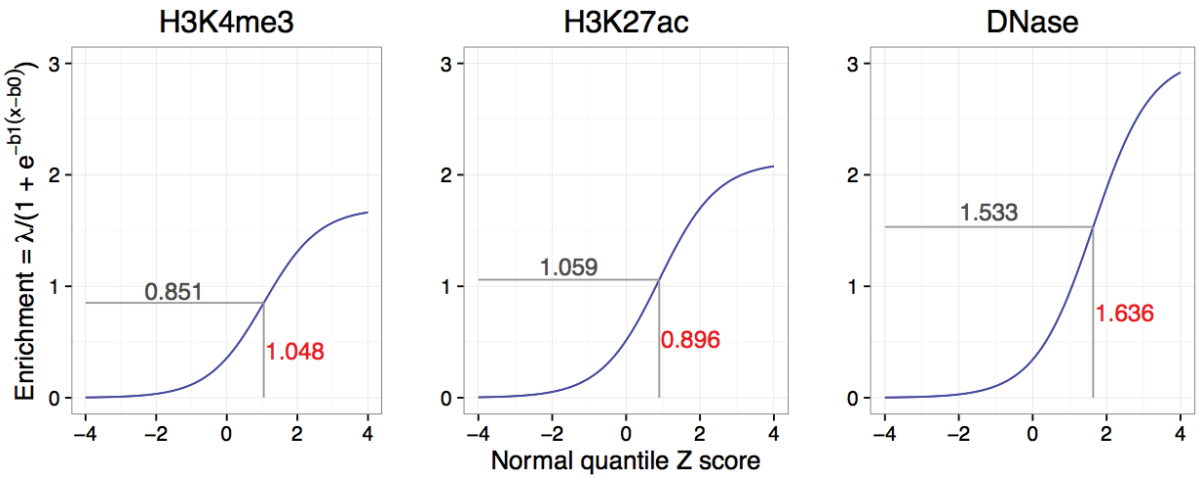


Figure 5: Parameters of the quantitative annotation model for 3 annotations. The x axis represents the normal quantile Z score across all SNPs; e.g. 95% of SNPs have annotation values between of +/- 1.96. The y axis represents the enrichment of SNPs with the highest scores relative to the lowest.

For DNase hypersensitivity, half-maximal enrichment is seen at a Z score of 1.636, which corresponds to the 95th percentile of all DNase values (Figure 5). In contrast, the half-maximal enrichments for H3K4me3 and H3K27ac occur at Z scores of 1.048 and 0.896, respectively, corresponding to the 85th and 82nd percentiles. In other words, only a small fraction of the top SNPs by DNase value are substantially enriched for causal eQTL variants, whereas enrichment of causal SNPs in H3K4me3 and H3K27ac annotations is distributed somewhat more broadly across the range of quantitative annotation values. A relevant factor

may be that DNase peaks are narrower and more numerous than both H3K27ac and H3K4me3 histone peaks. This also relates to the fraction of signal within peaks for these three annotations: whereas only 31% of DNase signal occurs within called peaks, the fraction is larger for H3K27ac (55%) and for H3K4me3 (61%). Since the global 3-level binary annotations were split into top/middle/bottom tertiles, for DNase this effectively allocated two enrichment parameters to values outside of peaks, and a single enrichment parameter within peaks (the top tertile). In contrast, for the histone modifications the split was closer to two parameters for values within peaks, and one parameter for values outside of peaks.

It is also worth noting that the 3-level within-peak annotations for H3K4me3 and H3K27ac were no better than a single binary peak annotation in terms of cross-validation model likelihood. Yet, the 3-level global annotations for the same ChIP-seq marks were considerably better. This indicates that substantial information about the location of causal variants is present in the level of these quantitative annotations *outside* of peak calls.

3.2.4 Imputed Roadmap data is more predictive for eQTLs than measured data

Two types of annotation data are provided in Roadmap Epigenomics: signal tracks from experimental assays, such as ChIP-seq and DNase-seq, and imputed signal tracks. An imputed signal track does not use any experimental data for the given tissue and assay, but instead predicts the signal based on (a) other assays in the same tissue, and (b) the same assay in different tissues. This prediction thus leverages correlations between assays in a given tissue, and between tissues for a given assay (Ernst and Kellis 2015).

LCLs were one of the cell types extensively profiled, i.e. with experimental assays and not only imputed assays. Since our eQTL training data was from LCLs, this enabled us to compare the performance of imputed and measured annotations for many assays. In most cases the imputed quantitative annotation achieved a higher model likelihood than the measured annotation for the same assay (Figure 6a), indicating that it was more informative for identifying likely causal eQTL variants. These improved likelihoods were accompanied by generally higher enrichments for the imputed annotations (Figure 6b). For the DNase hypersensitivity annotation, imputed and measured data performed similarly, while for the repressive histone marks H3K27me3 and H3K9me3, measured data performed slightly better than imputed data. This could indicate that whereas there is some redundancy in “activating” marks that can be used for imputation, repressive marks are imputed less effectively.

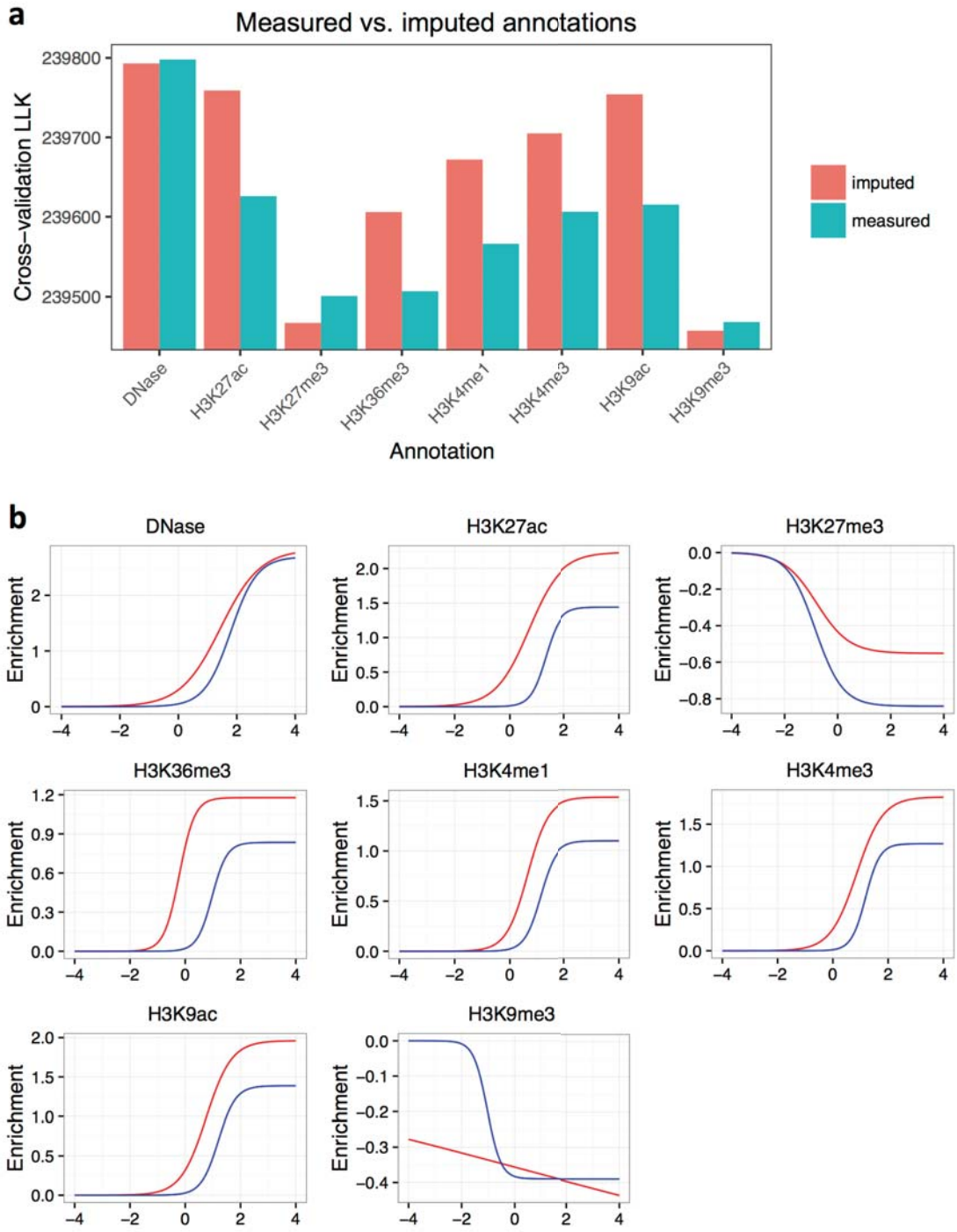


Figure 6: Imputed quantitative annotations from Roadmap Epigenomics outperform measured annotations. (**a**) Model log likelihoods, and (**b**) logistic curves defined by the optimal parameters for the same models; enrichments for imputed annotations are in red, measured annotations in blue. In all cases a 12-bin distance annotation was included. Results were similar when no distance annotation was used.

Based on these findings we chose to exclusively use imputed annotations, even though they may not be superior in every case. An important benefit is that imputed data are available for every annotation in every cell type profiled by Roadmap Epigenomics, which enabled us to extend our model to each of these cell types.

3.2.5 Interactions between annotations and gene distance

In the model implemented by fgwas, log-enrichments for SNPs are a linear combination of the enrichments for each individual annotation in which a SNP appears. We questioned whether an improvement could be made to this assumption for multi-annotation models. We might expect that certain histone marks are not equally informative at all distances from the gene TSS. For example, the histone mark H3K4me3 is enriched at active gene promoters, and is enriched for causal eQTL variants. However, when considering a given gene's expression, a high level of H3K4me3 at a distant gene is less likely to causally influence this gene than H3K4me3 at its own promoter. This represents an interaction between the histone mark annotation and a TSS distance annotation.

We hypothesized that annotation interactions with TSS distance might be widespread. We therefore created new annotations to model this interaction by splitting binary histone mark annotations into 3 distance bins: near (0 - 6,420 bp), medium (6,421 - 33,040 bp), and far (33,041 - 1 Mbp), corresponding with the first four, middle four, and last four bins of the 12-bin distance model. For example, for the "near-TSS" H3K4me3 annotation, a SNP would be assigned 1 if it is both near the TSS and in an H3K4me3 peak, and 0 otherwise. We first tested models that included both standard distance bins and these distance-interacting annotations for Roadmap binary segmentation annotations. In almost all cases, models with the distance-interacting annotations were slightly superior by cross-validation LLK to models with the binary annotation but no distance interaction (Figure 7a). The enrichment values also differed across the annotation/TSS distance interaction bins, indicating that the annotations have different levels of informativeness when they occur at different distances from a gene (Figure 7b).

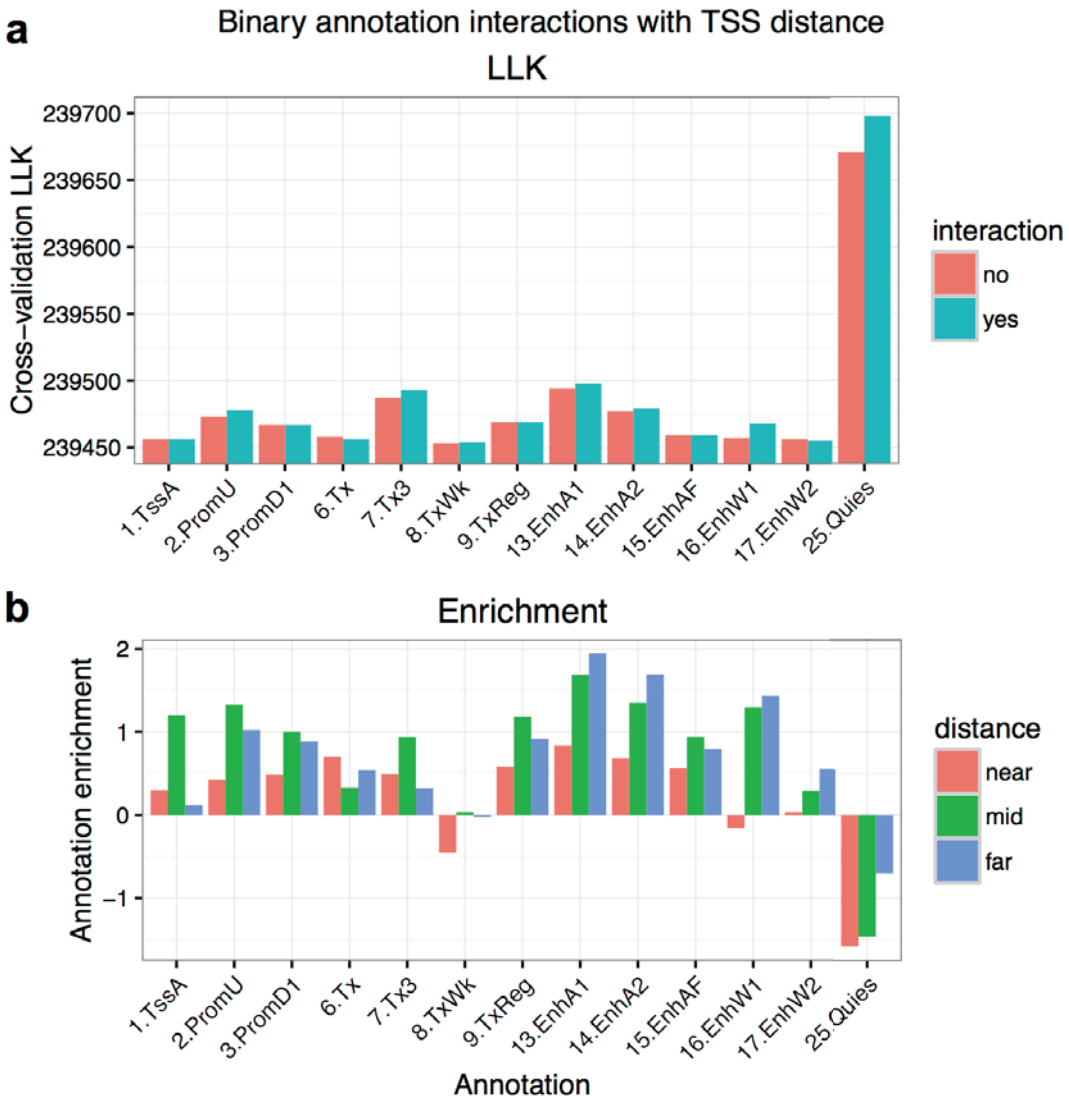


Figure 7: Modeling the interaction between binary annotations and distance to gene TSS slightly improved model performance in cross-validation. Shown are (**a**) model LLKs, and (**b**) the annotation enrichment in near/medium/far distance bins. A distance model is also included, so that the annotation*distance interaction does not reflect the general enrichment of causal SNPs near to the gene TSS.

We then included these distance-interacting annotations in the same model as quantitative annotations for the same histone marks. The results were highly similar to what was observed when only binary annotations were used, i.e. distance interaction annotations enable a small but notable improvement to model performance in cross-validation (Figure 8a).

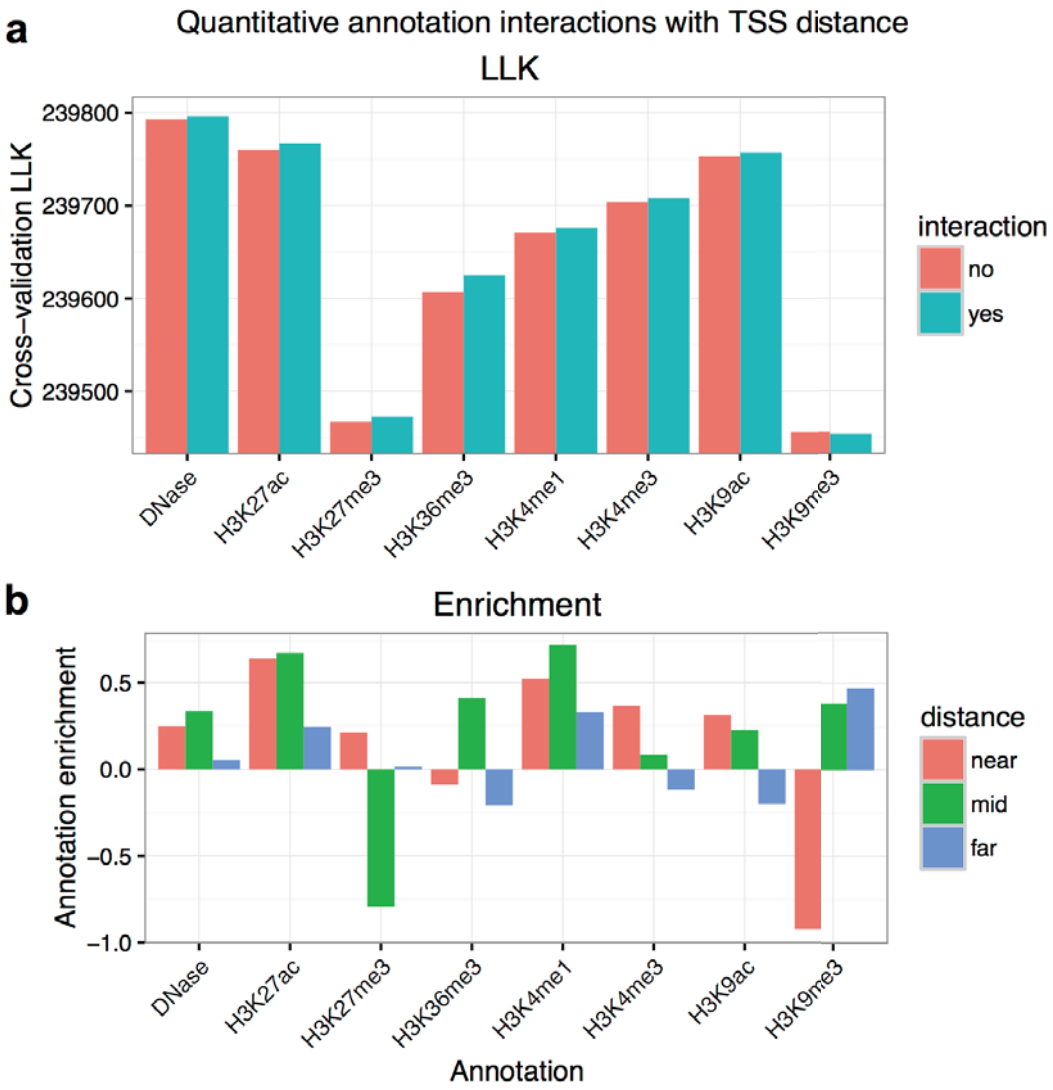


Figure 8: (**a**) For quantitative annotations, modeling the interaction between annotations and distance to gene TSS also slightly improves model performance in cross-validation. (**b**) Enrichments for causal SNPs in peaks for the quantitative annotations in (a), but split into three different distance bins (near/mid/far). A 12-bin distance model is also included.

The annotation enrichments across distance bins show some interesting patterns (Figure 7b). A group of the genome segmentation annotations show highest enrichment at medium distances (TssA, PromU, PromD1, Tx3', TxReg). In contrast, the enhancer segmentations (except for EnhAF) show highest enrichment when far from the TSS. Note that because a generic distance model was included, these are the enrichments observed over and above the general enrichment of eQTL SNPs near the gene TSS. Considering the quantitative annotations, H3K9me3 is the only assay that is not improved by considering a distance interaction. The greatest improvement is for H3K36me3 (Figure 8b), where we also see strong enrichment at medium distances, but no enrichment or mild depletion both near and far from the TSS. Since H3K36me3 reflects transcribed genomic regions, this may suggest

that causal SNPs for a focal gene's expression are slightly depleted in the transcribed regions of distal genes, and that causal SNPs are no more enriched in nearby transcribed regions than would be expected based on distance alone. For many of the other annotations, we also see a pattern of little to no enrichment in the farthest distance bin.

3.2.6 Building a multi-annotation model

3.2.6.1 Model building process

We used forward stepwise selection of annotations as outlined by (J. Pickrell 2013) to build a model containing multiple annotations, as illustrated in Figure 9. The procedure was as follows:

1. Begin with a model having 12 binary annotations for binned distance to gene.
2. Use fgwas to determine the likelihood of a with each annotation added individually.
3. Add to the model the single annotation the most improved upon the previous model's likelihood.
4. Repeat 2-4 until the model likelihood does not improve further.

At this point the model may be overfit, and so we switch to cross-validation:

5. Individually drop each annotation present in the model and determine the cross-validation likelihood.
6. Remove from the model the annotation that most improves the cross-validation likelihood when dropped (if any do).
7. Repeat 5-6 until the cross-validation likelihood does not improve further.

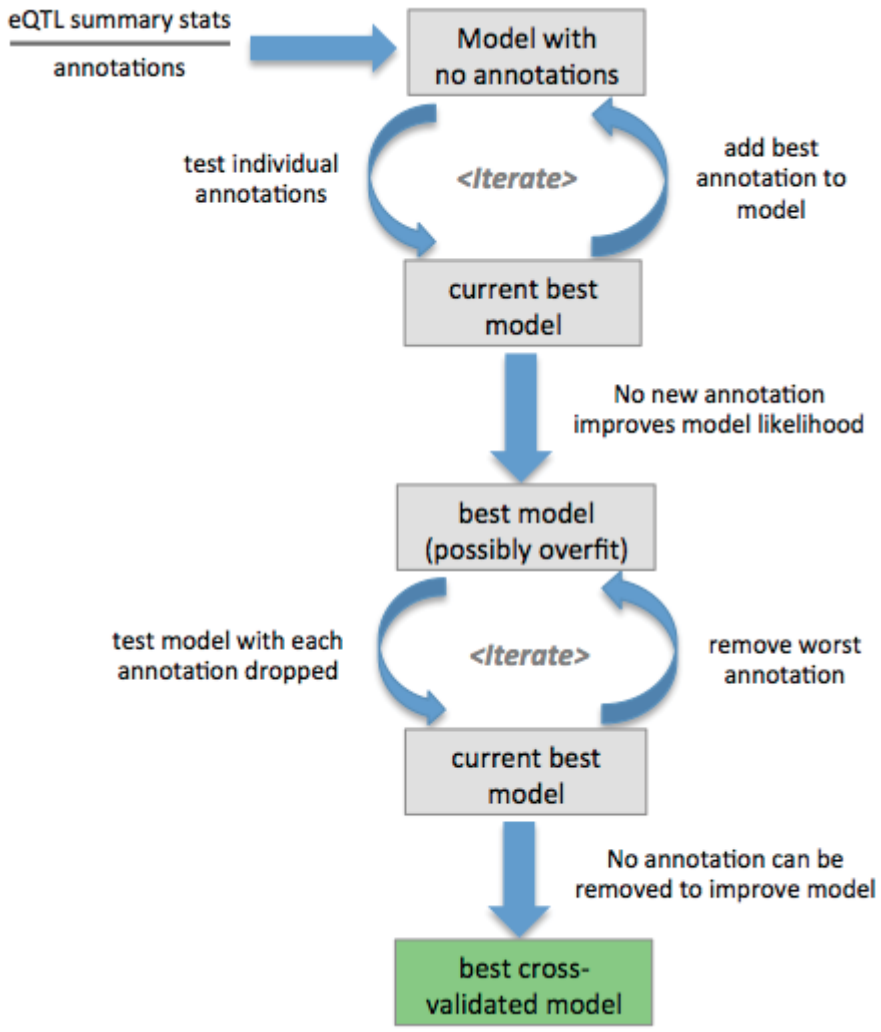


Figure 9: Schematic of model building process: forward stepwise selection to add annotations, followed by removal of annotations using cross-validation.

This model-building process is computationally intensive, as it evaluates hundreds models at each iteration, each one across thousands of genes, with up to a few thousand SNPs in the 2 Mb cis-window of each gene. To improve efficiency, we used code profiling to identify areas for optimisation in the fgwas code, and we added an additional stopping criterion that detects when the model fit is no longer improving (see Methods). These two improvements together reduced the average run-time to one third of what it was prior. Below, we describe additional design choices, involving selecting annotations and limiting the set of SNPs used in model training, that were essential to complete model-building in a reasonable time frame.

3.2.6.2 Selecting annotations

Because forward stepwise selection was used, the amount of computation needed scaled approximately linearly with the number of annotations. We therefore chose as input only annotations we deemed likely to be informative.

First, we used Roadmap quantitative annotations for which physical assays were performed in at least 30 samples (Table 2). Although imputed annotations are available for many more assays, including a number of histone modifications measured in just a handful of samples, we thought these less likely to generalise well across tissue types. Next, we selected specific binary annotations to split into TSS distance-interaction bins. We only split those annotations where an initial fgwas run with the single annotation showed an improved cross-validation likelihood when the annotation was split vs. unsplit. We used the suffixes t1/t2/t3 to indicate distance bins for annotations that are near/medium/far from the TSS (Table 3).

A group of annotations not yet described is the “centisnp” annotations, developed by the Pique-Regi group (Moyerbrailean et al. 2016), which predict the impact of genetic variants on transcription factor (TF) binding. These annotations are the only ones in our training set with resolution below that of a nucleosome (~200 nt). They are not cell type-specific, and apply only to variants present in the 1000 genomes project. However, this set of variants includes the majority of GWAS-associated variants. Centisnp refers to SNPs predicted to change a TF from bound to unbound as “switch SNPs”; those predicted to have a quantitative effect on binding are “effect SNPs”; and those within TF footprints but not predicted to affect binding are “footprint SNPs”.

To calculate the TSS distance annotation for a given variant and gene, we determined the minimum distance of the variant to FANTOM TSSes of the gene with an expression level of at least 2.0 transcripts per million (TPM). Some genes did not have any expressed FANTOM TSS, yet had nonzero expression in the Geuvadis LCLs. In this case we used the minimum distance to any Gencode TSS. The FANTOM consortium also reported that bidirectional transcription is a hallmark of active enhancers, and they produced a compendium of such enhancers in the same tissue types as their TSS definitions (Andersson et al. 2014). These enhancers include quantitative information on the level of transcription, and we used them as a quantitative annotation in our model.

We also included gene annotations from Gencode (Harrow et al. 2012), and evolutionary conservation values from GERP (Davydov et al. 2010). To maximize the informativeness of

gene annotations (UTR, coding, intron), we split these annotations depending on whether they are for the gene under consideration for a given eQTL or not, leading to annotations labeled e.g. “intron.samegene” and “intron.diffgene”. The full set of annotations used in model training is provided in Tables 1, 2, and 3.

Table 1: Distance annotations used in model training

TSSDist 0-285
TSSDist 286-1410
TSSDist 1411-3690
TSSDist 3691-6420
TSSDist 6421-10338
TSSDist 10339-15106
TSSDist 15107-21855
TSSDist 21856-33040
TSSDist 33041-49943
TSSDist 49944-85174
TSSDist 85175-164469

Table 2: Quantitative annotations used in model training

DNase	H3K27ac	effect-snp-num_motifs
H3K4me1	H3K27me3	footprint-snp-num_motifs
H3K4me3	H3K36me3	switch-snp-num_motifs
H3K9ac	DNAMethylSBS-fraction	Fantom enhancer TPM
H3K9me3	GerpRS-noncoding only	

Table 3: Binary annotations used in model training

Annotations beginning “Seg” are the 25-state Roadmap segmentation states.

Annotations ending t1/t2/t3 indicate that the annotation is only positive in the given distance bin from the TSS.

Gencode-antisense	Seg-6.Tx.t2	Seg-15.EnhAF.t2	effect-snp.t3
Gencode-coding.diffgene	Seg-6.Tx.t3	Seg-15.EnhAF.t3	footprint-snp
Gencode-coding.samegene	Seg-7.Tx3	Seg-16.EnhW1	switch-snp
Gencode-intron.diffgene	Seg-7.Tx3.t1	Seg-16.EnhW1.t1	DNase.t1
Gencode-intron.samegene	Seg-7.Tx3.t2	Seg-16.EnhW1.t2	DNase.t2
Gencode-lincRNA	Seg-7.Tx3.t3	Seg-16.EnhW1.t3	DNase.t3
Gencode-miRNA	Seg-8.TxWk	Seg-17.EnhW2	H3K27ac.t1
Gencode-rRNA	Seg-8.TxWk.t1	Seg-17.EnhW2.t1	H3K27ac.t2
Gencode-sense_intronic	Seg-8.TxWk.t2	Seg-17.EnhW2.t2	H3K27ac.t3
Gencode-sense_overlapping	Seg-8.TxWk.t3	Seg-17.EnhW2.t3	H3K27me3.t1
Gencode-snoRNA	Seg-9.TxReg	Seg-18.EnhAc	H3K27me3.t2
Gencode-snRNA	Seg-9.TxReg.t1	Seg-19.DNase	H3K27me3.t3
Gencode-UTR3.diffgene	Seg-9.TxReg.t2	Seg-2.PromU	H3K36me3.t1
Gencode-UTR3.samegene	Seg-9.TxReg.t3	Seg-2.PromU.t1	H3K36me3.t2
Gencode-UTR5.diffgene	Seg-10.TxEnh5	Seg-2.PromU.t2	H3K36me3.t3
Gencode-UTR5.samegene	Seg-11.TxEnh3	Seg-2.PromU.t3	H3K4me1.t1
Seg-1.TssA	Seg-12.TxEnhW	Seg-20.ZNF_Rpts	H3K4me1.t2
Seg-1.TssA.t1	Seg-13.EnhA1	Seg-21.Het	H3K4me1.t3
Seg-1.TssA.t2	Seg-13.EnhA1.t1	Seg-22.PromP	H3K4me3.t1

Seg-1.TssA.t3	Seg-13.EnhA1.t2	Seg-23.PromBiv	H3K4me3.t2
Seg-3.PromD1	Seg-13.EnhA1.t3	Seg-24.ReprPC	H3K4me3.t3
Seg-3.PromD1.t1	Seg-14.EnhA2	Seg-25.Quies	H3K9ac.t1
Seg-3.PromD1.t2	Seg-14.EnhA2.t1	Seg-25.Quies.t1	H3K9ac.t2
Seg-3.PromD1.t3	Seg-14.EnhA2.t2	Seg-25.Quies.t2	H3K9ac.t3
Seg-4.PromD2	Seg-14.EnhA2.t3	Seg-25.Quies.t3	H3K9me3.t1
Seg-5.Tx5	Seg-15.EnhAF	effect-snp.t1	H3K9me3.t2
Seg-6.Tx	Seg-15.EnhAF.t1	effect-snp.t2	H3K9me3.t3
Seg-6.Tx.t1			

3.2.6.3 Limiting training data to improve speed

For each of the 6,340 protein-coding genes used in model training, all SNPs in a 2 Mb window were tested for association with the gene's expression, a total of 39,566,693 tests. Even with the optimisations described above, running the model-building process with fgwas would take months to compute. While we could train the model on a small subset of genes, the results would depend more strongly on the particular genes selected. Moreover this would to a certain extent defeat the purpose of using eQTL data, where we have a large number of associations. We instead explored training the model using all genes, but with a subset of SNPs for which the association statistic was above a certain threshold. Because most SNPs are not associated with expression, this could improve the runtime dramatically.

To assess whether filtering variants based on association statistic would change model-building results, we selected 1,000 genes with a lead variant having $p < 1x10^{-12}$. We then determined the approximate Bayes factors (BFs) for variants, and created filtered datasets having only variants with BF > 10, or with a BF > 100. For a p value of $1x10^{-12}$, the equivalent BF is $\sim 4.7x10^{9}$, and so the variants filtered out are unlikely to be causal. Whereas the full 1,000-gene dataset had 6,362,813 variant tests, there were just 582,975 variants with BF > 10, and 383,442 variants with BF > 100. We applied the model-building process described previously, separately for the full and filtered datasets, stopping after 10 iterations. The annotations that were added to these three models are shown in Table 4.

	Full dataset	Filtered dataset BF > 10		Filtered dataset BF > 100	
	annotations added	annotations added	Order in full data	annotations added	Order in full data
1	DNase	DNase	1	DNase	1
2	H3K36me3	H3K36me3	2	State.25.Quies	5
3	intron.diffgene	intron.diffgene	3	UTR3.samegene	6
4	H3K9me3	H3K9me3	4	coding.samegene	10
5	State.25.Quies	Enh.Fantom	7	intron.diffgene	3
6	UTR3.samegene	UTR3.samegene	6	UTR5.diffgene	-
7	Enh.Fantom	H3K27ac	-	H3K27ac	-
8	effect-snp.nmotifs	coding.samegene	10	H3K27me3	-
9	H3K9ac.t3	UTR5.diffgene	-	State.1.TssA.t3	-
10	coding.samegene	effect-snp.nmotifs	8	Enh.Fantom	7

Table 4: The order in which annotations are added when model-building with three different training datasets. For annotations in the filtered datasets, we show the order in which the same annotation was added in the full 1,000-gene dataset.

Although the annotations added during model-building were similar, they were not identical. In addition, in the filtered datasets the enrichments reported are lower across all annotations than in the full dataset. We evaluated the performance of the three models shown in Table 4 using cross-validation, applied to either the same 1,000 genes or to a separate set of 1,000 genes. In these validation comparisons no variants were filtered out, and the only difference between models was which annotations were included. The models built using filtered data did not perform as well in cross-validation on the 1,000 genes they were trained on as did the model trained with all SNPs. However, for the independent set of genes, the model trained on the BF > 10 filtered dataset actually performed better in cross-validation than the model trained using all SNPs (Figure 10). Based on this, we believe that performing model-building with low-BF SNPs filtered out is an effective optimisation that is likely to result in a similar-performing model in external validation. We proceeded with building a full model on the 6,340 eQTL genes, using only variants with BF > 10.

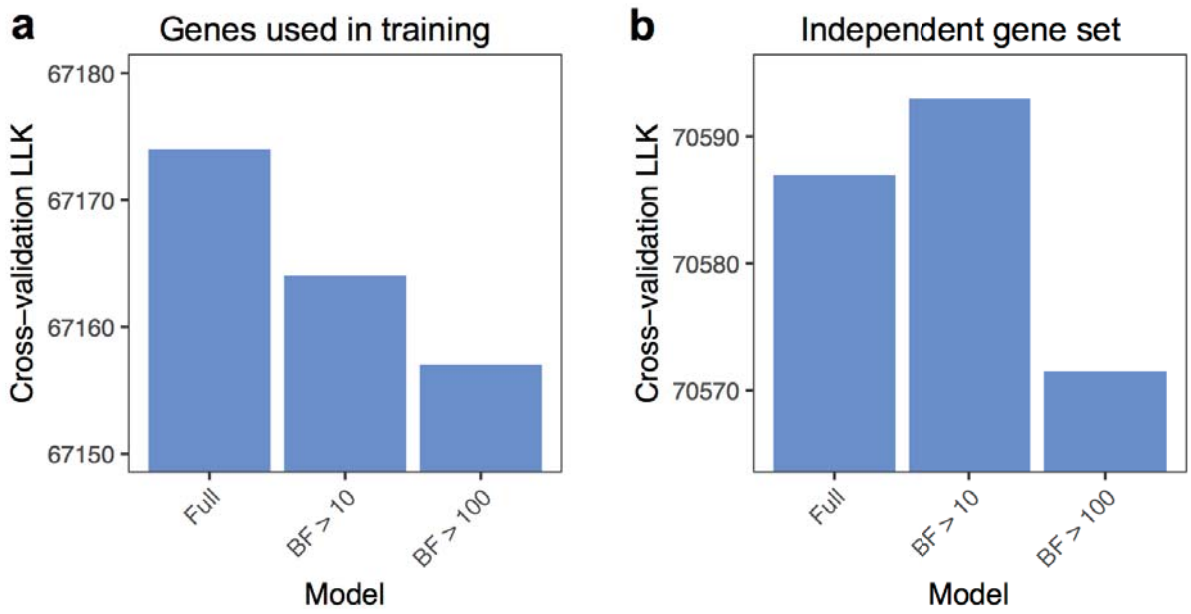


Figure 10: Cross-validation LLK of models shown in Table 4, for either (**a**) the set of 1,000 genes used in training, or (**b**) a separate 1,000 genes.

3.2.6.4 A final model with 38 annotations

Applying the model-building process described previously, we added annotations sequentially to the model for 40 iterations, after which the model likelihood no longer increased. Model LLKs plateaued once the 38th annotation was added (Figure 11).

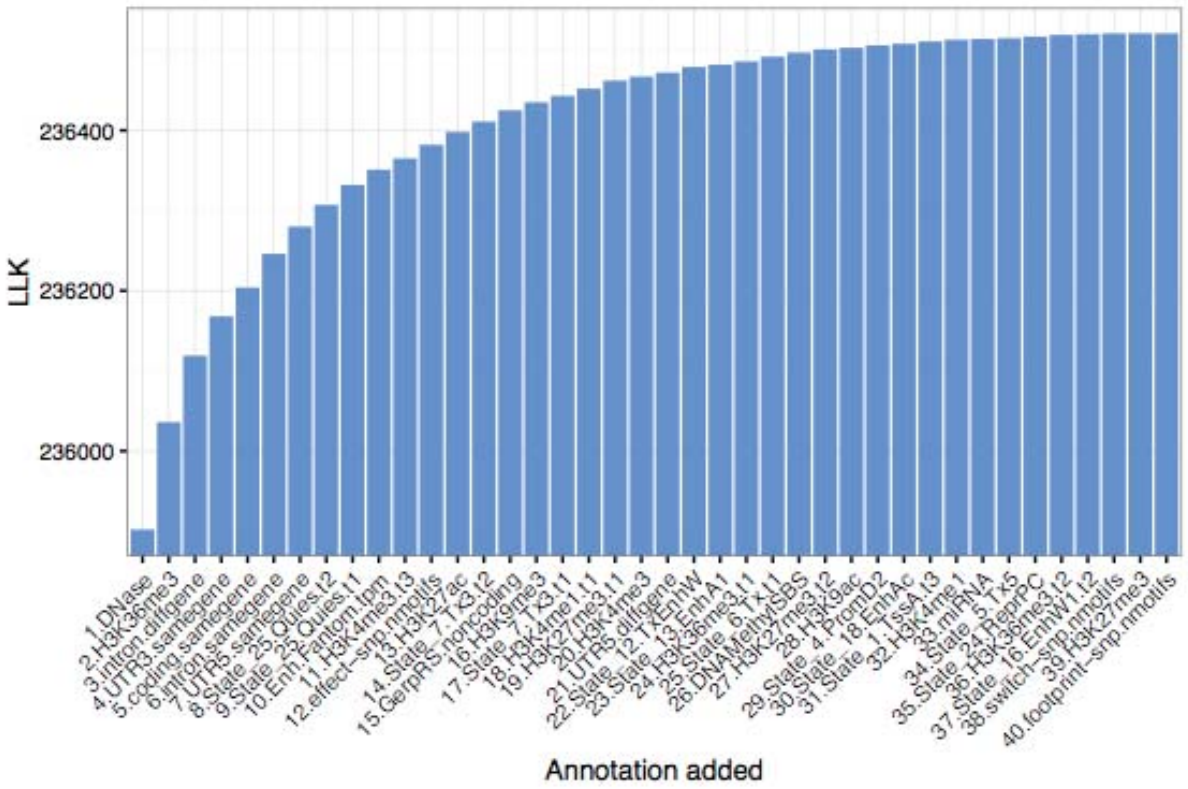


Figure 11: Model LLK with each annotation that was added over 40 iterations. Each model includes 12 distance bins, as well as all the annotations added prior to it.

We selected the first 38 annotations, and switched to using cross-validation, testing models with each annotation individually dropped. Surprisingly, none of the annotations could be removed without worsening the cross-validation likelihood, as shown in Figure 12. We also observed that when considering annotations to drop, the importance by cross-validation likelihood was not the same as the order in which they were added. This reflects the fact that annotations are correlated, and one annotation could be substituted by a combination of other annotations, which may have been added later on in the model-building process.

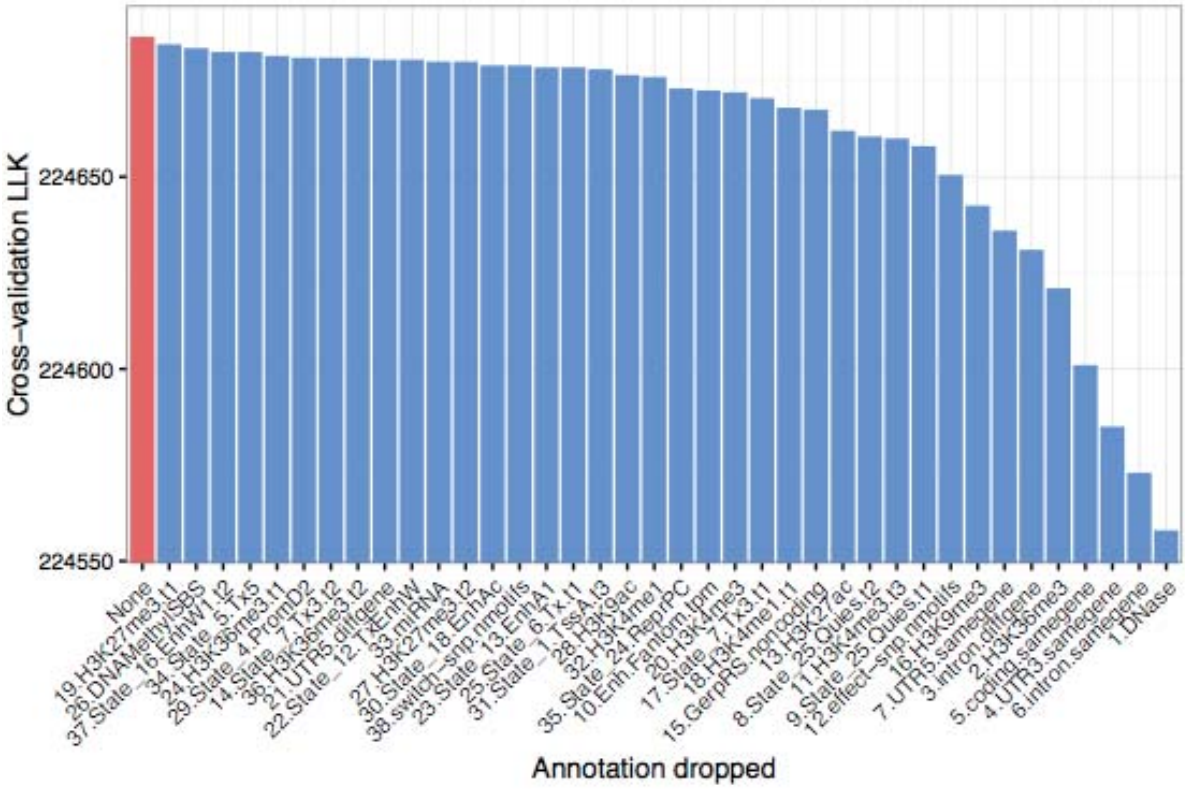


Figure 12: Model cross-validation LLK when each annotation is individually dropped. The full model with no annotations dropped is on the left, and no annotation can be dropped without reducing the LLK. For the x-axis labels, the number preceding each annotation name is the order in which it was added to the model.

Since no annotations could be dropped, the full set of 38 annotations and their associated enrichments is our most predictive model. Annotation enrichments are illustrated in Figure 13, and full details are reported in Appendix B. For some annotations, the confidence interval for their enrichment overlaps zero. While this argues for dropping them from the model, cross-validation supported keeping them in. We note that in a combined model with many annotations, the enrichment for each individual annotation is compensated by adjustments to other annotation enrichments.

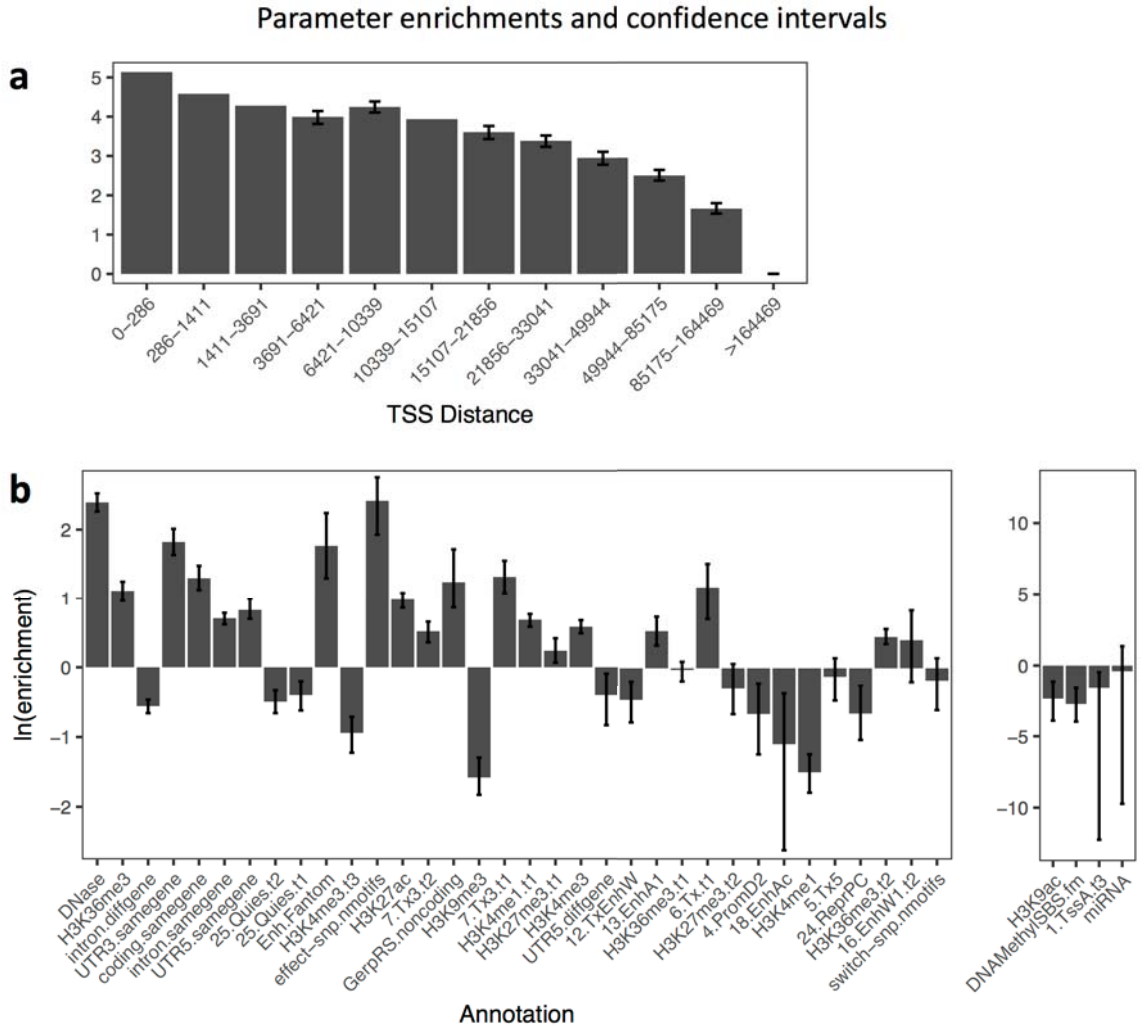


Figure 13: Enrichments for (**a**) TSS distance annotations, and (**b**) the 38 other annotations included in the final model. Confidence intervals were determined by individually adjusting annotation enrichments until the model LLK is decreased by 2 units. Fgwas failed when computing confidence intervals for some of the TSS distance annotations. In (b), four annotations are shown separately at right with a different scale, due to large confidence intervals.

Given enrichments for our 38 annotations, we can compute the PRF score for any variant in any of 119 Roadmap epigenomes. We next look at how PRF scores are defined based on the model parameters, and how they vary across the genome.

3.2.7 Distribution of PRF scores

The PRF score for a variant is the sum of enrichments for the variant's annotations. This is the value x_i in the equations below, repeated here for convenience.

$$\pi_{ik} = \frac{e^{x_i}}{\sum_{j \in S_k} e^{x_j}}$$ (Equation 2)

$$x_i = \sum_{l=1}^{L_2} \lambda_l I_{il}$$ (Equation 3)

The prior probability for a variant to be causally associated with gene expression, π_{ik}, involves x_i in the exponent, and thus the PRF score is proportional to the logarithm of the probability that the variant is associated. A variant's prior can only be computed relative to a defined set of SNPs, S_k, in a region around a gene of interest. This prior probability depends on the assumption that the causal variant is within the set of variants considered, and moreover it will change if the set of variants considered changes. The same is not true for the PRF score itself: although the PRF score for a variant depends on the gene being considered, it does not depend on the other variants considered.

We demonstrate some of important features of this approach in Figure 14, which shows the distribution of PRF scores in the vicinity of *SMAD3*. PRF scores tend to peak near the TSS of genes, and are higher in annotation-dense regions such as enhancers. PRF scores also tend to be higher within the body of the gene they are proposed to regulate. While *SMAD3* has many alternative annotated TSSes, these were not expressed in FANTOM LCLs, and so PRF scores are not elevated near these TSSes. This kind of information would be difficult to glean by manually exploring annotations in a genome browser. Zooming in to a 5 kb region upstream of *SMAD3*, shown in Figure 15, fine-grained variation in PRF scores is seen. The scores vary according to quantitative differences in histone modification levels, even at low values that might not be within called peaks.

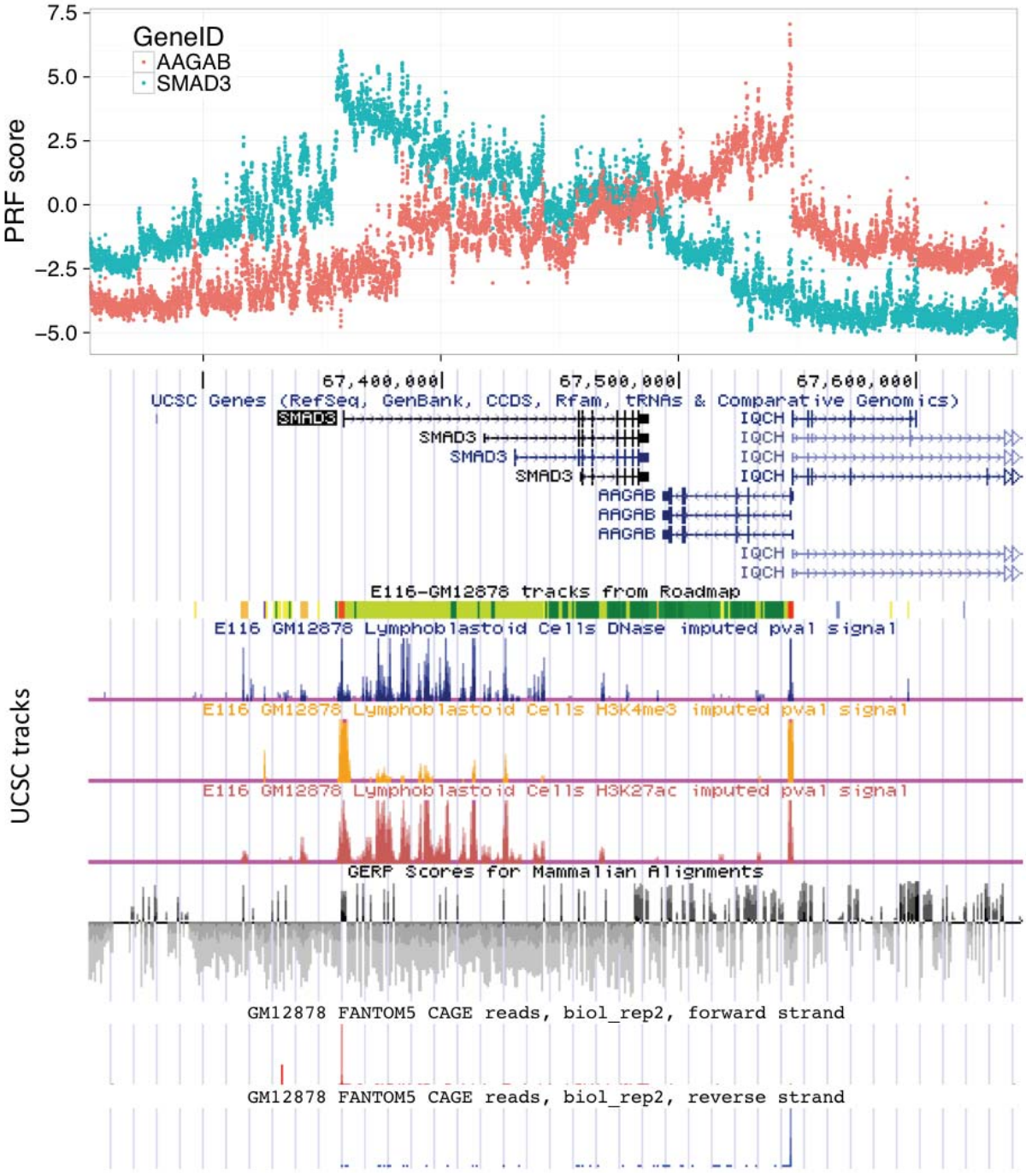


Figure 14: PRF scores for positions in a 400 kb window around *SMAD3*. Scores for two genes are shown, but multiple other genes within 1 Mb also have scores in the region. PRF scores peak towards the TSS of genes. Here, only two TSSes are used, which are visible in the FANTOM5 CAGE reads track.

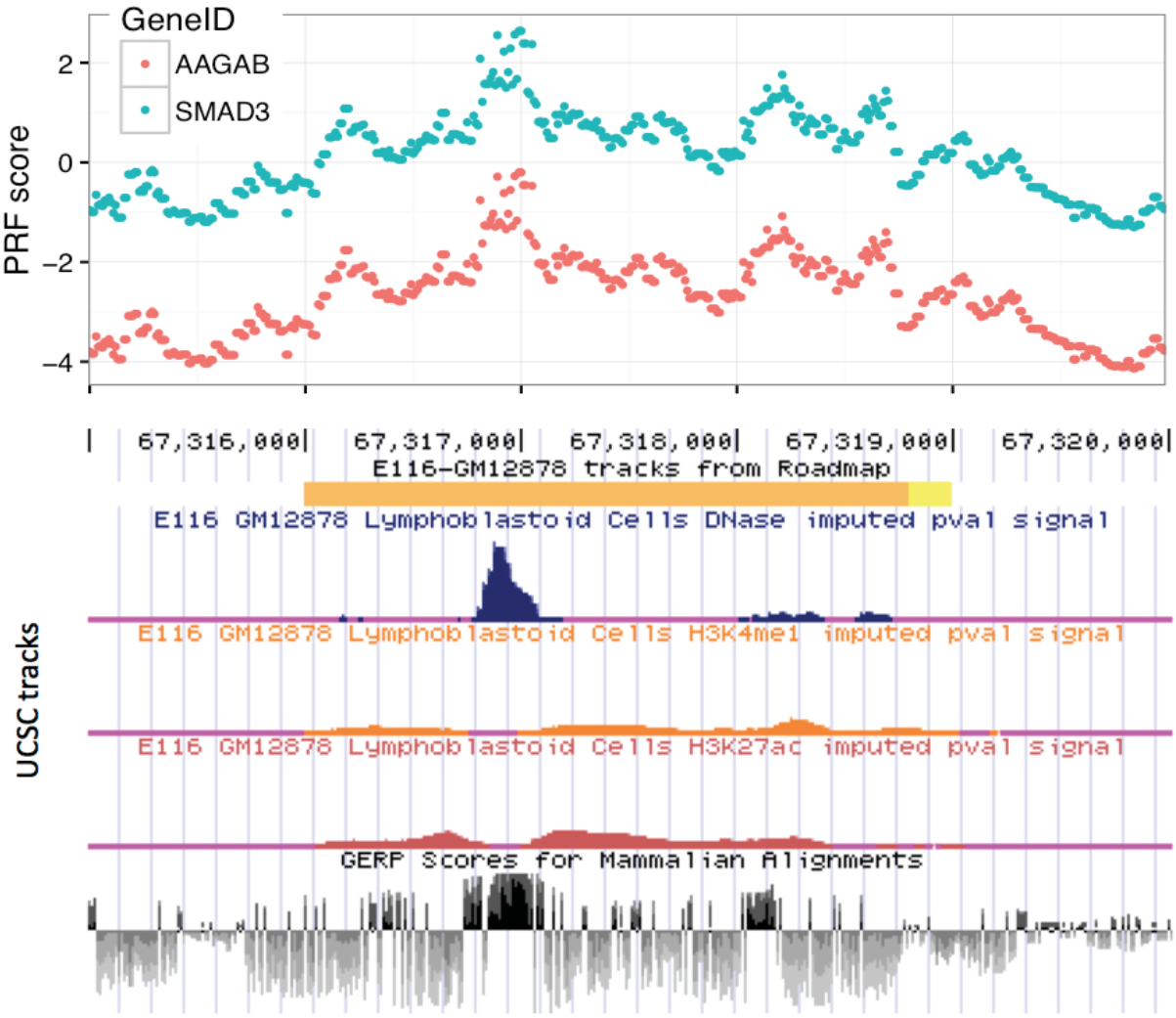


Figure 15: PRF scores in a 5 kb region upstream of *SMAD3*.

Each variant has a different PRF score for each gene within 1 Mb. This makes PRF scores well-suited to fine-mapping likely causal eQTL variants, but complicates the application to GWAS, where the relevant gene is not known at most associated loci. In the remainder of this chapter we discuss applying PRF scores to eQTL studies, and in Chapter 4 we apply PRF scores to GWAS.

3.3 Validation with eQTL data

3.3.1 Comparing score distributions

We wanted to show that PRF scores can be used to predict which genetic variants causally influence gene expression and, ideally, by extension influence complex traits. We refer to two types of PRF scores — “gene PRF” scores, in which the PRF score specific to a given gene is assigned to a variant, and “max PRF” scores, in which we assign to a variant the maximum PRF score for any gene in its 1 Mb window. GenePRF scores are useful when the relevant gene is known, but when it is unknown then maxPRF scores must be used.

We first compared the distribution of PRF scores for cis-eQTL variants with those of CADD and GWAVA, two leading methods providing genome-wide scores for non-coding variants. Since the PRF score model was trained on Geuvadis data, we used eQTL data from the GTEx project, beginning with subcutaneous adipose tissue. We selected the 2,493 adipose eGenes where the best variant had association $p < 1x10^{-12}$. For each eGene, we determined the posterior probability of association (PPA) for all tested variants, using the method of fgwas with statistical information only (i.e. no annotations). We used the adipose nuclei epigenome (Roadmap E063) to compute genePRF and maxPRF scores, and examined these scores for variants in different bins of posterior probability (Figure 16). PRF scores were higher on average for variants with higher PPA ($p < 1x10^{-300}$, Kruskal-Wallis test). This was also true for CADD and GWAVA, although the distributions of each score differed.

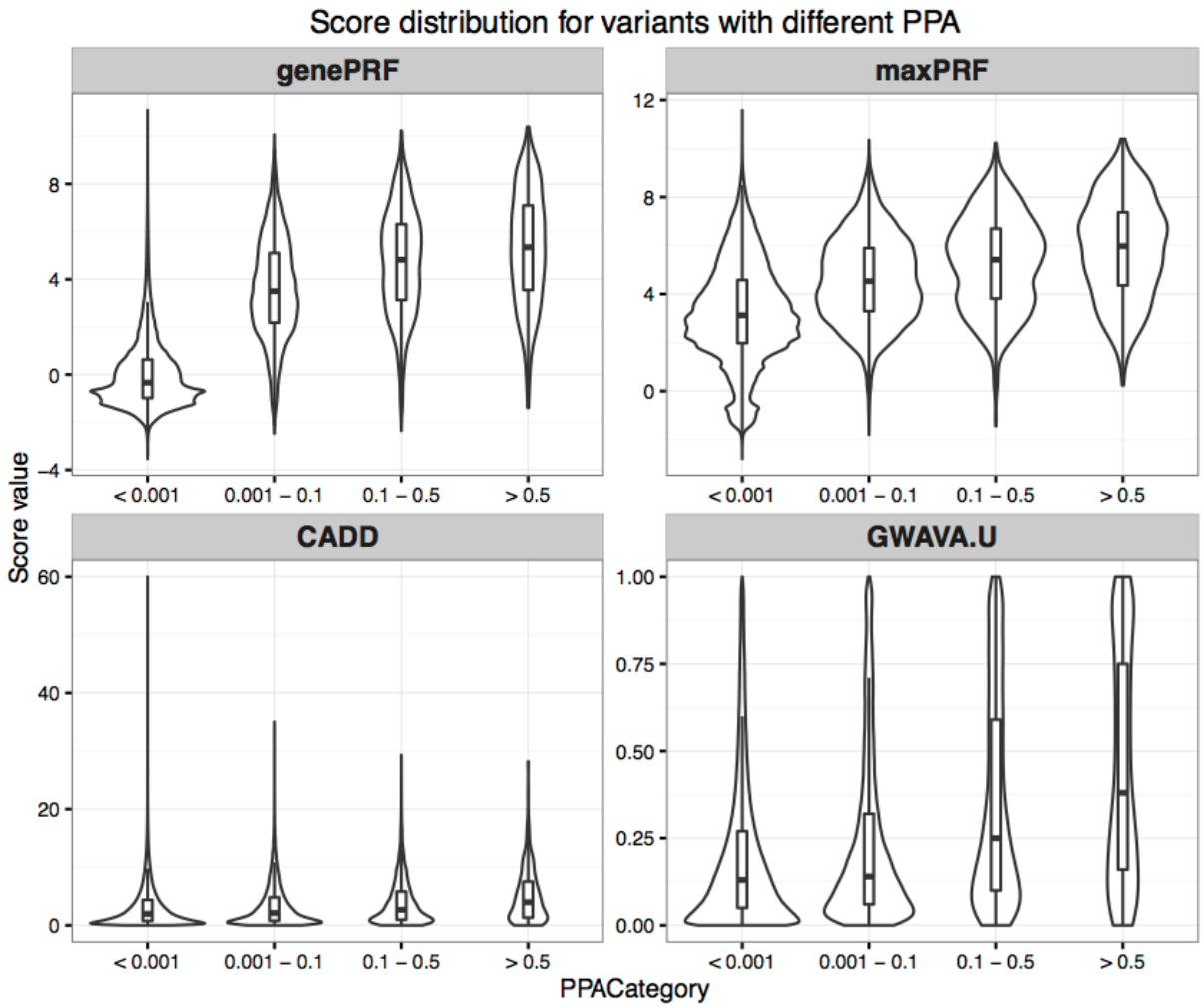


Figure 16: Scores for variants tested for association with gene expression in GTEx subcutaneous adipose tissue, stratified by posterior probability of association, for 4 different scoring methods: genePRF, maxPRF, CADD, and GWAVA.U. PRF scores were calculated using the E063 adipose nuclei epigenome. The numbers of variants in each PPA category are N=11,297,373 (PPA<0.001), N=41,574 (PPA 0.001-0.1), N=3,063 (PPA 0.1-0.5), and N=565 (PPA>0.5).

We next used more formal metrics to assess PRF score prediction performance. A brief introduction to these methods is provided in Appendix C.

3.3.2 Classifying lead variants

The PRF score can be treated as a binary classifier, with variants above some threshold score predicted as causal (“positive class”), and those below this score predicted as non-causal (“negative class”). Ideally, to define true positive cases we would use an external set of known expression-altering variants. However, there is no gold-standard set of genetic variants known to causally influence gene expression in specific cell types. In its absence, we must settle for a positive set that is enriched for causal variants. We therefore use lead

eQTL variants as a proxy for causal variants, and ask how well PRF scores discriminate lead variants from other variants. Many lead variants will not in fact be causal, and as a result we will likely underestimate PRF score prediction performance. In taking lead variants as causal, we also implicitly assume that there is a single causal variant per gene. Stepwise conditional regression has revealed that at current sample sizes, a significant minority of human genes are detectably regulated by multiple variants (Lappalainen et al. 2013). In such a case, even if the lead variant is causal, the presence of additional causal variants not in the positive ground truth set will lead to underestimation of prediction performance.

We compared genePRF and maxPRF scores with CADD and GWAVA using receiver-operating characteristic (ROC) curves; here, an area under the curve (AUC) above 0.5 indicates prediction performance better than chance. GWAVA defined scores for three classifiers, namely, GWAVA.TSS for a model that matched SNPs based on TSS distance, GWAVA.U which did not match on TSS distance, and GWAVA.R which matched on TSS distance and genomic region. For each of the scores we considered performance in identifying lead variants for GTEx subcutaneous adipose eGenes (with $p < 1x10^{-12}$) from among all variants within 1 Mb (Figure 17a). GenePRF scores (AUC=0.951) far outperformed other scores in prioritising lead variants (AUC 0.565 - 0.765). Achieving an AUC above 0.9 indicates very good classification performance, which may be surprising given that we expect only a modest fraction of lead variants to be causal. There is a simple explanation -- because PRF scores were trained using eQTLs, they heavily upweight variants near the TSS of genes, and lead eGene variants also cluster near the genes they regulate. Therefore, the problem of distinguishing lead variants is made easier because most distal variants can be discounted. Consistent with this, GWAVA.U scores, which weight TSS distance more heavily, performed better than the other GWAVA scores (Figure 17a).

An alternative performance measure considers precision (the fraction of cases predicted positive which are true positives) as a function of recall (the fraction of all true positives identified as such, also known as the true positive rate). Even with a high ROC AUC, the precision-recall curve for genePRF scores showed a precision of only 1% at a PRF score threshold where 25% of lead variants are identified (Figure 17b). The other scores similarly had very poor precision in their predictions. We are thus a long way from being able to precisely pinpoint causal variants from annotation data alone when considering a large number of candidate SNPs in a window around a gene.

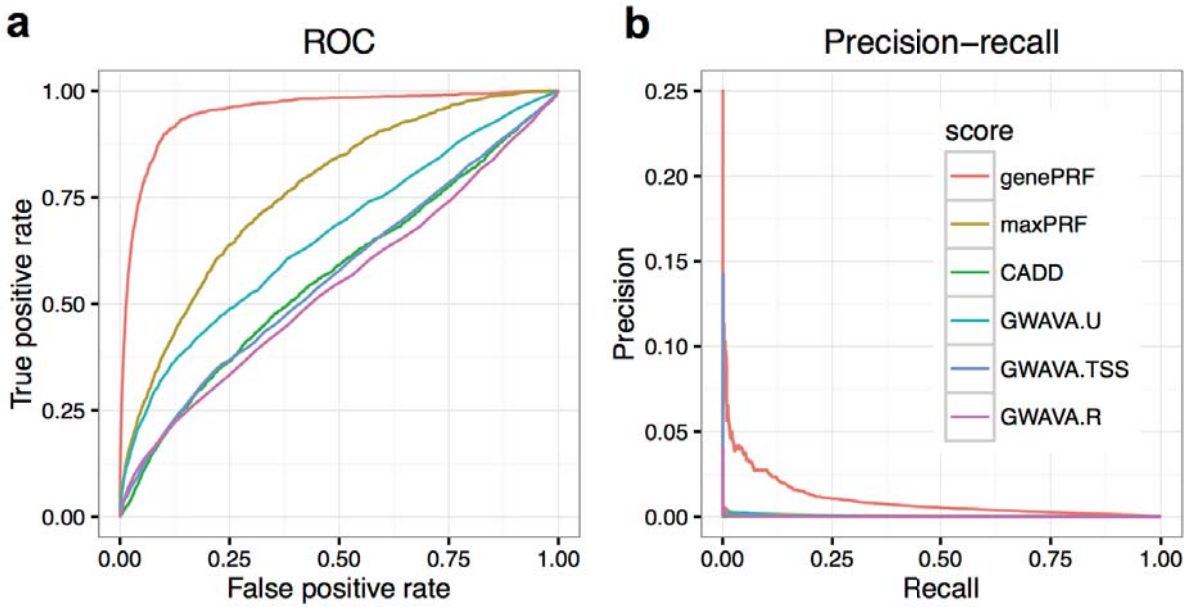


Figure 17: (**a**) ROC curves, and (**b**) precision-recall curves for identifying lead variants (with $p < 10^{-12}$) for eGenes in GTEx subcutaneous adipose tissue from among all variants within 1 Mb.

Rather than considering all variants in a large window, a more relevant measure of performance may be how well PRF scores discriminate lead variants from among candidate causal variants for the association. To assess this, we considered eGenes with a "confident causal" variant, defined as a single variant with PPA > 0.5 when using fgwas with no annotations. We plotted ROC and precision-recall curves for distinguishing the lead variant from the top 20 variants by statistical association for each eGene (Figure 18). Note that although we could use a threshold on PPA rather than fixing the number of variants at 20, we avoid this because the performance would be harder to interpret: some genes have dozens of variants with PPA > 0.01, whereas others have a single variant.

In this "fine-mapping" scenario, the ROC AUC for PRF scores (0.678) was dramatically worse than when all variants within 1 Mb are considered. The drop in performance is unsurprising, since TSS distance is less likely to distinguish among variants at a single association peak. Still, both gene-aware and gene-agnostic PRF scores performed slightly better than competing methods CADD and GWAVA. Interestingly, in this scenario PRF score precision improved to 17% when 25% of the lead variants were identified. This is because the positive and negative classes were less imbalanced in this scenario -- 1 in 20 variants was positive, compared with 1 in ~5000 when all variants in the 1 Mb cis-window of a gene were considered.

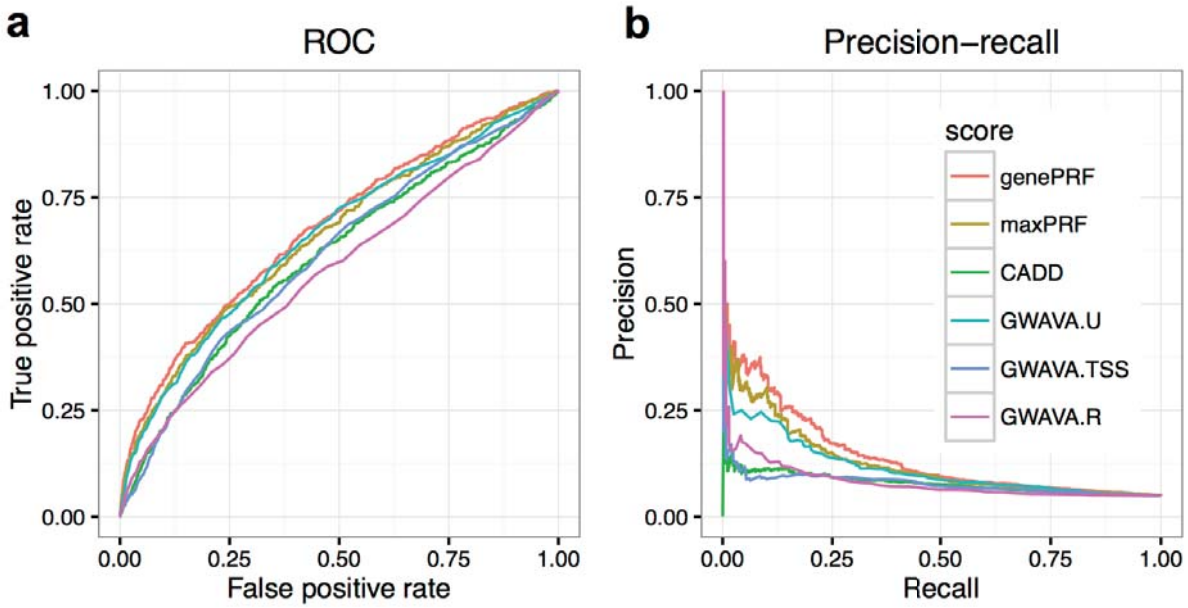


Figure 18: (**a**) ROC curves, and (**b**) precision-recall curves for identifying lead variants in GTEx subcutaneous adipose tissue for "confident eGenes" where the lead variant has a statistical PPA > 0.5. To represent a fine-mapping scenario, only the top 20 variants by statistical association are considered for each eGene.

A final metric that we considered was lift, which indicates how enriched the variants at a given prediction threshold are for true positives, relative to the same number of randomly chosen variants. When considering all cis-window variants for GTEx subcutaneous adipose eGenes, those in the top 1% of genePRF scores were 42-fold more likely than chance to be lead variants, while variants in the top 10% were 9-fold more likely (Figure 19a). Both genePRF and maxPRF scores considerably outperformed CADD (top 1% having lift 1.9) and GWAVA (top 1% having lift 10.4 for the best GWAVA score). When only the top 20 variants per gene were considered, those in the top 1% and 10% of genePRF scores were 7.3-fold and 3.1-fold more likely than chance to be lead variants (Figure 19b). In this scenario, maxPRF scores performed nearly as well as genePRF scores, since TSS distance was less informative as a predictor. GWAVA.U also performed well (top 1% and 10% of scores having lift of 5.0 and 2.7), but CADD performed poorly (top 1% and 10% of scores having lift of 2.5 and 2.0).

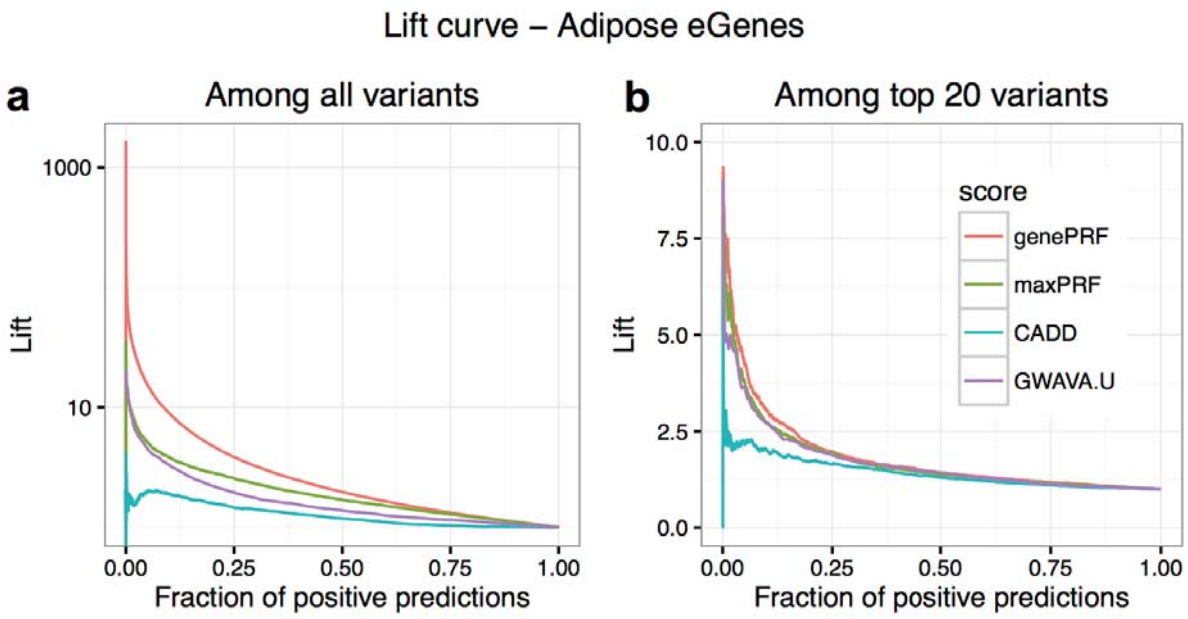


Figure 19: Lift curves for identifying lead eQTL variants in GTEx subcutaneous adipose from among (**a**) all variants in the 1 Mb cis-window or (**b**) the top 20 statistically associated variants. In panel (**a**) the lift values are plotted on a logged axis because they vary over orders of magnitude.

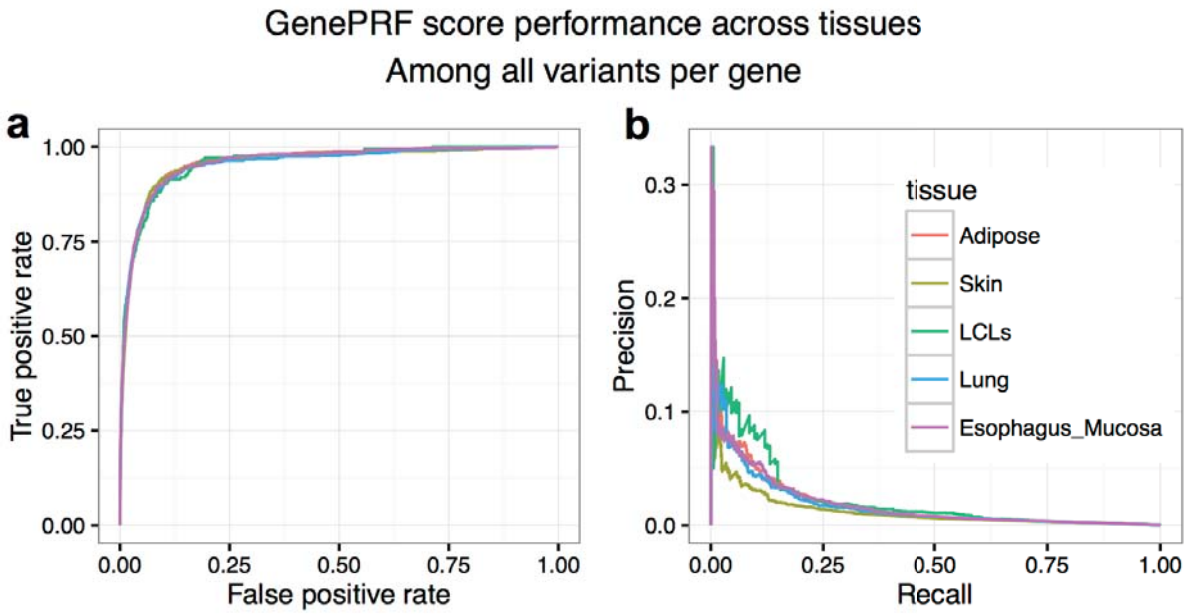


Figure 20: (**a**) ROC curves, and (**b**) precision-recall curves for genePRF scores in identifying lead variants in GTEx tissues from among all variants within 1 Mb. PRF scores were computed for each GTEx tissue using the most relevant epigenome. Performance on different GTEx tissues is indicated by color.

Although we used GTEx subcutaneous adipose tissue for these comparisons, the pattern was very similar for four other GTEx tissues which we investigated (Figure 20). These investigations show that PRF scores provide a genome-wide summary of regulatory information in specific cell types, which can be useful in predicting the locations of gene regulatory variants.

3.3.3 Fine-mapping: reducing credible set size

A measure of the utility of PRF scores is their ability to assist in fine-mapping causal variants. A common way to describe how finely an association signal has been localised is to consider the size of the credible set - the set of variants expected to contain the causal variant with a specified probability. The variants in the set can be determined by computing the PPA for each variant in the region, and then adding variants to the set (beginning with the most associated) until in sum they reach a specified probability of containing the causal variant, commonly either 95% or 99%. The credible set at a locus can be defined using statistical information alone, or with the inclusion of annotation information. When an annotation is informative we expect that statistical and annotation evidence should coincide, and thus incorporating annotations, summarised by PRF scores, should lead to smaller credible sets.

The PRF score for a variant is related to the log odds of a variant with those annotations causally influencing gene expression. As such, PRF scores for a set of eQTL variants can be directly used in a Bayesian framework to determine posterior probabilities of association for each variant (Equation 5).

$$\mathrm{PPA}_i = \frac{\pi_i BF_i}{\sum_{k \in S} \pi_k BF_k} \qquad \text{(Equation 5)}$$

This is identical to Equation 18 in Pickrell et al. (J. Pickrell 2013), except that here the PRF score is used directly to compute π_i, the prior probability of association for each SNP *i*, using Eq. 2 defined earlier. In conjunction with the eQTL summary statistics, this naturally integrates the statistical and annotation information to give posterior probabilities of association. This could also be done directly for a given eQTL study by using fgwas with individual annotations and the summary statistics. In developing PRF scores, we have summarised the complicated process of annotation selection, normalisation, and model optimisation.

We used summary statistics from GTEx subcutaneous adipose tissue to determine the 95% credible set of variants for each of the 2,493 eGenes with lead variant $p < 10^{-12}$. We also used genePRF scores to compute Bayesian priors for variants, and determined PRF score-adjusted 95% credible sets. For the majority of eGenes (67%), the size of the credible set was reduced when using PRF scores (Figure 21). For example, the number of variants in the median eGene credible set was 9 when using statistical information only, and 6 when incorporating PRF scores (Figure 21, top right panel).

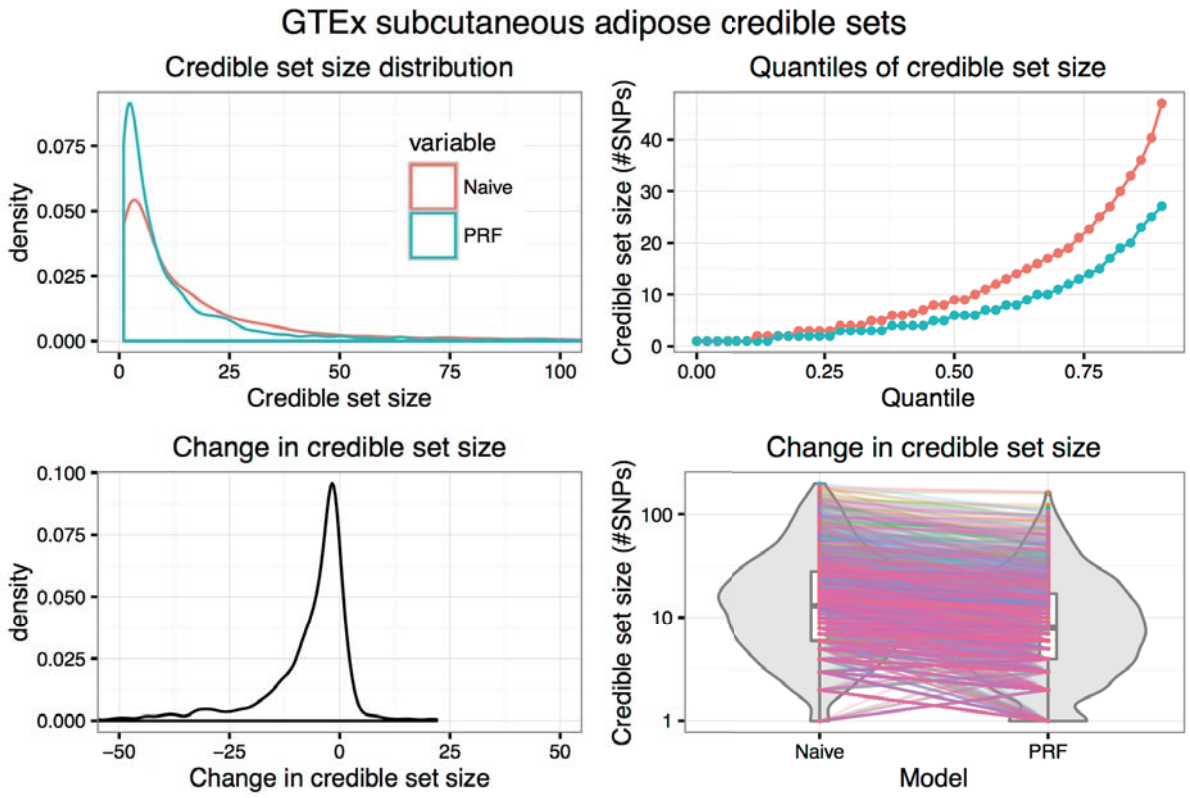


Figure 21: Sizes of 95% credible sets determined using either statistical information only, or combining statistical information with PRF scores. Each of the plots presents a different view showing that credible set sizes are reduced on average when PRF scores are incorporated.

3.4 Discussion

The human genome contains millions of common variants where alleles differ between individuals. An unknown fraction of these influence human phenotypes. While GWAS studies have identified thousands of associations linking variants with phenotypes, the subsequent

step of identifying the causal variants and mechanisms for these associations has proven extremely challenging. Methods that predict variant functionality using functional genomic data can be highly informative for fine-mapping (Spain and Barrett 2015), for burden tests with rare variants (S. Lee et al. 2014), and for identifying phenotype-relevant cell types (Finucane et al. 2015; Trynka et al. 2013).

PRF scores are the first genome-wide scores of regulatory potential based on eQTL data, which include thousands of associations where the regulated gene is known. Previous methods scoring variant functionality have been trained on either simulated data or Mendelian mutations, where the causal variants are known. To use eQTLs, we employed a Bayesian model that accounts for uncertainty in the location of the causal variant. Our PRF score model addresses a number of drawbacks of previous methods. First, whereas functional genomic data have quantitative values, to our knowledge all previous methods for prioritizing variants have exclusively used presence/absence annotations. By developing qfgwas, an extension to the fgwas software, we found that quantitative annotation values substantially improve model likelihoods for eQTL data, indicating that they provide improved performance over binary annotations for localising causal eQTLs. Second, we found that imputed data provided by the Roadmap Epigenomics project had greater predictive performance for eQTLs than the measured data. This finding may benefit others basing their work on Roadmap annotations, since the imputed data are available across all cell type epigenomes, whereas measured data are more sparse. However, since we only examined annotations assayed across more than 30 tissues, this result may not hold for annotations with very sparse sampling. Third, many tools require the user to collect and validate the utility of cell type-specific functional annotations. As a result, a relatively small set of annotations is usually used. With PRF scores, we have used a rigorous framework to integrate a wide range of annotations, producing cell type-specific scores of regulatory function for a large set of human tissues.

PRF scores showed better performance in identifying lead eQTL variants than the widely used methods CADD and GWAVA. However, all methods still showed relatively poor performance in discriminating likely causal eQTL variants from those with weaker statistical associations at the same loci. These evaluations are limited by the fact that we do not have a set of known “true causal” gene regulatory variants, and so we instead used variants where the statistical information alone provided good evidence of causality. Yet, even with this diluted set of true positives, far better prediction performance should be possible. This indicates that we still have a long way to go in deconstructing the grammar of gene regulation.

A number of factors may limit the prediction performance of PRF scores. Some of these are intrinsic to the method. The PRF score for a variant is a sum of log-odds annotation enrichments that were determined globally during model training. However, there may be many cases where a non-linear model would better capture the complexity of gene regulation. For example, TSS distance may not be as informative for a variant in a 3' UTR as for an intergenic variant; DNase hypersensitivity may not be as informative for a splice site variant as for an enhancer variant. As additional genomic data is collected and used for prediction, the need to model non-linear combinations of annotations may grow in importance. In addition, since our model was trained using fgwas, we implicitly assumed that each eQTL was due to a single causal variant. Although in principle this should not bias the estimate of enrichments, it is unknown to what extent modelling multiple causal variants per eGene could improve our enrichment estimates.

Another factor that may limit PRF score performance is the lack of nucleotide-resolution features in our annotation data; most of the annotations used, such as histone modifications and genome segmentations, are limited to 200 nucleotides in resolution. In contrast, there are a growing number of examples of single nucleotide sequence changes that influence transcription factor binding, gene expression, and complex traits. Based on this lack of relevant input features, PRF scores are not allele-specific. This may pose a particular problem for variants that introduce new transcription factor binding sites, thereby altering chromatin accessibility and other epigenomic features. If such a variant is not present in the individuals for whom the reference data was gathered, then no reference epigenomic annotations will overlap the locus, leading to a low PRF score.

The PRF score model was trained on the Geuvadis eQTLs, and so it is possible that the annotation enrichments are to a certain extent overfit on this dataset. The GTEx project now has reasonably large sample sizes for many tissues, and so it would be worthwhile to evaluate how well a model trained based on one tissue extends to the other tissues. It would also be possible to use eQTLs from multiple tissues in model training, which could improve the precision of parameter estimates while also focusing on those annotations that translate well across tissues and datasets.

In principle, many additional features could be included in the PRF score model, which could improve its predictive utility. For example, methods that predict changes to open chromatin, such as Basset and deltaSVM, could used to produce inputs with nucleotide-level resolution for PRF score model training. In addition, although the large volume of existing transcription

factor ChIP-seq data is unevenly distributed across cell types, this is a rich data source that can be integrated with binding motifs to provide more precise predictions of variant effects. By using the centisnp annotation as input we have incorporated some measure of variant effects on TFBS. However, this was available only for 1000 genomes SNPs, and was not applied in a cell type-specific manner. Finally, as genome-wide datasets of chromosome conformation capture (e.g. Hi-C) become available across more cell types, these may help to identify distal regulatory regions. Linking regulatory regions with specific gene promoters could also improve the ability of PRF scores to distinguish the relevant genes for a given variant. Since distance to TSS is the primary annotation linking variants to genes in our PRF score model, we are unable to identify cases where the regulated gene is not the nearest gene.

The coupling of PRF scores with the Roadmap epigenomes is both a strength and a weakness. Computing PRF scores is straightforward across 119 different epigenomes, including a number of cell lines routinely used for molecular assays. However, the dependence on these annotations means that the model is not easily extendable to additional cell types. This precludes the use of PRF scores for specific cell types that are thought to be relevant to some diseases, such as pancreatic islets for type 2 diabetes (Thurner et al. 2017), or regulatory T cells for autoimmune diseases (Carbone et al. 2014).

In summary, our results indicate that a careful treatment of different types of annotations can maximize how informative they are in predicting SNP regulatory potential, and PRF scores integrate these annotations in a novel way across many cell types. As will be described in Chapter 4, PRF scores can be used both for identifying cell types relevant to complex traits, and fine-mapping individual associations. We believe that there remains considerable potential for integrating additional annotations to increase PRF score predictive performance.

3.5 Methods

URLs

Roadmap peaks:
http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidatedImputed
Roadmap signal tracks:
http://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidatedImputed
FANTOM TSS: http://fantom.gsc.riken.jp/5/datafiles/phase1.0/extra/CAGE_peaks/

FANTOM Enhancers: http://fantom.gsc.riken.jp/5/datafiles/phase2.0/extra/Enhancers/
Gencode: https://www.gencodegenes.org/releases/19.html
GERP: http://hgdownload.soe.ucsc.edu/gbdb/hg19/bbi/All_hg19_RS.bw
Centisnp: http://genome.grid.wayne.edu/centisnps/
CADD scores:
http://krishna.gs.washington.edu/download/CADD/v1.0/whole_genome_SNVs.tsv.gz
GWAVA scores: (only available for 1000 genomes SNPs)
ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/VEP_plugin/gwava_scores.bed.gz

EQTL and annotation data for model building

We downloaded genotype data for GEUVADIS samples from the 1000 genomes phase 1 release (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/) and fastq files for RNA-seq on the same samples from ENA (http://www.ebi.ac.uk/ena/, PRJEB3366). For the 343 samples of European descent, we aligned RNA-seq reads to GRCh37 and the Ensembl 69 transcriptome using Bowtie 2 and TopHat, and used custom code to count reads which overlapped the union of exons across transcripts for each gene. We used RASQUAL to associate read counts with all SNPs (imputation INFO>0.7) within 1 Mb of the TSS for each protein-coding gene in Gencode v19. We selected the 6,340 protein-coding genes for which the lead eQTL SNP had $p < 10^{-6}$, and used their association statistics (for 39,566,692 SNPs) as input to fgwas (J. Pickrell 2013) in fine-mapping mode (option -fine).

We annotated each SNP with the distance to the nearest FANTOM TSS expressed at >= 2 transcripts per million (TPM) in LCLs. For binary annotations we determined SNP overlap using bedtools, while for quantitative annotations we used bigWigAverageOverBed to extract the signal value at the SNP. Quantitative annotation values were transformed to normal quantiles based on the distribution of values across all tested SNPs. To split binary annotations into near/medium/far bins of TSS distance (either 0-6420, 6421-33040, or >33040), we created new annotations with the same values as the original annotation, but set to zero outside of the desired distance bin.

The centisnp annotations list the number of motifs which are altered by a given SNP. We used the number of motifs as a quantitative annotation.

Quantitative annotation parameter regularization

A pitfall of using the logistic function for quantitative annotation enrichment is that the model is not always identifiable; that is, different combinations of the logistic's three parameters can give equivalent model likelihoods because they define nearly identical curves over relevant

subsets of the range. A solution to this optimisation problem is to apply a penalty to the logistic coefficients that prevents them from becoming too large. This constrains the search space to what we consider reasonable parameterisations. We use an L2 penalty on the squared parameter value, similar to ridge regression. A penalty of 0.01 on the squared logistic parameter values leads to a cost of ~0.1 units of log-likelihood (LLK) for a parameter value of 3, but of ~1 unit LLK for a parameter value of 10. Experimentation with different penalty values indicated that this level of penalty had a very modest effect on the model LLK after optimisation, as well as on the parameter values for most quantitative annotations, but dramatically improved convergence speed in specific cases.

Fgwas efficiency improvements

We implemented two changes to improve the computational efficiency of fgwas, which are included in the qfgwas version available on Github. Normally fgwas stops the Nelder-Mead optimisation procedure after the optimisation step size has reached a sufficiently small (fixed) value such that further improvement to the model is unlikely, or alternatively after a maximum number of iterations is reached. We noticed that in many cases the step size never became sufficiently small to halt optimisation, and yet the model likelihood did not improve over thousands of iterations. We did not wish to lower the maximum number of iterations, as that might prematurely halt optimisation for models that could still be improved. We thus implemented an additional stopping criterion: when the model LLK is not improved by at least 0.2 units over 400 iterations. Examining many optimisation runs showed that the final model was never significantly changed due to early stopping, yet compute time was considerably reduced for many runs.

To further improve the runtime efficiency of fgwas, we applied code profiling to identify areas for improvement. This highlighted a single function, which sums annotation enrichments for a given SNP at each optimisation iteration, that consumed the majority of the compute time. By precomputing the enrichments once for each iteration rather than for each SNP, we cut the runtime of fgwas for multi-annotation models by approximately 50%. These two improvements are revealed in the run time for models with increasing numbers of parameters (Figure M1).

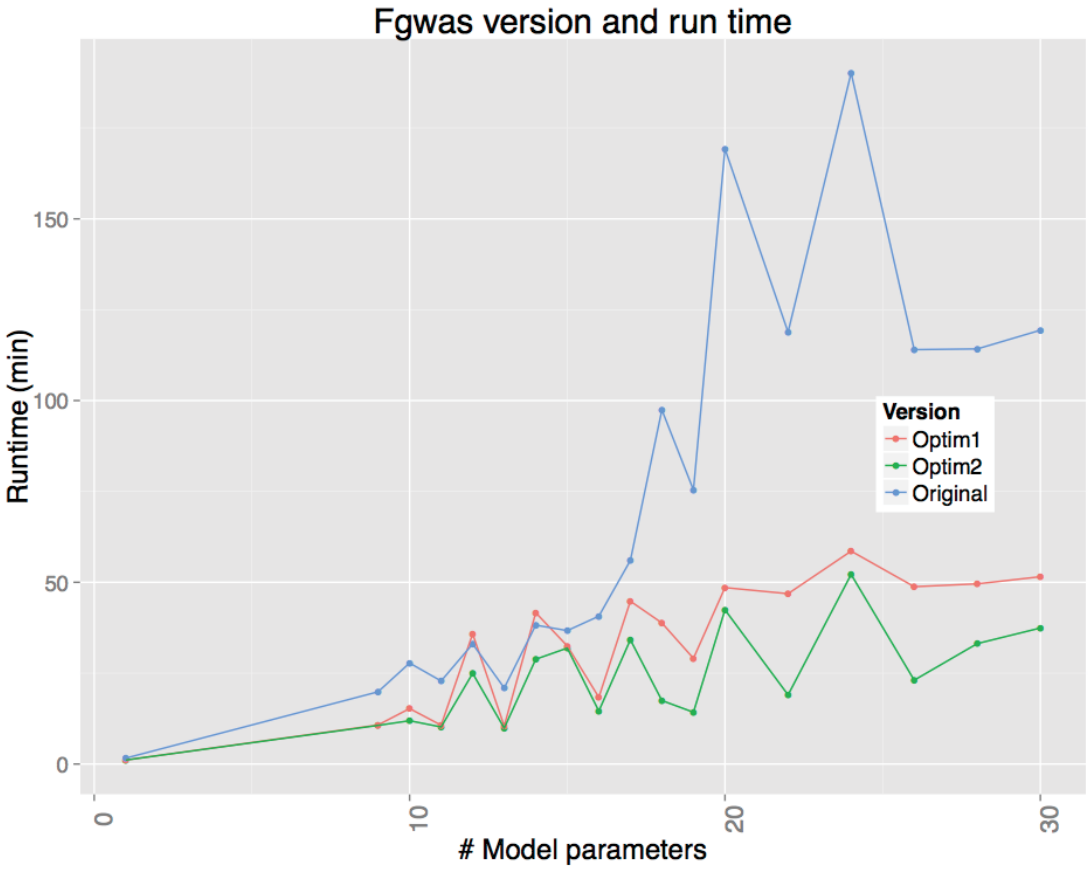


Figure M1: Fgwas running time for models with different numbers of parameters after addition of a new stopping criterion (Optim1) or code optimisations based on profiling as well as the stopping criterion (Optim2).

Model building

To subset SNPs for model training, we first used the Wakefield approximation (Wakefield 2009) to derive approximate Bayes factors from SNP Z scores and MAF, and filtered to retain only SNPs with BF > 10. For model building we used forward stepwise selection to add annotations sequentially to the model, arriving at a 38-annotation model as described in the main text. For cross-validation, we began by tuning the fgwas penalty parameter as described by Pickrell (J. Pickrell 2013); this maximized the cross-validation likelihood with a penalty parameter of 0.01. We then tested 38 models by cross-validation where each single annotation was dropped, but none of these had higher likelihood than the full model.

Computing PRF scores

PRFCalc: software to compute PRF scores

PRF scores are defined for any position in a window of +/- 1 Mb around each protein-coding gene's TSS, in each cell type from the Roadmap Epigenomics project. The number of PRF scores that could potentially be computed is thus:

2×10^6 positions/gene * ~2×10^4 genes * 119 epigenomes = 4.76×10^{12} PRF scores

In addition, for a given position, we would like to be able to provide a breakdown of the PRF score to reveal the individual annotations contributing. With TSS distance plus 38 annotations in the model, the number of values we need to access or compute is:

39 * 4.76×10^{12} = 1.9×10^{14} values

This is approximately 100 terabytes of data, an amount that is not feasible to store and access quickly without considerable infrastructure. We therefore provide software that calculates PRF scores from the required annotation data for each epigenome. This "prfcalc" software is available at https://github.com/Jeremy37/prfcalc. PRFCalc solves the problem of extending the annotation enrichments determined for LCLs to all of the Roadmap epigenomes.

Matching FANTOM and Roadmap tissue types

We use FANTOM TSS and enhancer definitions, yet the FANTOM consortium did not assay the same samples as Roadmap Epigenomics. To be able to compute PRF scores for the Roadmap tissues, we mapped the tissues profiled by FANTOM onto equivalent Roadmap epigenomes. A good match was available for all epigenomes, except for E018 to E022, which are from induced pluripotent stem cell (iPSC) lines and iPSC-derived cell lines. With no matching FANTOM cell types, PRF scores are not available for these epigenomes. When more than one FANTOM tissue was a good match to a Roadmap epigenome, we combined the FANTOM tissues by averaging the transcription levels for a given TSS or enhancer across samples, weighted by the FANTOM sample read depth.

Cell type specificity of PRF scores is determined partly by the set of genes expressed in a cell type, used for TSS distance calculation, and partly by cell type-specific annotations. To focus on genes active in a cell type, we include only those with FANTOM expression of at least 2.0 TPM for TSS distance calculations.

Mapping annotation values to fixed normal quantiles

When training the PRF scores model, annotation values for the set of variants used in training were first transformed with a quantile normal transformation. Because these were broadly distributed across the genome, they reflected the genome-wide distribution of relevant annotation values. However, when computing PRF scores, we may want the score for a single variant. The transformed annotation value for this variant should reflect its quantile among the annotation values used during model training, not its quantile among the variants whose PRF scores are computed at a given time. To do this we created a table, for each annotation, that discretizes the mapping from annotation value to normal quantile into 20,000 bins. When computing the PRF score for a variant, we first retrieve its raw annotation values from Roadmap epigenome files, and transform the values by lookup in these tables.

Determining credible sets

To compute the 95% credible set for an eQTL, we used SNP BFs computed with the Wakefield method. For naive credible sets, we used a flat prior across variants to determine PPAs using Equation 5, i.e. with all π_i set to 1, and assuming a single causal variant among those tested for association. For functionally fine-mapped credible sets, π_i was set to the genePRF or maxPRF score for each variant. We sorted variants by their PPA, and defined the credible set as the minimal number of top variants whose PPA sums to at least 0.95.

3.6 Appendix A - Fgwas equations

We briefly describe the model likelihood computed in fgwas. We assume a standard linear regression between *y*, a vector of quantitative phenotypes (e.g. a gene's expression), and *g*, genotypes for the same individuals. The evidence against the null hypothesis that there is no association between genotype and phenotype can be represented by the Bayes factor; since we use summary statistics here, we compute the approximate Bayes factor as described by Wakefield (Wakefield 2009). The model likelihood is:

$$L(\vec{y}|\theta) = \prod_{k=1}^{K} \sum_{i=1}^{N_k} \pi_{ik} BF_{ik} \qquad \text{(Equation A1)}$$

where the product is over each of K genes, and the sum is over each of N_k variants tested for a given gene, and θ contains all the parameters of the model, i.e. annotation enrichments. The implicit assumption is that each gene has one causal variant influencing its expression. Here, *i* and *k* denote the i^{th} SNP tested for the the k^{th} gene, and the SNP prior probability to be associated was defined in the main text as:

$$\pi_{ik} = \frac{e^{x_i}}{\sum_{j \in S_k} e^{x_j}} \quad \text{(Equation 2)}$$

$$x_i = \sum_{l=1}^{L_2} \lambda_l I_{il} \quad \text{(Equation 3)}$$

where S_k is the set of SNPs tested for gene k, L_2 is the number of annotations in the model, λ_l is the effect of SNP annotation l. For a binary annotation, I_{il} is 1 if the SNP falls in annotation l or 0 otherwise. For a quantitative annotation, I_{il} depends on the annotation value z, and contributes parameters β_0 and β_1 defining a logistic function:

$$I_{il} = \frac{1}{1 + e^{-\beta_1(z-\beta_0)}} \quad \text{(Equation 4)}$$

The likelihood in Equation A1 is maximized by a search across the parameter space using the Nelder-Mead algorithm. When comparing models using cross-validation, we instead maximize a penalized likelihood function:

$$\ln(L^*(\vec{y}|\theta)) = \ln\left(\prod_{k=1}^{K} \sum_{i=1}^{N_k} \pi_{ik} BF_{ik}\right) - p\left(\sum_{l=1}^{L_2} {\lambda_l}^2\right)$$

3.7 Appendix B - model annotation enrichments

Show below are annotations enrichments and parameters from the full 38-annotation model used to determine PRF scores.

TSS Distance	λ
0-285	5.1363
286-1410	4.5813
1411-3690	4.2803
3691-6420	3.9961
6421-10338	4.2492
10339-15106	3.9361
15107-21855	3.6071
21856-33040	3.3870
33041-49943	2.9527
49944-85174	2.5003
85175-164468	1.6573

Binary annotation	λ
intron.diffgene	-0.5568
UTR3.samegene	1.8221
coding.samegene	1.3009
intron.samegene	0.7043
UTR5.samegene	0.8240
State_25.Quies.t2	-0.4929
State_25.Quies.t1	-0.3982
H3K4me3.t3	-0.9406
State_7.Tx3.t2	0.5172
State_7.Tx3.t1	1.3189
H3K4me1.t1	0.6831
H3K27me3.t1	0.2369
UTR5.diffgene	-0.3979
State_12.TxEnhW	-0.4628
State_13.EnhA1	0.5350
H3K36me3.t1	-0.0245
State_6.Tx.t1	1.1575
H3K27me3.t2	-0.2999
State_4.PromD2	-0.6657
State_18.EnhAc	-1.1001
State_1.TssA.t3	-1.6091
miRNA	-0.3995
State_5.Tx5	-0.1243
State_24.ReprPC	-0.6603
H3K36me3.t2	0.4493
State_16.EnhW1.t2	0.4043

Quantitative annotation	λ	β0	β1
DNase	2.3911	2.0624	2.0614
H3K36me3	1.1160	0.0392	2.4245
lcl.Enh.Fantom.tpm	1.7666	0.6441	0.5540
effect-snp.nmotifs	2.4137	2.7872	1.4456
H3K27ac	0.9785	0.4803	3.8216
GerpRS.noncoding	1.2420	2.8310	3.2616
H3K9me3	-1.5800	-0.5448	0.7292
H3K4me3	0.5823	0.6817	4.6988
DNAMethylSBS.fm	-2.7399	3.4404	0.7903
H3K9ac	-2.3670	3.9121	2.4937
H3K4me1	-1.5025	1.6209	0.7241
switch-snp.nmotifs	-0.1794	4.3033	-2.0631

3.8 Appendix C - Classifiers

The PRF score can be treated as a binary classifier, with variants above some threshold score predicted as causal (“positive class”), and those below this score predicted as non-causal (“negative class”). A variety of metrics can be used to assess the performance of binary classifiers, with tradeoffs as to how informative they are in different circumstances. In describing these metrics, it is useful to refer to the confusion matrix, which is a 2x2 contingency table describing the possible combinations of classifier prediction (positive/negative) and ground truth (positive/negative) (Table A1).

		Predicted condition	
		predicted positive	predicted negative
Ground truth	condition positive	true positive (TP)	false negative (FN)
	condition negative	false positive (FP)	true negative (TN)

Table A1: Confusion matrix representing the possible classifier predictions and true conditions in binary classification.

A simple metric is accuracy, which is the fraction of correctly classified cases, (TP + TN) / Total cases. A major problem with using accuracy to evaluate classifiers is that when the classes are unbalanced, then the accuracy can be very high even when the predictions are not useful. For example, if 99% of cases are true negatives, then a classifier would have an accuracy of 99% simply by predicting every case as a negative. However, the sensitivity of this classifier, also known as the true positive rate, TP / (TP + FN), would be zero. This scenario reflects the case with genetic variation, where only a small fraction of variants influence molecular or organismal phenotypes. The accuracy of a classifier is thus dependent on the prevalence of the two classes in the data.

In classification we are concerned with how well both positive and negative cases are identified. A common way to relate these quantities to each other is to plot the true positive rate (TPR) against the false positive rate (FPR = FP / (FP + TN)) as the classifier threshold is varied. This is called the receiver operating characteristic (ROC) curve, examples of which are shown in Figure A1 (left plot). A classifier that makes predictions randomly would

produce a curve (line) along the diagonal and would have an area under the curve (AUC) of 0.5. A good classifier would have a curve bending towards the upper left, with 0.5 < AUC <= 1, indicating a higher true positive rate than false positive rate. Unlike accuracy, the TPR and FPR are theoretically independent of the prevalence of the two classes in the data, as their values depend only on the fraction of negative or positive cases correctly identified.

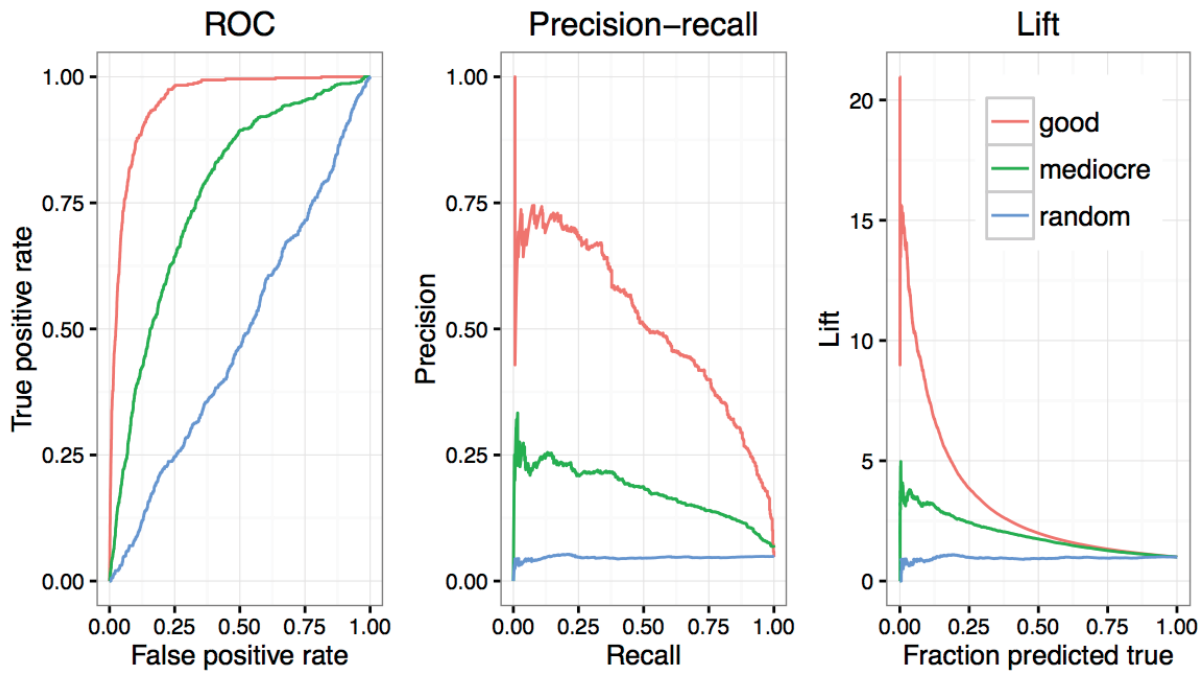


Figure A1: Performance metrics for three classifiers based on simulated data, which are labeled as good (red line), mediocre (green line), or random (blue line). The random classifier has random prediction with respect to the true classification. Shown are (**a**) ROC curves, (**b**) precision-recall curves, and (**c**) lift curves. AUCs for the ROC curves in (a) are 0.95, 0.77, and 0.50 for the good, mediocre, and random classifiers.

While the ROC curve informs on how much better the classifier performs than chance, it fails to reveal how confident we should be in the classification at a given score threshold. The precision, TP / (TP + FP), tells us the fraction of cases predicted as positive which are true positives. A high ROC curve AUC can be achieved even when the precision is low across most of the score range. The TPR, also known as recall, tells us the fraction of all positives that are identified. Precision is often plotted against recall across the range of classifier scores, producing a precision-recall curve (Figure A1, middle panel). A good classifier would produce a curve traveling through the upper right part of the plot, indicating that a large fraction of positive cases can be identified without sacrificing precision.

A final measure of classification performance is lift, which indicates how much better than random the classification performs at different score thresholds. For example, a lift value of 10 at a score threshold where 5% of of the data is predicted true would indicate that among the top 5% of scores there are ten times as many true positives as expected by picking cases randomly. Plotting lift versus the fraction of the dataset above the threshold can reveal over what range of scores the classifier is particularly informative. Lift values always trend towards 1, since large fractions of the dataset can by definition not be highly enriched for positives.

4 Applying PRF scores to GWAS: identifying cell types and fine-mapping

Collaboration note

The work described in this chapter is solely my work, with advisory input from Daniel Gaffney.

4.1 Introduction

Identifying causal variants at GWAS-associated genomic loci is challenging. Purely statistical approaches to fine-mapping have limited resolution when there is extensive LD between variants. Methods that combine statistical and annotation data are available, but require manual selection of a set of potentially relevant functional annotations, and are limited by the power of the GWAS in how well annotation enrichments can be estimated. These barriers have limited the identification of the functional variants, genes, and cell types at GWAS loci.

In Chapter 3, we developed PRF scores, which integrate a diverse set of functional annotations, and can be computed for a large number of cell types with extensive epigenomic data available. Here we apply PRF scores to two outstanding problems in post-GWAS analysis: (i) identification of the relevant cell types of interest, for individual loci and genome-wide, and (ii) fine-mapping to identify plausible candidate causal variants. Because the same annotations underlie PRF scores in both steps, the cell types identified as enriched should be directly useful for subsequent fine-mapping.

To achieve these aims, we solve a number of distinct challenges. First, because epigenomic annotations differ globally across tissues, we faced a problem of normalisation and centering of PRF scores. Second, because the causal variants and genes at each locus are unknown, we faced the related problems of summarizing PRF scores in each locus and cell type, and then aggregating this information genome-wide. We tested a range of alternative solutions to these problems, and then apply PRF scores to identify enriched cell types and to fine-map causal variants for six complex traits. We highlight the results for a number of individual loci, including likely causal variants near *IL2RA* in rheumatoid arthritis and *SMAD3* in inflammatory bowel disease. We also show examples where fine-mapping fails, highlighting areas for improvement. Finally, we summarise the genome-wide outcomes of PRF score fine-mapping on the credible sets of causal variants across loci and on the genes most strongly implicated.

4.2 Identifying cell types for complex traits

If a cell type is relevant for a complex trait, then we expect that causal regulatory variants will be located in active regulatory marks in that cell type, and thus will have higher PRF scores than in irrelevant cell types. In principle, identifying relevant cell types is a matter of testing whether variants that are statistically more likely to be causal have higher PRF scores in that cell type. A key assumption of this approach is that there are no other confounding factors that could influence a given variant's PRF score. However, preliminary analysis suggested that this assumption might be invalid, and that PRF scores showed some global systematic biases across cell types. We therefore sought to first characterise and correct for these differences.

4.2.1 Average PRF scores differ across epigenomes

We noted that mean genome-wide PRF scores differed across epigenomes (Figure 1a), with embryonic stem cell lines and some immune cell types having higher mean scores. In initial testing these cell types dominated estimates of cell type enrichment in many individual traits. We found that a major driver of these differences was that, even though the same enrichments are used for equivalent annotations across epigenomes, these annotations differ in their genomic coverage and average quantitative values. For example, the length of genomic sequence in peaks of DNase hypersensitivity varied from 37 Mb to 107 Mb, while the number of expressed FANTOM TSSes varied from ~13,000 to 23,000. Interestingly, despite strong enrichment parameters, differences in DNase hypersensitivity and number of FANTOM TSSes showed only weak correlations with epigenome mean PRF score (Figure 1b). Instead, we found that a small number of quantitative annotations were highly correlated with mean PRF score, and with each other (Figure 1c).

Because quantitative annotation enrichments are computed for every variant, small global differences in these values between epigenomes contributes to differences in mean PRF scores. These differences are observed across the full distribution of PRF scores (Figure 2a). To make scores comparable across epigenomes, we centered each epigenome's scores by subtracting the difference between the epigenome's mean PRF score and the global mean score across epigenomes. Following this normalisation the distribution of PRF scores still showed minor differences between epigenomes (Figure 2b), but no epigenome dominated cell type rankings across multiple traits. Therefore, all calculations using PRF scores to determine cell type specificity are mean-normalised.

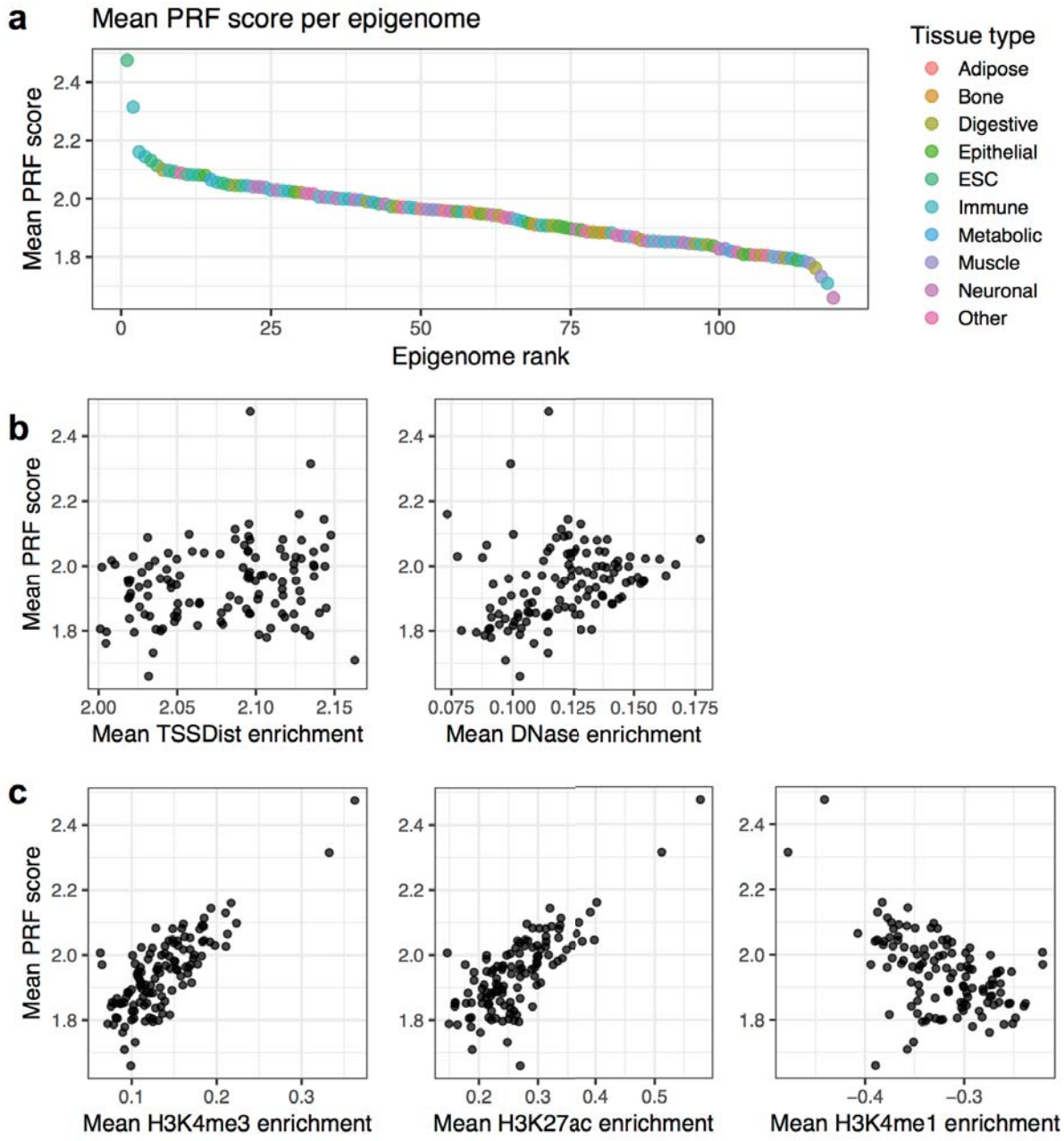


Figure 1: Global differences between epigenomes. (**a**) Mean PRF score for each epigenome across ~288,000 evenly spaced genomic positions. (**b**) Mean TSSDist and DNase hypersensitivity enrichments are only weakly correlated with epigenome mean PRF scores. (**c**) Moderate to strong correlations were found between mean PRF scores and enrichments due to quantitative histone modification annotations H3K4me3 (Pearson r^2=0.62, p<2x10^{-16}), H3K27ac (r^2=0.58, p<2x10^{-16}), H3K4me1 (r^2=0.25, p=3x10^{-9}) and H3K36me3 (not shown; r^2=0.41, p=4x10^{-15}).

An alternative strategy for normalizing quantitative annotations across epigenomes is possible. In the PRF score model described here, quantitative annotations are quantile normalized with respect to the annotation values present in the Geuvadis training dataset, such that each annotation quantile maps to a particular enrichment. It would be possible to

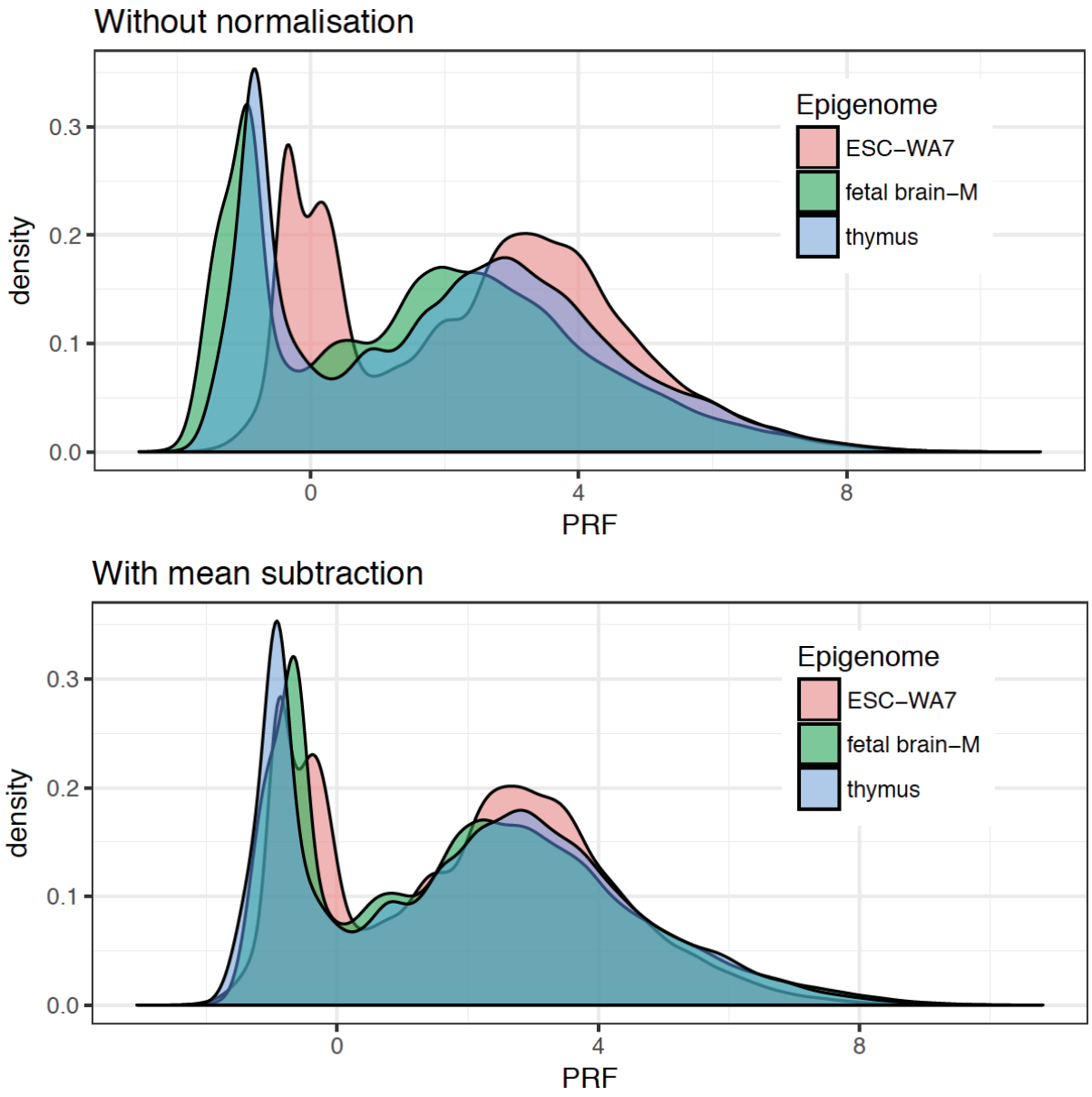


Figure 2: (**a**) Mean PRF scores differ among epigenomes without normalisation. Three representative epigenomes are shown. (**b**) After normalisation to make means equivalent, PRF score distributions are highly similar but not identical across epigenomes.

instead determine quantiles for each annotation separately for each Roadmap epigenome; in this way, the same annotation value might receive a different enrichment in two tissues, but across all variants the distribution of enrichments should be identical between tissues. Although we have not used this normalization method, it could provide a more elegant solution than mean-centring the scores.

4.2.2 PRF score summarisation at a locus

We sought to identify relevant cell types for disease at two levels: first, at the level of an individual locus, and second, genome-wide. When the causal variant at a GWAS locus acts by altering gene regulation, it should have greater overlap with regulatory annotations in highly relevant cell types than in irrelevant cell types, and will therefore have a higher PRF score in relevant cell types. However, in general we do not know the causal variant at a locus.

We explored two different methods to overcome the problem of uncertainty in the location of causal variants, which we call the weightedPRF score and the maxPRF score. For both methods, we began by determining the posterior probability of each variant to be causal using the WTCCC fine-mapping method on GWAS summary statistics at the locus (Wellcome Trust Case Control Consortium et al. 2012). We next defined the credible set of variants that together comprised a >=95% probability of containing the causal variant, with the restriction that only variants with at least a 1% causal probability were included.

The **weightedPRF score** is the sum of PRF scores for variants in the credible set, weighted by their causal probabilities:

$$\text{weightedPRF} = \sum_i PPA_i PRF_i$$

The **maxPRF score** is the maximum PRF score for any variant in the credible set:

$$\text{maxPRF} = \max_i(PRF_i)$$

The weightedPRF score includes a contribution from every variant in the credible set. If GWAS association statistics were noiseless and unbiased, then the weightedPRF score would be optimal, as it assigns more weight to variants more likely to be causal. However, GWAS associations are subject to noise from multiple sources, including sampling noise, and genotyping and imputation inaccuracies. Because many non-causal variants may be included in the credible set, fluctuation in their PRF scores will contribute to noise in the weightedPRF score. The maxPRF score is less subject to this kind of noise since the PRF score from a single variant is used; however, it sacrifices the statistical information which distinguishes among credible set variants.

To explore the utility of these scores in ranking epigenomes at individual loci, we applied them to summary statistics from a GWAS of the autoimmune disease rheumatoid arthritis

(RA) (Okada et al. 2014). Figure 3 depicts the distribution of weightedPRF and maxPRF scores across epigenomes for ten illustrative loci, with three immune cell types highlighted. Three important observations can be made from looking at these scores across loci. First, the epigenomes of cell types already known to be relevant to RA, such as T and B cells, are highly ranked at some loci but not others. Second, average PRF score across epigenomes differ greatly from locus to locus. Third, within many loci, epigenome scores are concentrated in a narrow range. WeightedPRF scores had a smaller spread on average than did maxPRF scores (mean within locus standard deviation of 0.47 vs. 0.70).

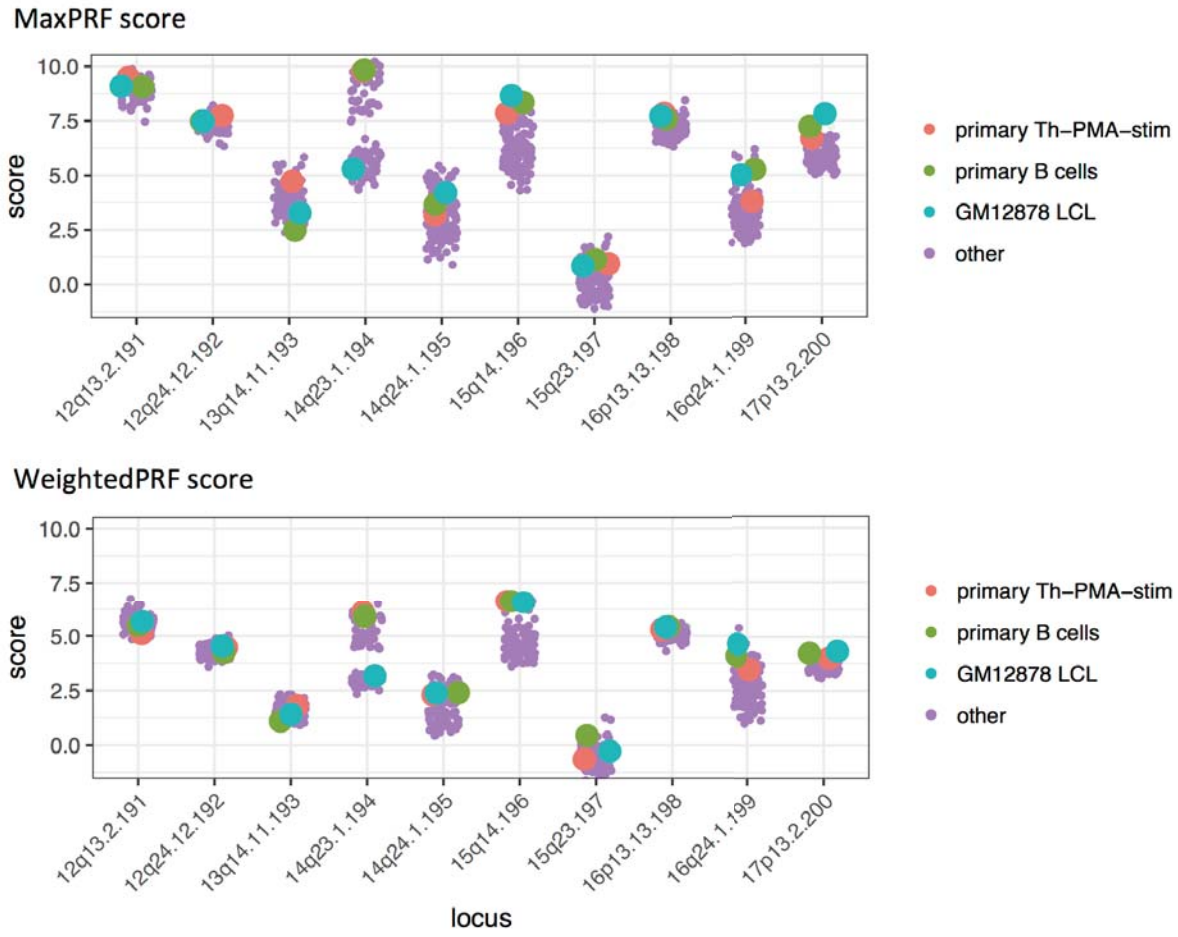


Figure 3: PRF scores across epigenomes for ten loci of an RA GWAS (Okada et al. 2014). Loci are labelled on the x axis. All 119 epigenomes are shown as points for each locus, but three epigenomes are highlighted: PMA-stimulated T-helper cells, primary B cells from peripheral blood, and GM12878 LCLs.

The differences in PRF scores across loci arise because some GWAS associations occur near to genes in regions with strong epigenetic marks, whereas for others this is not the case. The narrow clustering of scores at some loci may reflect the fact that expression of most genes is not highly cell type-specific (GTEx Consortium 2015), and so some level of

regulatory activity is present across many cell types. Two loci illustrate these trends. Locus 15q14 (Figure 3) overlaps an intronic region of the candidate autoimmune disease gene *RASGRP1*, whose expression is limited to certain tissues, and immune cell types occupy the top ranks by both maxPRF and weightedPRF score. In contrast, genes at the 13q14 locus are broadly expressed, and immune cell types are not among the top ranks. The top-ranked cell types at the locus do not have a clear functional grouping, and include umbilical vein endothelial cells, mesenchymal stem cell-derived chondrocytes, and neural progenitor cells.

At many individual RA loci, various immune cell types were among the top ranks by both maxPRF and weightedPRF scores. Because the scores of related cell types are often very similar, it would be imprudent to conclude that the top-ranked cell type at a locus is necessarily the most biologically relevant. Noise in the individual assays underlying a variant's PRF score could change the ranking of cell types at a locus. However, when there are many associated loci, we expect that truly relevant cell types will have higher weightedPRF or maxPRF scores on average.

We sought to use information across all associated loci to more precisely identify disease-relevant cell types. We explored two methods to do this: first, computing the mean maxPRF score across loci, and second, combining epigenome ranks across loci with robust rank aggregation.

4.2.3 Ranking cell types by mean PRF score across loci

A straightforward method to identify relevant cell types is to determine, for each epigenome, the mean maxPRF score across loci. Figure 4 shows this applied to summary statistics from six GWAS, using the WTCCC method at each locus to identify 95% credible set SNPs. For each GWAS, we ordered all epigenomes by the mean of their maxPRF scores across loci. Many of the high-scoring epigenomes reflected well-established enrichments in trait-specific cell types. For example, adult liver was the top epigenome for HDL cholesterol levels, followed by the HepG2 hepatocellular carcinoma cell line, and adipose nuclei. For RA and inflammatory bowel disease (IBD), the top 20 epigenomes were all immune cell types. For educational attainment, the top two cell types were ESC-derived neurons and fetal brain, suggesting a role for early neural development. These enrichments are concordant with those suggested based on candidate gene and expression analyses in the original GWAS analyses (de Lange et al. 2017; J. Z. Liu et al. 2015; Global Lipids Genetics Consortium et al. 2013; Okbay et al. 2016; Okada et al. 2014).

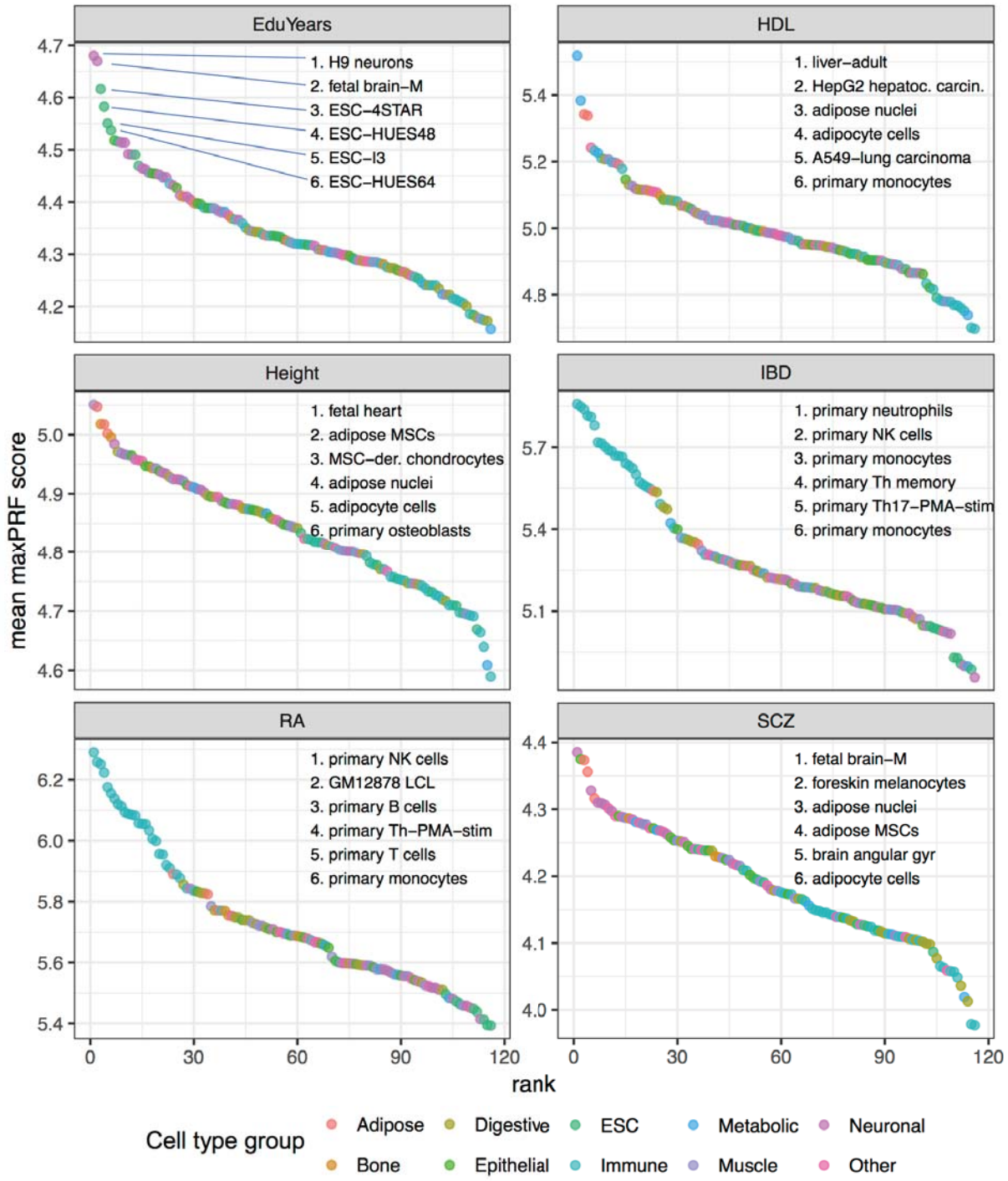


Figure 4: Ranking of epigenomes using the mean normalised maxPRF score across associated loci for six GWAS. The top six epigenomes are labeled for each GWAS, and epigenomes are ordered based on their mean scores. Points are coloured by epigenome tissue type.

Applying the same calculation using the weightedPRF score gave highly similar, but not identical, enrichments to that obtained with maxPRF. Considering that epigenome maxPRF and weightedPRF scores were fairly well correlated across loci (Pearson r^2 = 0.54 - 0.77), in subsequent analyses we used only the maxPRF score as a locus summarisation for simplicity.

Although examining mean PRF scores across loci is simple and clearly identifies relevant cell types among the top ranks, it offers no measure of statistical significance. For example, for schizophrenia (SCZ) the top epigenome is fetal brain, yet subsequent epigenomes ranked by mean score are somewhat surprising - foreskin melanocytes and adipose. Are any of these epigenomes enriched beyond chance levels? We next describe an alternative method of identifying trait-associated epigenomes that assigns statistical significance to the enrichment.

4.2.4 Ranking cell types with robust rank aggregation

As seen in Figure 3 above, both maxPRF and weightedPRF scores differ substantially between loci. Moreover, because biological mechanisms differ between loci, an epigenome that is globally relevant to the trait may not have a high score at every locus. Still at a subset of loci we expect the maxPRF score of a trait-relevant epigenome to be higher than for irrelevant epigenomes. A commonly used way to test for this kind of enrichment is to compare scores for a given set of variants against those of background sets of SNPs. However, there are several drawbacks to this approach. First, the background SNP sets need to be closely matched to to the focal SNPs on properties such as distance to gene, LD, and allele frequency, otherwise enrichment tests tend to show strong enrichment even where none is truly present (Trynka et al. 2015). Second, obtaining such background SNP sets is computationally intensive. Finally, it is sometimes impossible to obtain sufficient matched SNPs.

We use an alternative approach to computing epigenome enrichments. We first rank epigenomes at each locus by their maxPRF scores, and then, using robust rank aggregation (Kolde et al. 2012), test whether some epigenomes have higher ranks across loci than expected by chance. Figure 5 shows the enrichment of epigenomes determined using this method for the six GWAS discussed previously. Notably, P values obtained from applying robust rank aggregation gave an ordering of epigenomes that was highly similar to that obtained based on the mean maxPRF score across loci for each trait.

We also applied the robust rank aggregation method to GWAS of Crohn’s disease (CD) and ulcerative colitis (UC), which are related but distinct autoimmune diseases together described as inflammatory bowel disease (IBD). We found that while CD was associated almost exclusively with immune cell types, UC was strongly enriched for both gastrointestinal and immune cell types (Figure 6), an observation reported recently using stratified LD score regression (Finucane et al. 2015).

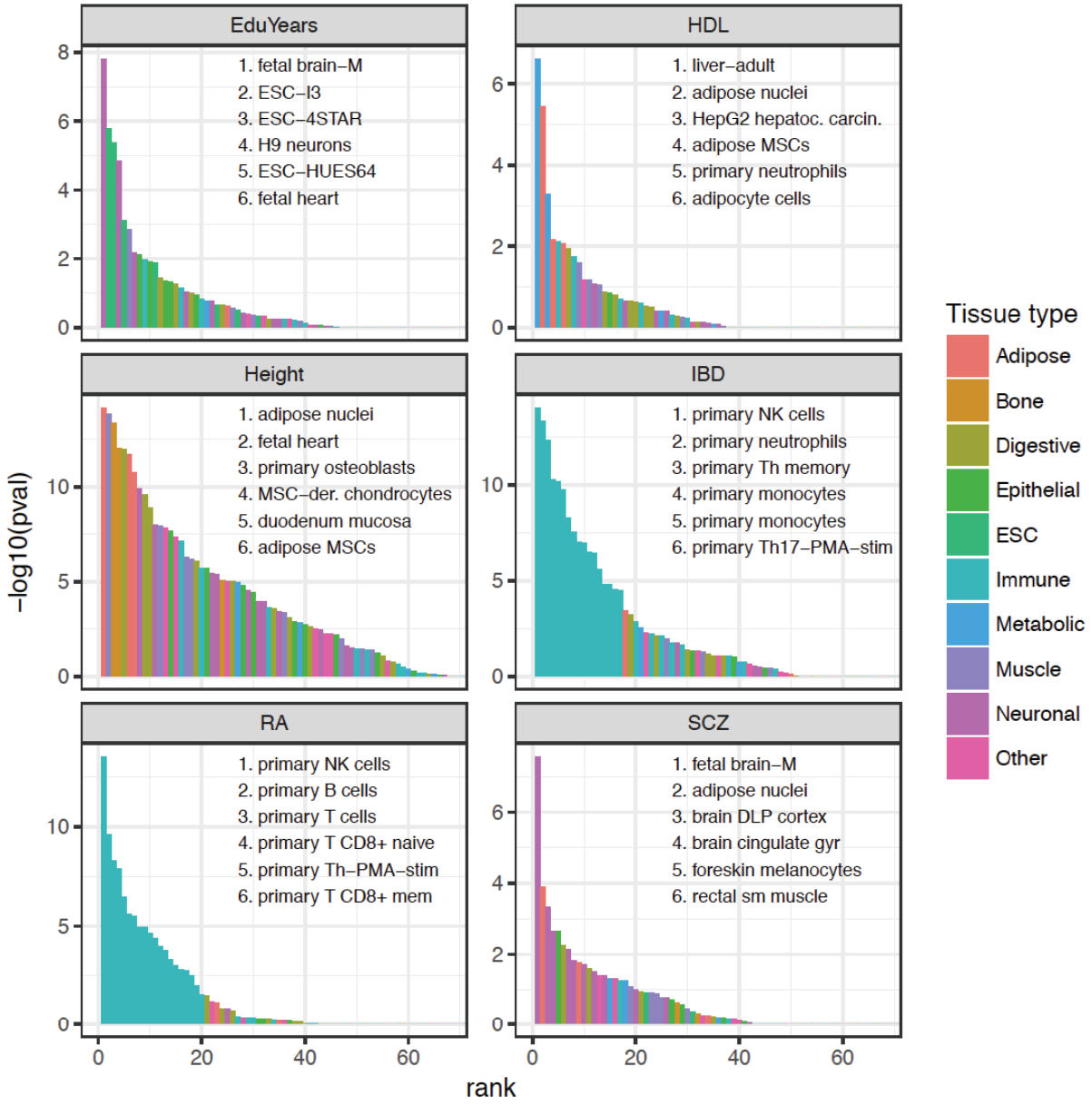


Figure 5: Enriched epigenomes identified by robust rank aggregation with maxPRF score for six GWAS.

A caveat of our method is that we are not independently testing each epigenome for *absolute* enrichment of high scores among trait-associated variants. Rather, with robust rank aggregation we identify epigenomes that have higher scores *relative* to other epigenomes at associated loci. This approach has certain benefits and drawbacks. One benefit is that we avoid the need for background SNP sets, and the potential associated problems such as finding enrichment across all tests. A second benefit is interpretability: while it is possible for *all* cell types to be enriched for high PRF scores at trait-associated variants, due to the

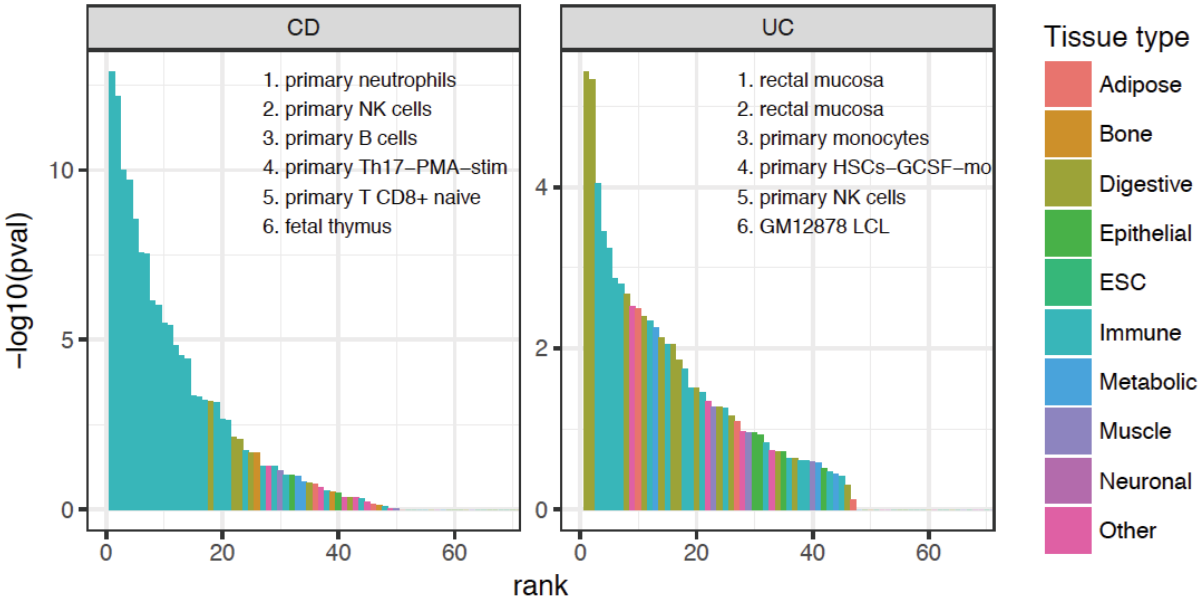


Figure 6: Enriched epigenomes in GWAS of ulcerative colitis and Crohn's disease.

ubiquity of gene regulatory mechanisms, only the most relevant cell types can be enriched *relative* to others. However, a potential drawback is that by using ranks rather than absolute scores, we lose information at highly informative loci, and give equal weight to loci where all epigenomes have similar scores. Furthermore, because enrichment is computed relative to other epigenomes, the significance of each epigenome's enrichment is at least somewhat influenced by the other epigenomes in the comparison.

We compared cell types enrichments determined using PRF scores with stratified LD score regression (LDSC), a widely used alternative method. Stratified LDSC determines cell types associated with GWAS traits by using summary statistics to estimate the fraction of heritability attributable to variants in specific annotations, relative to a background set of annotations. Cell type enrichments determined across five GWAS traits were highly concordant between the two methods (Figure 7, three traits shown). For HDL cholesterol, both methods reported liver as the cell type most enriched for associations. Similarly, for schizophrenia, fetal brain was the top cell type for both methods, although stratified LDSC highlighted primarily brain cell types as the most enriched, and PRF score enrichments were more diverse. Both methods also indicated that immune cell types were highly enriched for RA associations, but the top cell types differed: whereas PRF scores highlighted natural killer (NK) cells, stratified LDSC reported stimulated T-helper cells and T-regulatory cells as most enriched. Notably, the CD56 marker used to isolate NK cells may also include subsets of B and T cells, and so the PRF score result may indicate a general enrichment of multiple immune cell types. Since causal mechanisms and cell types have not yet been elucidated for

most loci associated with these traits, there is no benchmark for evaluating the performance of different methods. Differences between the methods may relate to the fact that stratified LDSC's results are based on heritability, so will be influenced primarily by the strongest GWAS associations, whereas the PRF score method gives all associations equal weight.

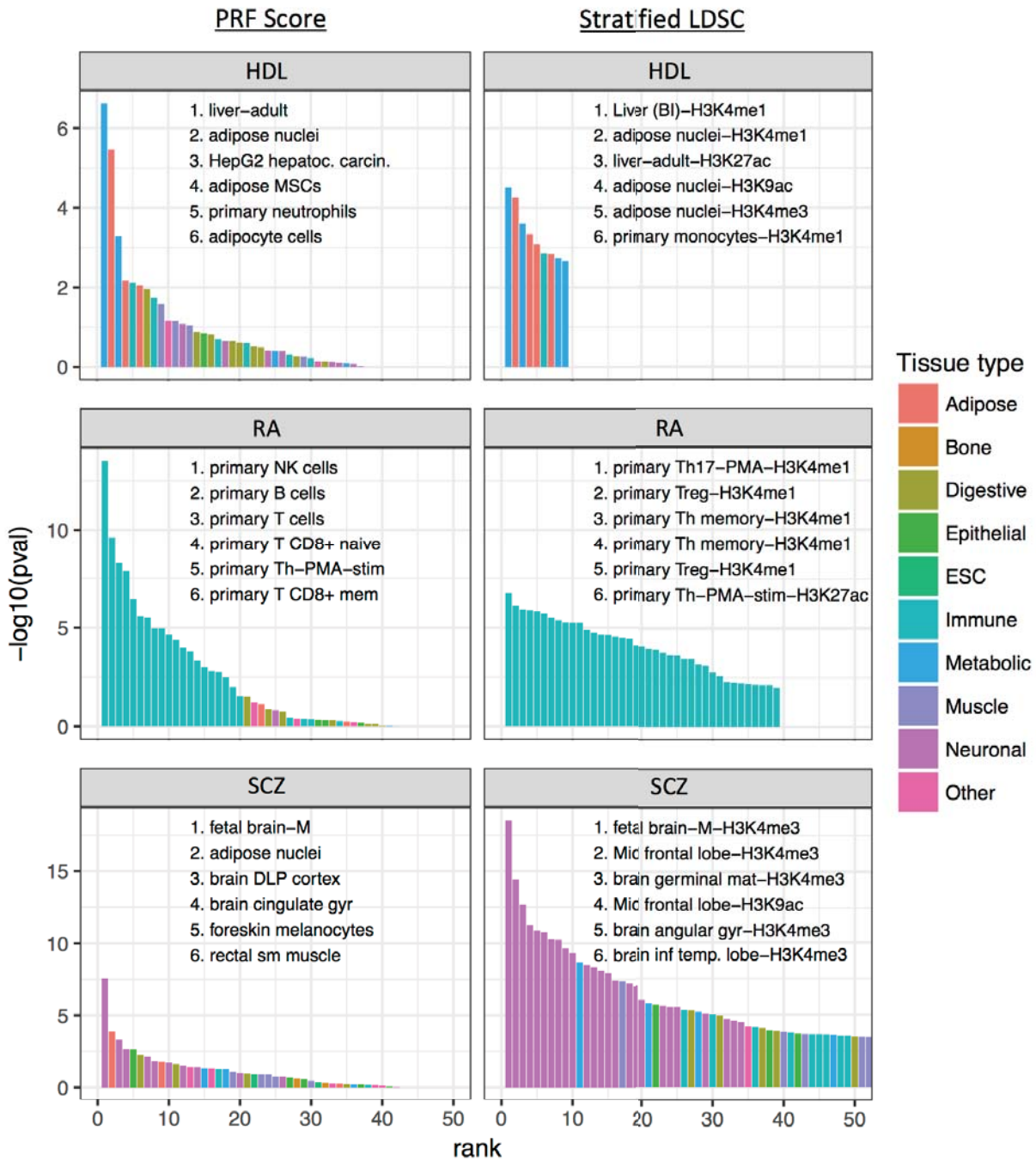


Figure 7: Comparison of the cell type enrichments discovered using PRF scores or stratified LD score regression (LDSC), across three GWAS traits: HDL cholesterol, rheumatoid arthritis, and schizophrenia. The log p value axes differ between traits, but are the same in each case for PRF scores and stratified LDSC.

4.3 Fine-mapping with PRF scores

4.3.1 PRF scores are higher for trait-associated variants

Fine-mapping causal variants is a distinct problem from identifying enriched cell types. Although we have shown that PRF scores at associated loci are higher in trait-relevant epigenomes, it could be that all variants at these loci tend to have more regulatory activity in these epigenomes. For PRF scores to be useful in fine-mapping, they must be higher for causal than for non-causal variants at the same loci. To confirm that this is the case, we applied Bayesian fine-mapping (Wellcome Trust Case Control Consortium et al. 2012) to each associated locus of five GWAS, with the assumption of a single causal variant per locus. For each trait, we selected a single epigenome from among those identified previously as enriched for that trait, and computed PRF scores using these epigenomes (Table 1). We excluded the GWAS for height because a wide variety of cell types are enriched at height loci, and because in a combined analysis the large number of height associations would dominate the results.

GWAS Trait	**Selected epigenome**
Rheumatoid arthritis	E041 - Primary T helper cells PMA-I stimulated
Inflammatory bowel disease	E046 - Primary natural killer cells from peripheral blood
Schizophrenia	E073 - Brain dorsolateral prefrontal cortex
HDL cholesterol	E066 - Adult liver
Educational attainment	E053 - Cortex derived primary cultured neurospheres

Table 1: Selected epigenomes for fine-mapping in five GWAS.

Across the five traits, PRF scores in trait-relevant epigenomes were higher for variants with higher statistical posterior probability of being causal (Figure 7a; $p=2x10^{-21}$, linear regression, PRF scores higher by 2.4 for 1 unit PPA). This suggests that causal variants have higher PRF scores; however, such a pattern could also be observed if the *loci* where high-PPA variants are found differ systematically from other loci, such as by being closer to genes or having broader regulatory regions. If this were the case, then PRF scores for likely causal variants would be no higher than average PRF scores at the same loci. We therefore examined *relative* PRF scores obtained after subtracting the median PRF score among all

variants at each locus. Relative PRF scores were also significantly higher for variants with high PPA than for likely non-causal variants (Figure 7b; $p = 6x10^{-33}$, linear regression), indicating that PRF scores contain information relevant to fine-mapping.

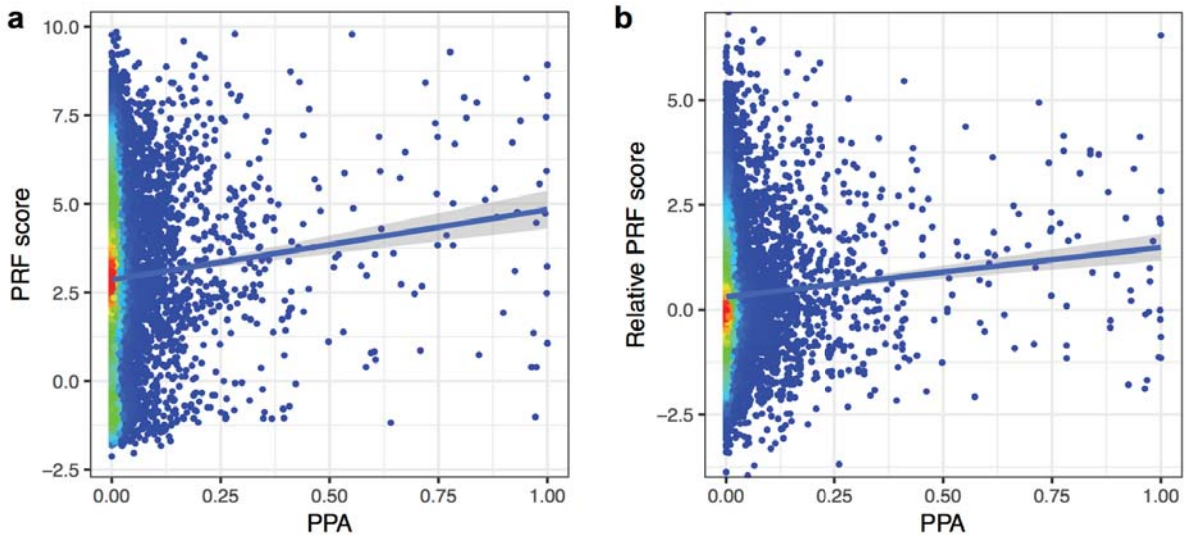


Figure 7: Scatter plots with linear fits of the relationship between PRF scores and the posterior probability of association, across all loci of five GWAS. The shaded region represents the 95% confidence interval for the fit. For plotting purposes, points with PPA < 0.1 were downsampled. (**a**) PRF scores are higher for variants more likely to be causally associated with traits (linear regression, $p = 2x10^{-21}$). (**b**) Relative PRF scores, computed at each locus by subtracting the median score, are also higher for variants more likely to be causal ($p = 6x10^{-33}$).

Although highly significant, the relationship between PRF score and variant PPA is very weak (Pearson r^2=0.00023), implying that the power to discriminate likely causal variants from non-associated variants is poor. For eQTLs, gene-specific PRF scores could clearly discriminate likely causal variants from presumed non-causal variants (see Chapter 3, Figure 17), based largely on the gene TSS distance annotation. For GWAS, we do not know the causal gene at a locus, and so we instead use the maximum PRF score at each variant for nearby genes. Fortunately, to be useful we do not need PRF scores to discriminate causal variants from among all variants, but only to discriminate causal variants from among credible set variants.

The cell type-specificity of PRF scores enables identifying relevant cell types across GWAS loci, but it complicates fine-mapping, since for each locus we must select a relevant epigenome and compute the PRF scores we will use. An important question is whether this cell type specificity enhances fine-mapping performance, and whether any gain is sufficient

to justify the additional complexity. To evaluate this, we compared PRF scores for likely causal variants in relevant epigenomes with those in three epigenomes that seem unlikely to to be relevant for any of the traits considered: placenta, foreskin keratinocytes, and adipose-derived mesenchymal stem cells. Considering only the 72 trait-associated loci with a likely causal variant (PPA > 0.5), and only the top 20 variants by statistical association at each locus, we compared the median relative PRF score for likely causal variants to the median relative PRF score of the remaining variants. In trait-relevant epigenomes this difference was much larger than in trait-irrelevant epigenomes (Figure 8), suggesting that using a more relevant cell type will aid fine-mapping. Importantly, by considering relative PRF scores this comparison addresses whether cell type-specific information is relevant for fine-mapping, and would not be affected if, for some loci, all variants had higher scores in trait-relevant epigenomes. The PRF score difference for likely causal variants in Figure 8 may be an underestimate of the value of using the “relevant” epigenome at a locus for two reasons: first, not all variants with PPA > 0.5 are causal; and second, here we have used a single epigenome across all loci for each trait, but we expect that different epigenomes may be most relevant at different loci.

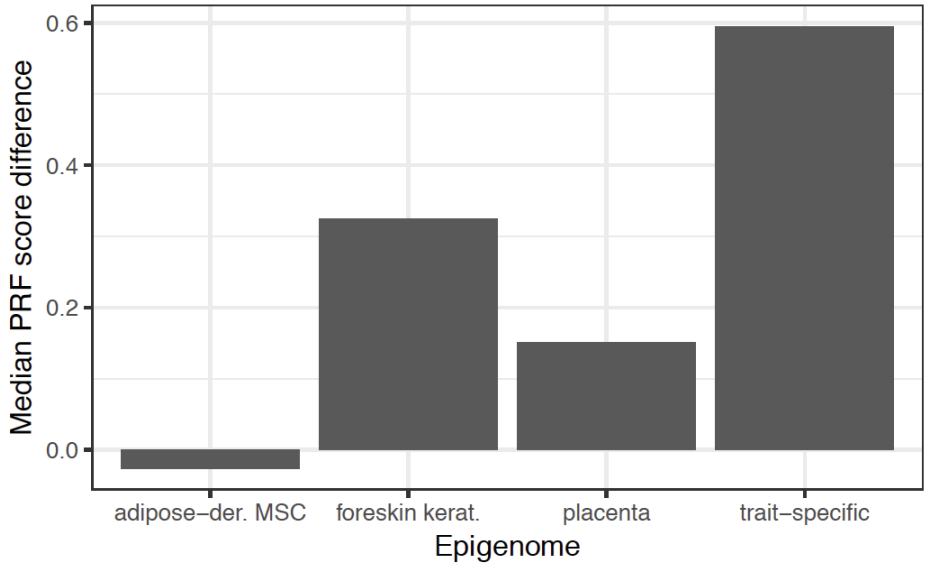


Figure 8: Bar plot of the median difference between PRF scores for variants with statistical PPA > 0.5 and the top 20 variants by statistical association at each locus, computed across five GWAS either for trait-relevant epigenomes (noted in Table 1) or for 3 trait-irrelevant epigenomes.

4.3.2 Fine-mapping individual loci

The aim of integrating PRF scores with GWAS is to identify likely causal variants and genes at individual loci. To enable this application, we developed software and plotting tools that

integrate GWAS summary statistics with PRF scores, and display the results in a transparent manner. To illustrate how these tools can be used, and also to show the limitations of PRF score fine-mapping, we focus on examples of four individual loci.

1. At the *IL2RA* locus association with RA, PRF scores strengthen support for the lead SNP, which is a likely causal variant.
2. At the *SMAD3* locus associated with IBD, PRF scores strongly support the SNP statistically ranked fourth; this likely causal SNP has been experimentally validated as altering *SMAD3* expression and AP-1 binding.
3. At the MEF2C locus associated with educational attainment, PRF scores fail to identify a likely causal SNP due to missing relevant genomic annotations, showing that manual examination of individual loci is important.
4. At the BLK locus associated with RA, the causal SNP is missing from the set of variants considered, and PRF scores highlight an alternative SNP. This describes an underappreciated general problem of fine-mapping analyses which is general to all methods.

Additional PRF score fine-mapping examples are included in the Appendix.

PRF score fine-mapping for GWAS uses the same approach as we described for eQTLs to compute posterior association probabilities incorporating functional annotations (Chapter 3, Equation 5). For GWAS, this requires the assumption that the trait association is driven by a regulatory variant with similar genomic properties to those discovered in steady-state eQTL studies. Further, because we do not know the causal gene at each locus, we use the maxPRF score across nearby genes, rather than the gene-specific PRF score. As before, we assume that the association signal is driven by a single causal variant.

We used PRF scores to fine-map 482 associated loci using summary statistics from five GWAS traits mentioned previously. At each locus we limited the set of variants considered to those with PPA > 0.001. Typically this included many variants not in the 95% credible set, and so it was possible for PRF scores to prioritise variants outside the credible set, thereby increasing the credible set size.

4.3.2.1 IL2RA locus - strengthening support for the lead SNP

We first consider an RA association near *IL2RA* (Figure 9), a gene with associations to multiple autoimmune diseases, and which is the target of the multiple sclerosis therapy Daclizumab (Bielekova et al. 2006).

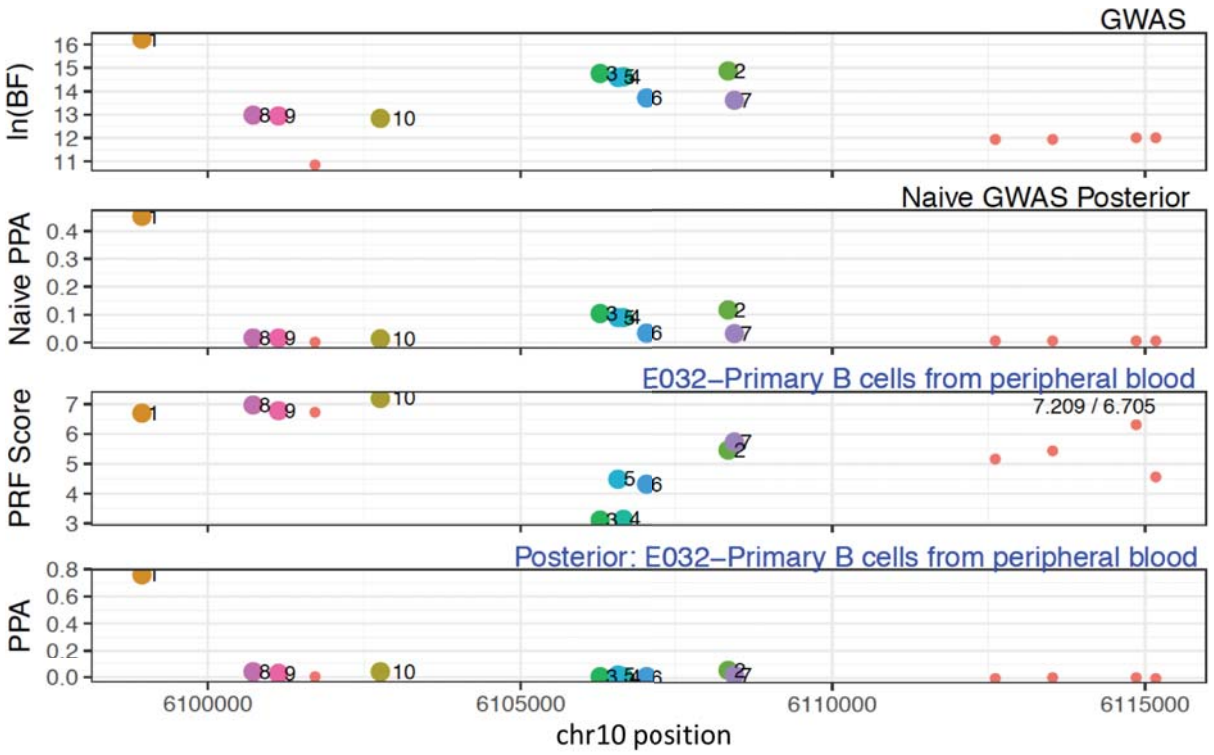


Figure 9: PRF score fine-mapping plot of a rheumatoid arthritis GWAS association at the *IL2RA* locus using the epigenome E032-Primary B cells from peripheral blood. Variants with naive PPA > 0.001 are shown; variants are numbered according to their statistical association, with number 1 being the most associated variant. Top panel: natural log of the approximate Bayes factor for each variant in the credible set. Second panel: naive PPA, determined with the WTCCC Bayesian method assuming a single causal variant. Third panel: PRF scores for credible set variants in the indicated epigenome. Last panel: functional PPA, computed by integrating association Bayes factors and PRF scores from the epigenome in panel three.

The lead variant at the locus, rs706778, achieved a PPA of 0.45 considering statistical information alone, and this was increased to 0.81 when fine-mapping with PRF scores from the PMA-stimulated T-helper cell epigenome (Figure 9, bottom panel). This variant is located in a FANTOM enhancer, and its PRF score was boosted by presence in a DNase hypersensitivity peak, along with histone modifications H3K4me3, H3K27ac, and H3K36me3. In what follows, we refer to the PPA obtained from statistical information alone as the "naive PPA", and to that incorporating functional priors from PRF scores as the "functional PPA". It is noteworthy that the functional PPA of the lead variant was increased even though a number of other credible set SNPs have similarly high PRF scores. This is because the weaker statistical association of these SNPs (e.g. labels 8, 9, 10 in Figure 9) were not boosted sufficiently by their high PRF scores to give them a high functional PPA; in contrast, the lower PRF scores of more strongly associated variants (e.g. labels 2 - 7 in

Figure 9) reduced their functional PPA, thereby boosting confidence in the lead variant as being causal.

A key strength of the PRF score model is that the annotation enrichments contributing to each variant's score are transparent. These enrichments can be visualised in a bar plot of the top variants at each locus (Figure 10). Here we show a breakdown of PRF score enrichments for all variants among either the top six by statistical PPA or the top six by functional PPA.

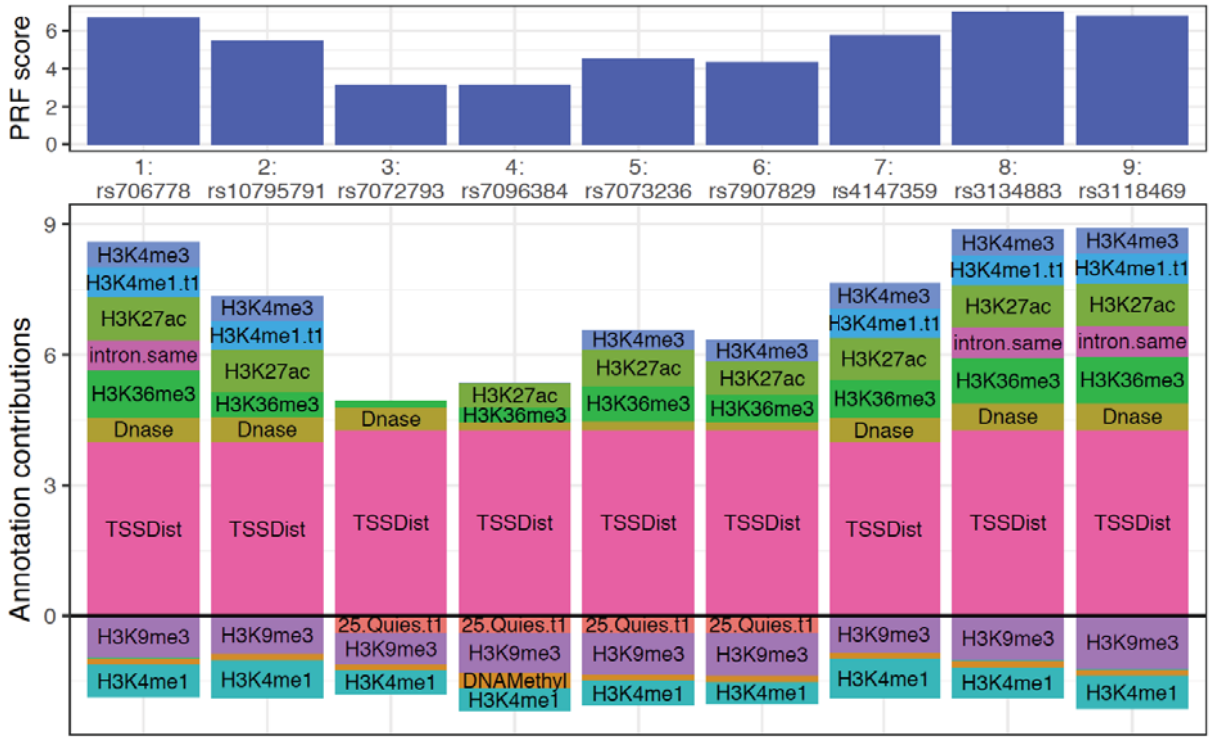


Figure 10: PRF score bar plot showing annotation contributions to the score for each variant in the epigenome E032-Primary B cells from peripheral blood. Annotation labels above zero contribute positively to the PRF score, whereas those below zero subtract from the PRF score. Many of these enrichments reflect the quantitative level of a given ChIP-seq annotation, and do not necessarily indicate a ChIP-seq peak overlapping the variant.

4.3.2.2 SMAD3 locus - a causal variant that is not the lead SNP

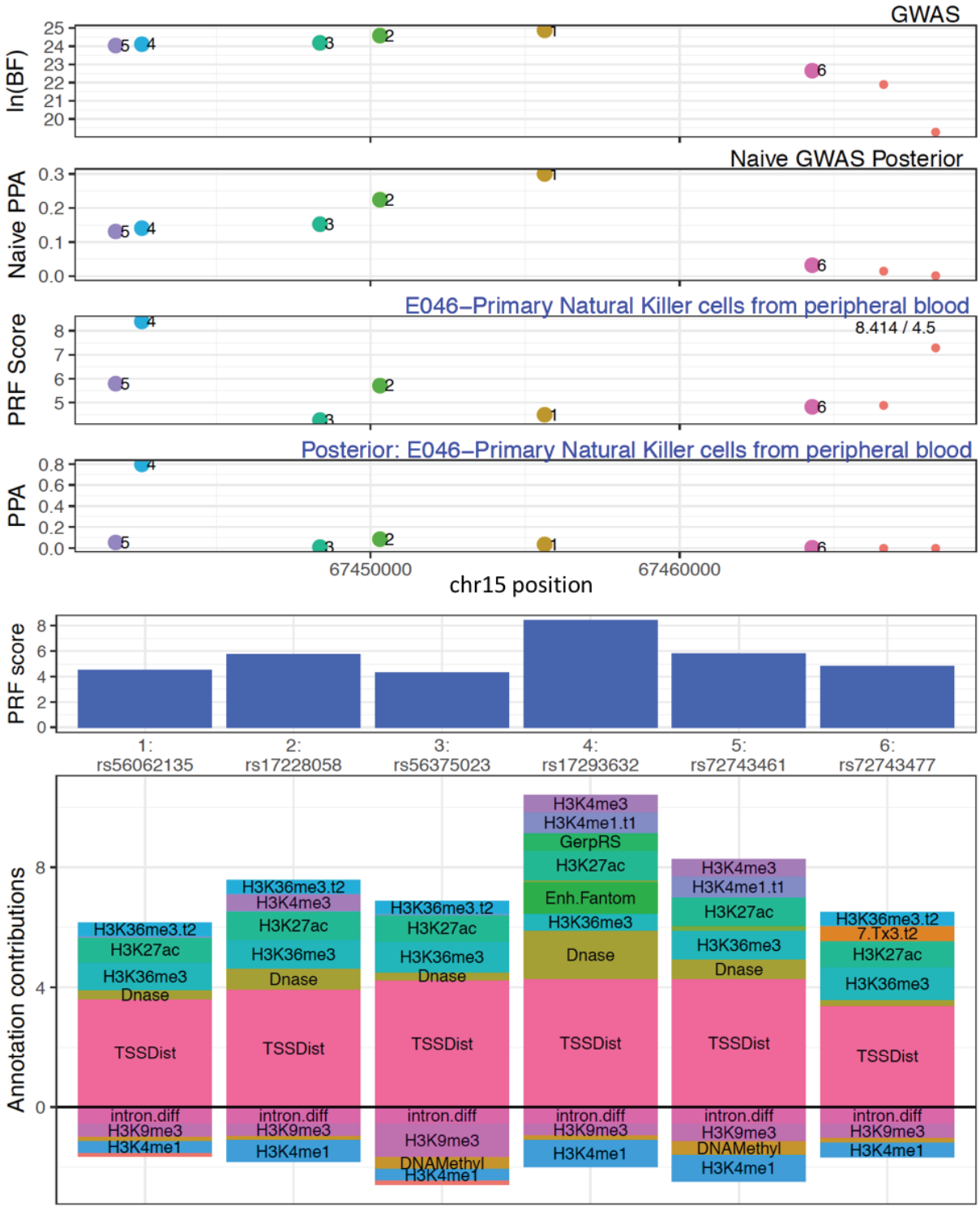


Figure 11: PRF score fine-mapping of an IBD association at the *SMAD3* locus using the epigenome E046-natural killer cells.

A different pattern occurred at the *SMAD3* locus in IBD (Figure 11). Here, high LD resulted in five SNPs having naive PPAs in the range 0.13 - 0.30. *SMAD3*, a strong candidate gene for

IBD, is an intracellular signal transducer and transcriptional modulator activated by TGF-β. Fine-mapping with PRF scores indicated that the fourth variant by statistical association, rs17293632, was a more plausible causal candidate than the other credible set variants, with a PPA of 0.80 using PRF scores from the primary natural killer cells epigenome. This variant is located in a FANTOM enhancer in the first intron of *SMAD3*, and its PRF score is boosted by high nucleotide conservation among mammals, as well as presence in a DNase hypersensitivity peak, along with histone modifications H3K4me3, H3K27ac, and H3K36me3. Other supporting evidence, not considered in the PRF score, is that rs17293632 is located at a nearly invariant position of a JUND binding motif within an AP-1 ChIP-seq peak in K562 cells. Interestingly, although not the top SNP by statistical association in the IBD GWAS we used for fine-mapping (J. Z. Liu et al. 2015), this SNP was reported as the lead SNP in other IBD GWAS, and experimental data have shown that it is an eQTL for *SMAD3* and has an allele-specific effect on AP-1 binding (Turner et al. 2016).

PRF score fine-mapping requires us to use a specific epigenome; to understand the effect of this choice, we also performed fine-mapping at this locus using three other epigenomes: E041-PMA-stimulated Th cells, E030-neutrophils, and E029-monocytes. These gave similar results, although the quantitative scores differed slightly, and hence functional PPAs differed. For example, in all cases rs17293632 was preferred over the lead SNP, but its functional PPA varied from 0.45 in E041-PMA-stimulated Th cells to 0.93 in E030-neutrophils. Indeed, because *SMAD3* is widely expressed, it is difficult to know which cell type is the most appropriate for fine-mapping. This example shows that PRF scores can be effective in prioritising likely causal variants from among a number of statistically-associated variants in strong LD. Also, related epigenomes give generally concordant results, but with some variation in the confidence assigned to different variants.

4.3.2.3 MEF2C locus - failed fine-mapping due to a missing annotation

The *MEF2C* locus associated with educational attainment illustrates how PRF scores can fail to highlight a likely causal variant (Figure 12). Here, the lead SNP obtained a naive PPA of 0.975, while just seven other SNPs have PPA above 0.001. After PRF score fine-mapping, the credible set increased from a single SNP to 24 SNPs. At this locus, however, there are reasons to believe that the lead SNP is causal.

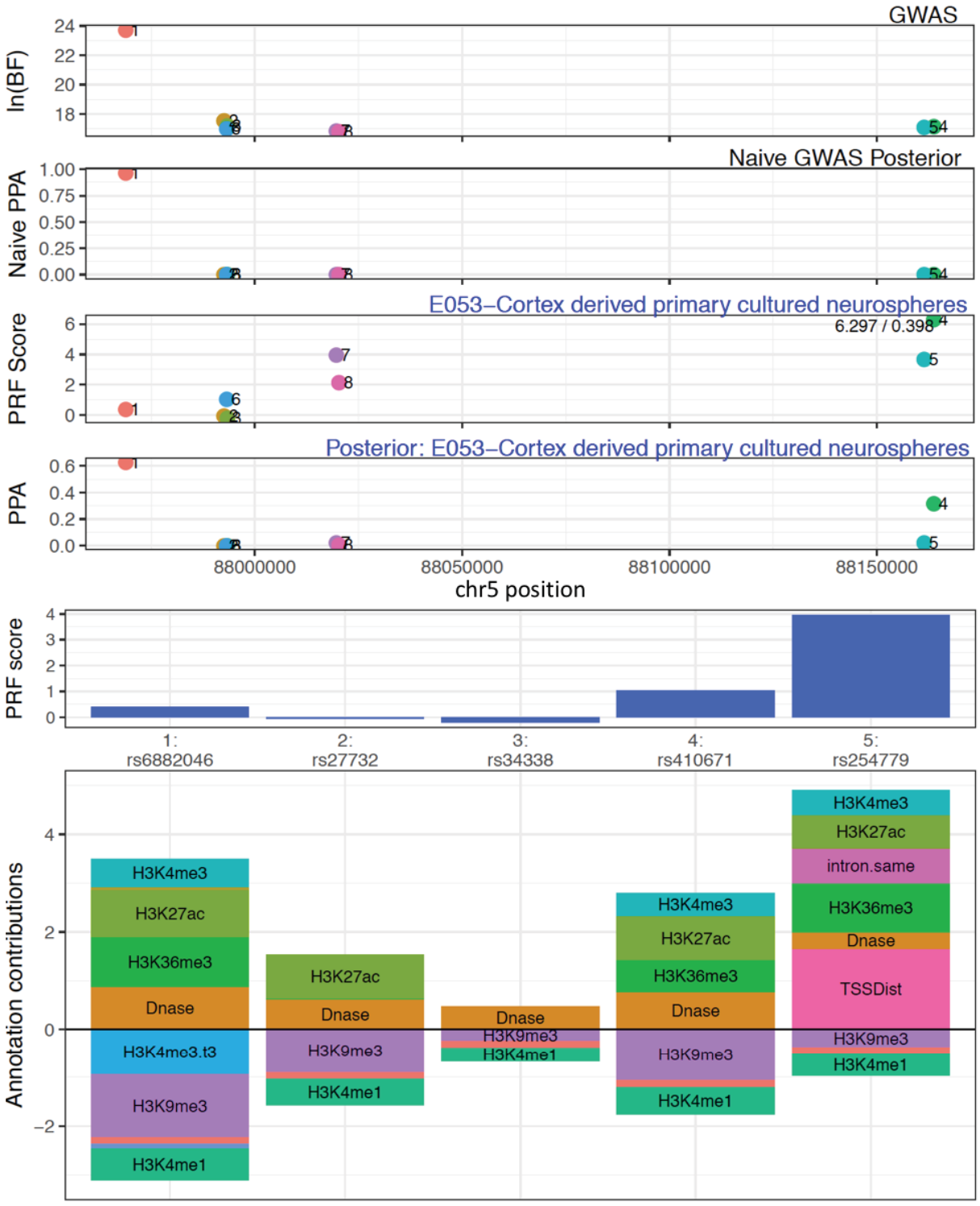


Figure 12: PRF score fine-mapping of an educational attainment association at the *MEF2C* locus using the epigenome E053−Cortex derived primary cultured neurospheres.

The lead SNP, rs6882046, occurs at a moderately conserved position (GERP score 2.53) in the 5' exon of the long noncoding RNA (lncRNA) gene *LINC00461*, within 50 bp of a

FANTOM-annotated TSS. When fine-mapping with epigenome E053−Cortex derived primary cultured neurospheres, this variant received a fairly low PRF score of 0.4. Despite modest enrichments for DNase hypersensitivity and histone marks, it received no annotation enrichment for TSS distance, since lncRNAs were not included in the PRF score model, and the nearest protein-coding TSS is 230 kb away for *MEF2C*. Some low-ranked SNPs at the locus received higher PRF scores due to being nearer to *MEF2C*. As a result, support for the lead SNP was weakened, and its functional PPA reduced to 0.62. One of the highly ranked SNPs, rs61104616, occurs at a highly conserved nucleotide (GERP score 4.44) in the first intron of *MEF2C*, with modest levels of DNase hypersensitivity and histone modifications.

Although the alternative variants cannot be assumed to be non-causal, it is noteworthy that if a significant enrichment for TSS distance were given to the lead SNP, its PPA would have remained very high. Still, it's not clear that lncRNAs should generally be treated as genes, since the fraction of lncRNAs which are functional is unknown. The GENCODE release we used (v19) annotates ~14,000 lncRNA genes, and other sources have estimated their number at more than 50,000 (Iyer et al. 2015). Whereas the majority of protein-coding genes show high sequence conservation and are presumed to be functional, the same is not true of lncRNAs (Palazzo and Lee 2015). In addition, expression-altering variants are strongly enriched towards the TSSes of protein-coding genes, but this has not been established for lncRNAs. For these reasons, we did not include lncRNAs as genes in the PRF score model.

The *MEF2C* locus illustrates that when factors missing from the PRF score model are relevant at a locus, the model can reduce confidence in variants that are likely to be causal. Because TSS distance is a heavily weighted annotation for PRF scores, missing gene annotations can have an especially large effect. This also reflects a general caveat that must be considered with any fine-mapping method using functional genomic data - annotation resources are neither complete nor perfectly accurate. As annotation and model training datasets improve in the future, integrative models such as PRF scores can be improved to more often identify all relevant information at each locus.

4.3.2.4 BLK locus - failed fine-mapping due to a missing variant

The *BLK* locus illustrates what can happen with PRF score fine-mapping when the causal variant is not among those considered. The most recent RA GWAS (Okada et al. 2014) used genotypes imputed to the 1000 genomes phase 1 reference panel. Newer data from phase 3 of the 1000 genomes project include an indel variant, rs558245864, which was absent from

phase 1, and which is in high LD with the reported lead SNP (European $R^2 > 0.95$). Recent work from our group (Kumasaka et al. in prep) showed that rs558245864 is a chromatin accessibility QTL in lymphoblastoid cells, and moreover incorporating Mendelian randomisation to evaluate the causal relationship between variants in ATAC-seq peaks showed that this variant is far more likely to causally influence chromatin accessibility. Finally, CRISPR-Cas9 allelic replacement showed that this variant influences both chromatin accessibility in the region and expression of *BLK*.

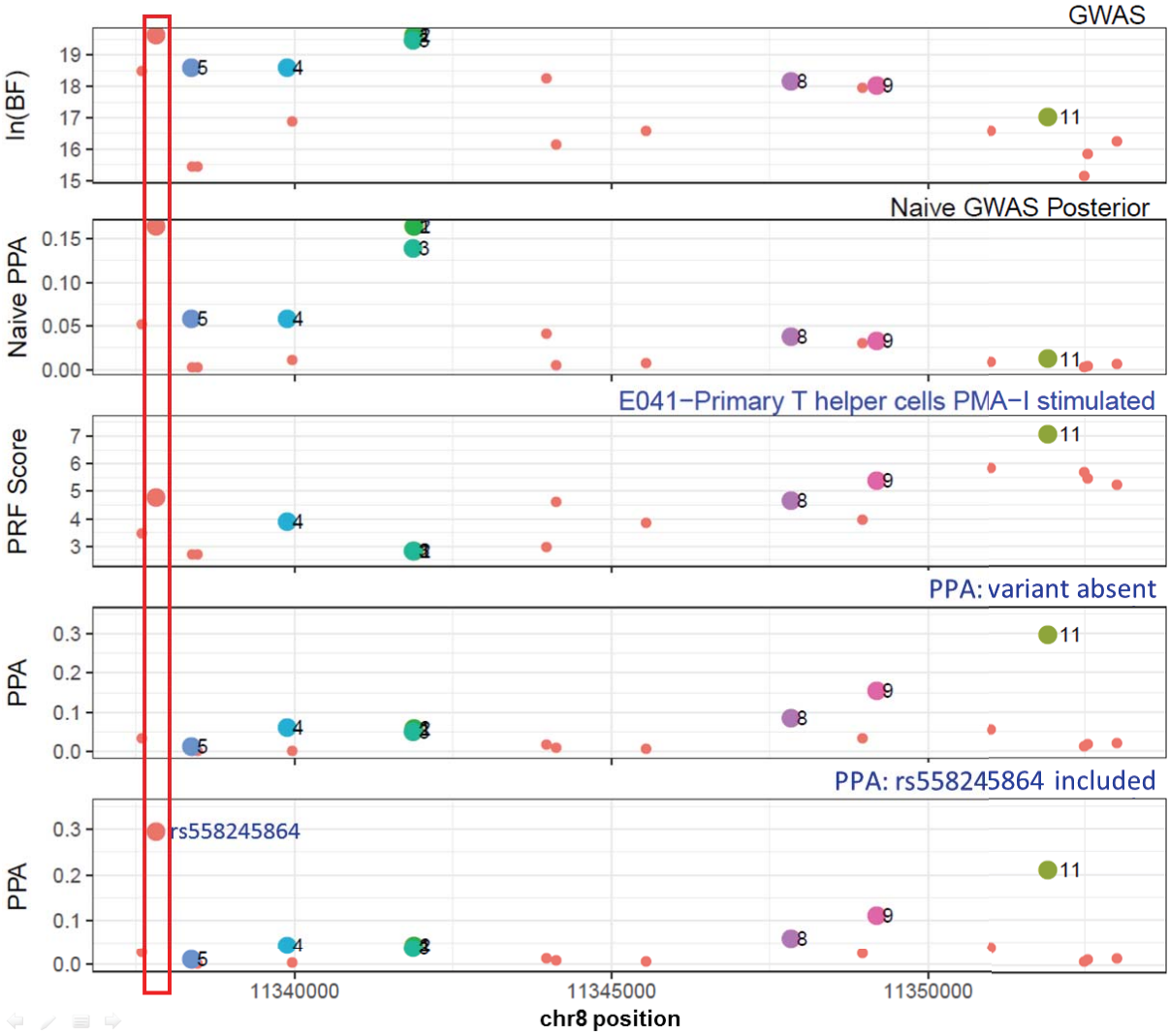


Figure 13: PRF score fine-mapping of a rheumatoid arthritis association at the *BLK* locus using the epigenome E041-PMA-I-stimulated T helper cells. The position of rs558245864 is outlined in red. The bottom panel shows functional PPAs computed when rs558245864 is included, whereas the second-to-bottom panel shows functional PPAs computed when rs558245864 is absent.

Figure 13 shows fine-mapping with PRF scores, either with or without rs558245864. When this variant is included and assumed to have an association Bayes Factor equal to the reported lead SNP, then it receives the highest PPA among associated variants (Table 2).

When this variant is absent, then SNP rs922483 is favoured, due to its altering a conserved nucleotide in the 5' UTR of *BLK*, with additional enrichments from DNase hypersensitivity and histone modifications. Although we do not have specific evidence that rs922483 is non-causal, and indeed there could be more than one causal variant in high LD, a more parsimonious explanation is that rs558245864 is the single causal variant for the association.

Variant ID	GWAS p value	GWAS rank	fPPA	fPPA (rs558245864 incl.)
rs558245864	4.80E-12**	**	**	**0.325**
rs2736337	4.80E-12	1	0.069	0.047
rs2736338	4.80E-12	2	0.069	0.047
rs2736336	5.70E-12	3	0.058	0.039
rs2061831	1.40E-11	4	0.070	0.047
rs2618444	1.40E-11	5	0.017	0.011
rs2409780	1.60E-11	6	0.040	0.027
rs2736340	2.00E-11	7	0.019	0.013
rs1478901	2.20E-11	8	0.098	0.067
rs13277113	2.50E-11	9	0.179	0.121
rs9693589	2.80E-11	10	0.037	0.025
rs922483	6.90E-11	11	**0.340**	**0.230**

Table 2: PPAs of top variants at the BLK locus, either when rs558245864 is included or absent. **rs558245864 is assumed to have a p value equivalent to the lead SNP.

When applying any fine-mapping method, it should therefore be kept in mind that the causal variant may not be in the set considered at all. Although genotyping and imputation of SNPs has improved greatly in accuracy and sensitivity, the sensitivity at which indels and structural variants are detected is still far lower.

4.3.3 Changes to credible sets

Only at a minority of GWAS loci can the association signal be fine-mapped to a single likely causal variant. However, at many loci the size of the credible set can be reduced, as described for a number of fine-mapping methods (Chen et al. 2015; Y. Li and Kellis 2016; Kichaev et al. 2014), and this can inform the selection of variants for more detailed experimental investigation.

To explore the effects of PRF score fine-mapping across traits and across loci, we examined the size of 95% credible sets before and after fine-mapping for the five traits considered

previously, using 317 associations with a minimum p value below $1x10^{-8}$ (Figure 14). The median credible set size was reduced from 18 to 15, with 214 credible sets becoming smaller, 50 staying the same size, and 53 becoming larger. This reduction in average credible set size is similar to that reported by previous methods: PAINTOR achieved an average reduction from 12.3 to 10.4 variants in 90% credible sets from simulated data when using priors based on functional annotations (Kichaev et al. 2014).

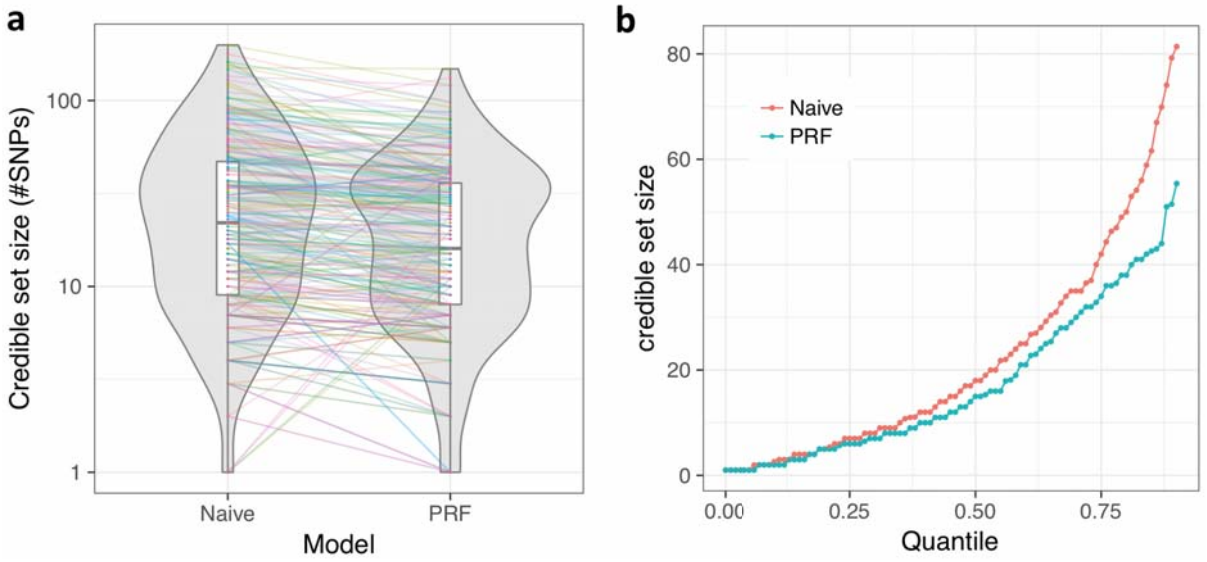


Figure 14: Credible set sizes are reduced when applying PRF score fine-mapping. (**a**) Violin plot of 95% credible set sizes across all loci either using statistical fine-mapping ("naive") or PRF score fine-mapping. Lines show the change in credible set size for each associated locus, with different loci represented by different line colours. (**b**) Quantiles of the distribution of credible set sizes for naive and PRF score fine-mapping. Only quantiles up to the 90th percentile are shown, since credible sets at the tail of the distribution are very large.

We note that it is possible for credible set size to be decreased even when PRF scores support a non-causal variant. Because of LD, many non-causal variants will have some level of statistical association with the trait. If any one of these variants happens to have a high PRF score, its functional PPA may be boosted sufficiently to "crowd out" other variants from the credible set. For example, a case where this may be a particular problem is when a variant at the promoter of a gene happens to be in LD with the causal variant; because in general we do not know the causal gene, such a variant would usually have a high PRF score. To explore the extent to which this can occur, we performed the same analysis as above, but with PRF scores permuted at each locus among all variants with naive PPA > 0.001. Comparing the results of fine-mapping with true versus permuted PRF scores revealed that credible set sizes were reduced to a similar extent with permuted scores (Figure 15).

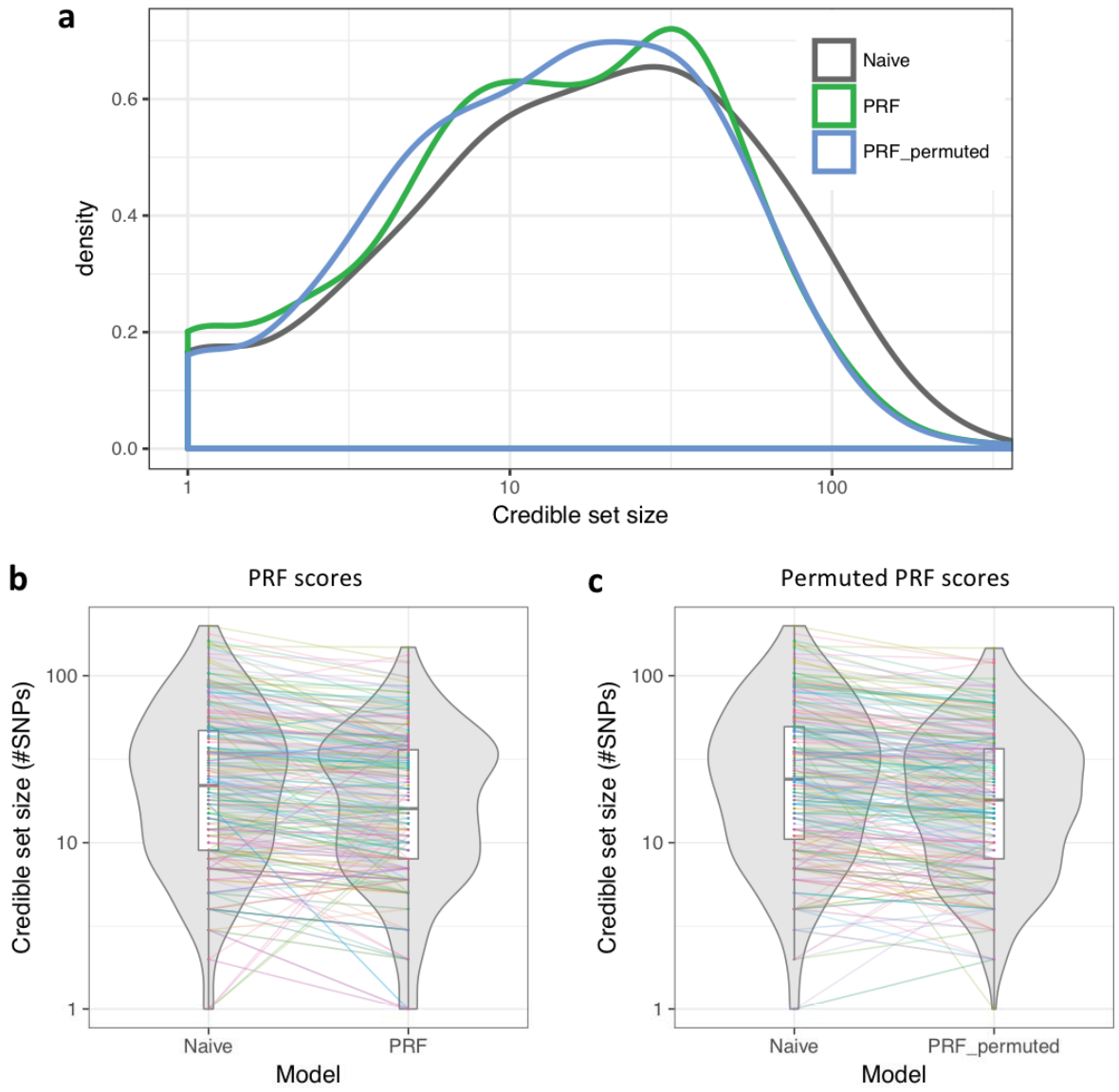


Figure 15: (**a**) Density plot of credible set sizes from fine-mapping 317 trait-associated loci using either true PRF scores or PRF scores permuted among variants with naive PPA > 0.001 at each locus, compared with credible sets from naive statistical fine-mapping. Using both true scores and permuted scores reduced credible set sizes by a similar amount (blue and green distributions are shifted to the left). (**b,c**) Violin plot of credible set size changes when fine-mapping with PRF scores (b), or fine-mapping with permuted PRF scores (c).

The reduction in credible set sizes with permuted PRF scores did not hold when considering associations with a “confident” lead variant having naive PPA > 0.5. Here, whereas true PRF scores slightly reduced credible sets or left them unchanged, permuted PRF scores slightly increased them (Figure 16). To check whether these patterns were dependent on the particular distribution of scores, we repeated the analysis using randomly generated PRF scores drawn from a normal distribution with mean and standard deviation equal to that of PRF scores. The results were concordant with those from the permuted data, with credible

set sizes slightly increased for associations with a confident lead variant, and considerably reduced for the remainder. These observations indicate that reduction in credible set size is not a good indicator of whether a fine-mapping method is accurate, either at a single locus or globally.

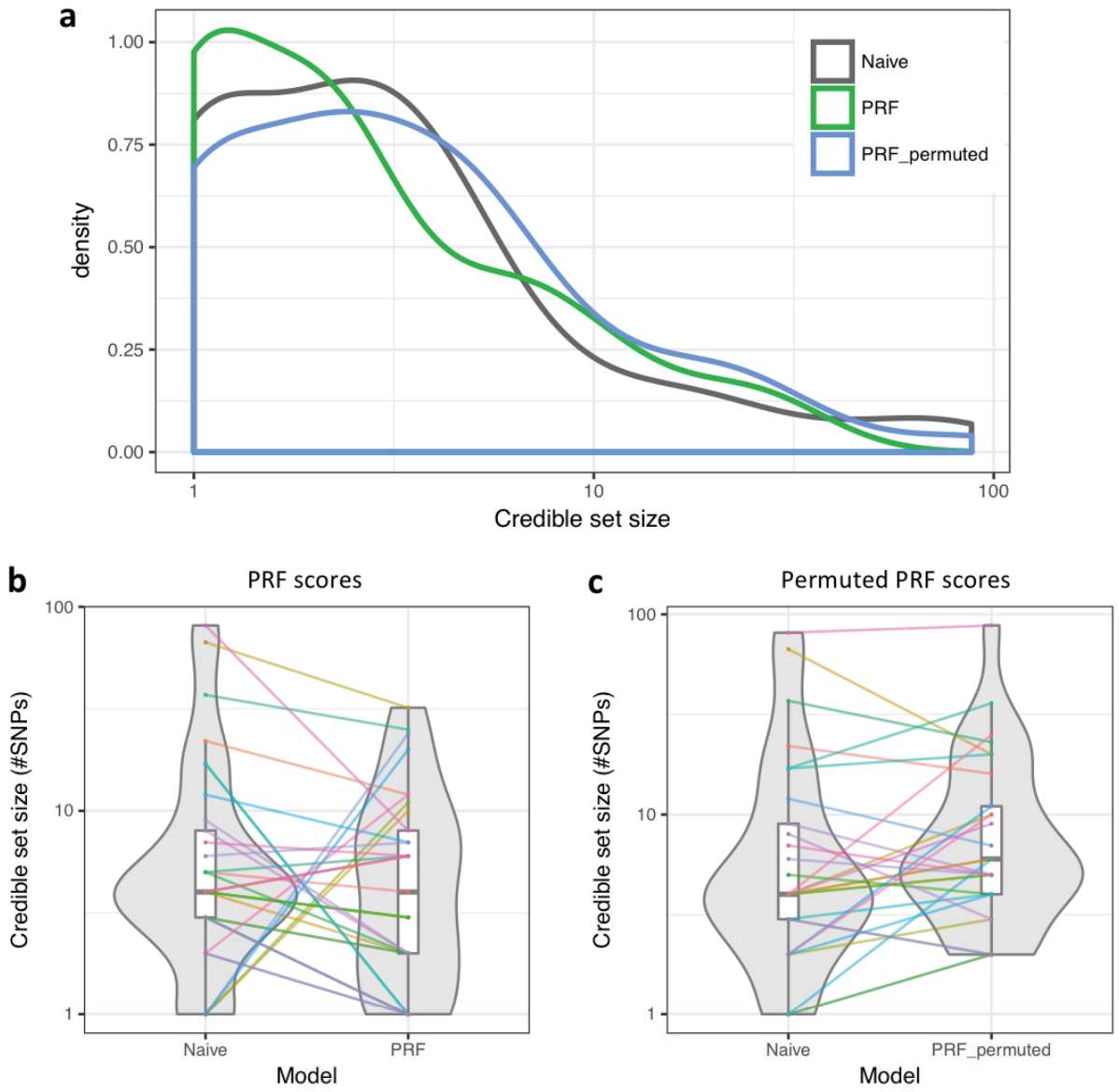


Figure 16: Fine-mapping of 62 loci with a lead variant having naïve PPA > 0.5, using either true or permuted PRF scores. (**a**) Density plot of credible set sizes from fine-mapping. True PRF scores give a slightly higher density of small credible sets, while permuted PRF scores have a slightly higher density at larger credible sets. (**b,c**) Violin plots of credible set size changes when fine-mapping with true PRF scores (b), or with permuted PRF scores (c). Only loci where the credible set size changed are shown. Credible set sizes remained similar or were slightly reduced with true PRF scores, but were slightly increased with permuted PRF scores.

4.3.4 Changes to implicated genes

A key goal of GWAS is to discover genes whose modulation influences risk for disease, and which are therefore potential therapeutic targets. One approach is to assume that the nearest gene to a lead variant causally influences disease risk. A handful of high-profile examples have demonstrated that this is not always true, and that long-distance gene regulatory variants can influence complex traits (Claussnitzer et al. 2015; Guenther et al. 2014; Musunuru et al. 2010). With fine-mapping we hope to discover not only causal variants but also the genes they regulate.

To explore the effect of PRF score fine-mapping on implicated genes, we examined the distribution of lead variants around genes, determined by either naive PPA or functional PPA (Figure 17a). Across 1,002 associations for the six GWAS traits, most naive lead variants were located within a gene body (552, 55%), including its introns; for functional lead variants this was even more often the case (663, 66%). Among lead variants not within genes, functional lead variants tended to be closer to the nearest gene. However, the nearest gene to the lead variant was changed in only 170 cases, despite the fact that for 595 of the associations the lead variant was changed by fine-mapping. One reason for this is that two thirds of the time a lead variant was changed, it was by less than 50 kb (Figure 17b).

PRF scores are implicitly tied to specific genes, which is particularly useful for fine-mapping eQTLs, where the regulated gene is known. When fine-mapping GWAS, the PRF score used is the maximum score for a variant across nearby genes. This most often ends up being the score for the closest gene, because distance to gene TSS is the most heavily weighted annotation, and few other annotations specifically tie a variant to a gene. However, there are two ways in which PRF scores may implicate an alternative gene at a locus. First, the functional lead variant may be located at the promoter of an alternative gene; this gene would be a strong candidate to be causally implicated in the association. Second, in substantial minority of cases, the nearest gene is not expressed in the epigenome used to compute PRF scores, and so the PRF score refers to the nearest expressed gene.

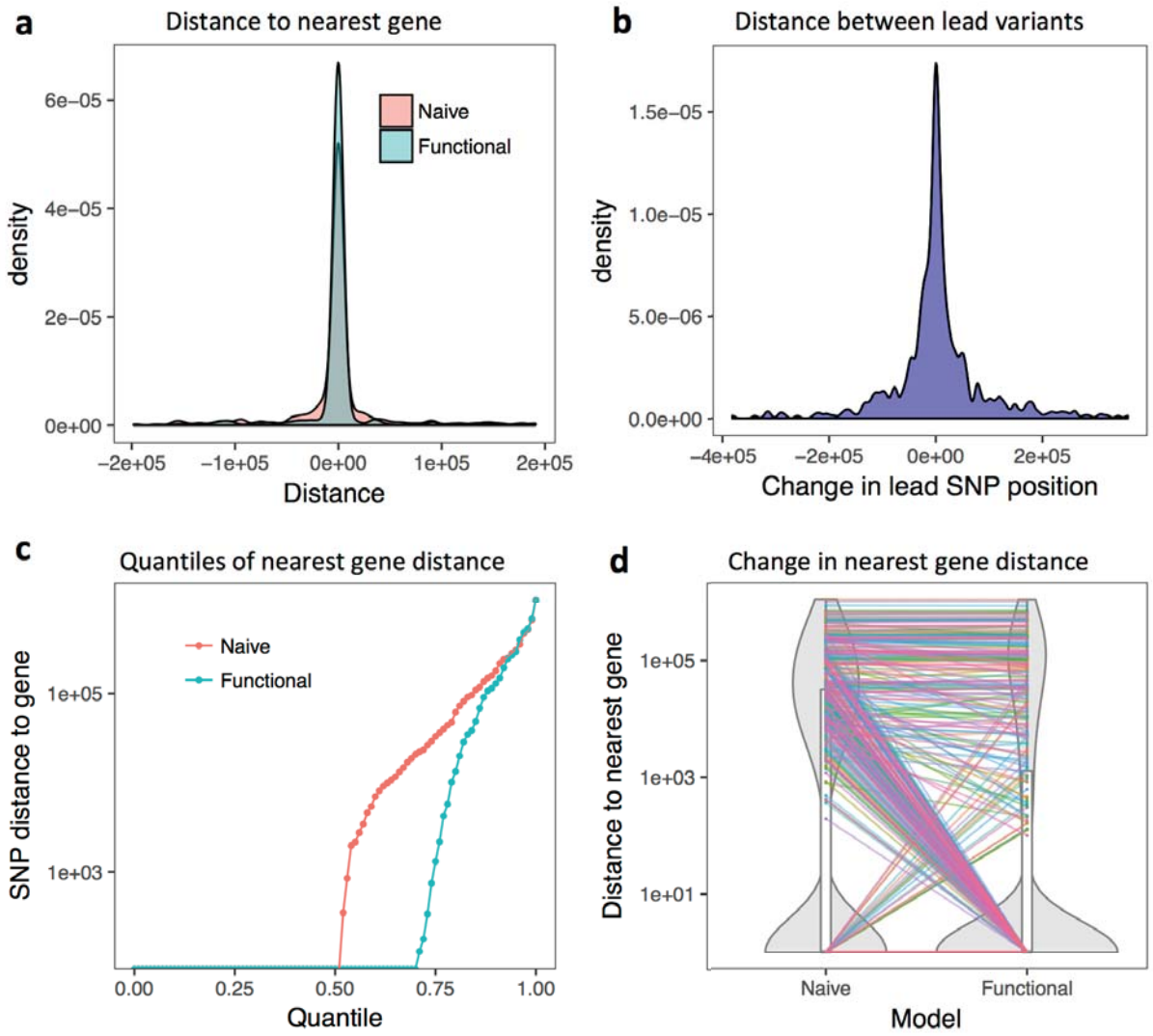


Figure 17: How implicated genes are changed by PRF score fine-mapping for 1,002 GWAS associations across six traits. (**a**) Density plot of distance to the nearest gene for lead variants either pre- or post-fine-mapping. (**b**) Density plot of the distance from the functional lead variant to the naive lead variant, for 595 cases where the lead variant was changed by fine-mapping. 67% of lead variants were within 50 kb of each other. (**c,d**) Distance to nearest gene across quantiles of the distribution (c) or as a violin plot (d) showing the change in distance for the 595 associations with changed lead variants. Functional lead variants were within genes more often than naive lead variants, and the remainder were also closer to the nearest gene.

An example of the former case occurs at an IBD-associated locus on chromosome 1. Here, the naive lead SNP (rs7523335) is nearly 100 kb from *ERFFI1*, but the functional lead SNP (rs17523802) is at the promoter of *PARK7*, in a region of dense transcription factor binding. Due to extensive LD in the region, rs17523802 was 18th-ranked by association statistic, yet its naive PPA (0.02) was only slightly lower than that of the lead SNP (0.035); after fine-mapping, the functional PPA of rs17523802 was boosted to 0.59 (Figure 18). Additional evidence supporting rs17523802 as a candidate causal variant is that it is a stronger eQTL

for *PARK7* in GTEx whole blood than rs7523335 is for *PARK7* in any tissue. *PARK7* is a multifunctional protein that translocates between the mitochondrion, cytoplasm and nucleus in response to oxidative stress. While mutations in *PARK7* have long been associated with familial Parkinson's disease, recent studies suggest that it has roles in inflammation and T cell migration (W. Liu et al. 2015; Ashley et al. 2016; Jung et al. 2014), functions with relevance to IBD.

In 144 cases the nearest gene to the lead functional variant differed from the gene associated with the variant's PRF score. In nearly all cases this occurred because the nearest gene was not expressed in the epigenome used to compute PRF scores. Although in some cases the lack of a gene's expression may be informative, such situations need to be interpreted by looking at fine-mapping results using more than one epigenome. It is possible that the causal gene at a locus is simply not expressed in some cell types selected for PRF score fine-mapping.

The gene-specific nature of PRF scores might be a larger benefit if more annotations that inform on long-range genomic interactions were included in the model. For example, a distal gene could potentially be prioritized if Hi-C data (or promoter-capture Hi-C) indicated that GWAS-associated SNPs located in an enhancer had high contact frequency with the distal gene's promoter. These data were not included in the PRF score model for two reasons. First, an initial attempt at including data from the highest-resolution Hi-C experiment to date (Rao et al. 2014) did not improve PRF score predictions for the Geuvadis eQTLs, despite being a good match for the cell type used in Geuvadis. Potential reasons for this are that Hi-C data may require special handling to extract relevant signal, or that the data simply were not yet high enough resolution. Second, Hi-C data are not yet broadly available across cell types. However, this may change in the future as consortia such as BLUEPRINT produce high-quality promoter-capture Hi-C data across multiple cell types (Javierre et al. 2016).

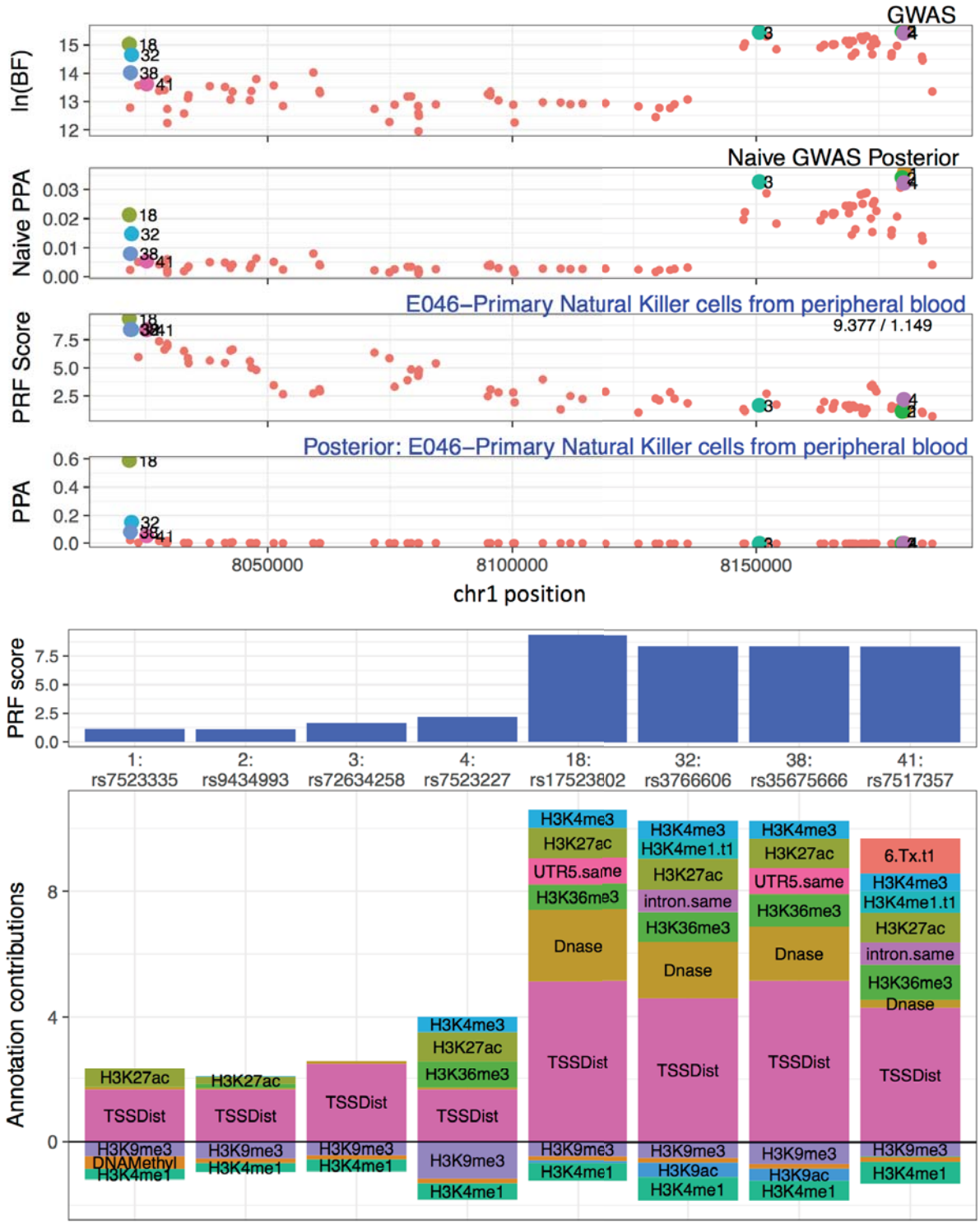


Figure 18: Fine-mapping an IBD association at the *PARK7-ERFFI1* locus localises a candidate causal variant at the promoter of *PARK7*.

4.4 Discussion

We have described a novel method applicable to two challenges in post-GWAS analysis: identifying associated cell types and fine-mapping causal variants at non-coding GWAS loci. Previous methods that identify associated cell types have primarily focused on individual annotations, such as H3K4me3 (EpiGWAS) (Trynka et al. 2013), H3K27ac (PICS) (Farh et al. 2015), or DNase hypersensitivity (FORGE) (Dunham et al. 2015). Other methods, such as fgwas (J. Pickrell 2013) (upon which PRF scores are based), GARFIELD (Iotchkova et al. 2016), and stratified LD score regression (Finucane et al. 2015) allow any individual annotation to be used, but do not integrate these together. By using PRF scores, we automatically integrate a large set of cell type-specific annotations for each of 119 epigenomes.

There are advantages and disadvantages to integrating multiple annotations to determine cell type enrichments for a GWAS trait. An expected benefit is that cell type associations should be more accurate, because causal variants that appear in different cell type-specific annotations are incorporated into a single enrichment test. Furthermore, the integration of multiple annotations should reduce noise in comparison with enrichment from a single assay. Across five GWAS traits, we observed cell type enrichments consistent with prior knowledge of relevant cell types. However, a benefit of methods that test annotations individually is that the enrichment seen using one annotation can be “validated” if other annotations from the same cell type give a similar result. Because PRF scores are based exclusively on Roadmap Epigenomic annotations, the only comparable validation is by observing whether related cell types are also enriched. In addition, it is not straightforward to extend PRF scores beyond the Roadmap epigenomes, because the same set of annotations would be needed in any additional cell types.

One advantage of using PRF scores to detect cell type enrichments is that no LD reference panel is needed. Methods that require knowledge of LD, such as EpiGWAS, PICS, GARFIELD, and stratified LD score regression, depend upon there being a good match between the reference panel and the study population. However, unlike these methods, PRF score cell type enrichment depends on full summary statistics from well-imputed data, since we assume that the causal variant is present among those associated at the locus.

Our method for determining cell type enrichments integrates naturally with subsequent fine-mapping. Enrichment of an epigenome’s PRF scores directly indicates that across multiple loci, there is a correspondence between high PRF scores and credible set variants.

Epigenomes can also be ranked at individual loci, and it is possible to select different epigenomes for fine-mapping different loci. A limitation of our method is that because our cell type enrichments are based on associated loci only, we may be missing useful genome-wide signal from sub-threshold associations. A potentially interesting alternative that we have not yet explored would be to use the genome-wide method stratified LD score regression with each epigenome's PRF scores as an annotation input.

Our application of PRF scores to fine-mapping is distinct from prior methods that incorporate functional annotations. First, most such methods estimate annotation enrichments directly from GWAS data (Y. Li and Kellis 2016; J. Pickrell 2013; Kichaev et al. 2014). Because the accuracy of enrichment estimates can be limited by the number of GWAS associations, one of our motivations for developing PRF scores was to leverage the much larger number of eQTL associations. PRF scores therefore include accurate annotation enrichments, and are suitable for loci where altered gene regulation is a suspected mechanisms for the association. Since most GWAS associations do not appear to be explained by coding associations, this includes most GWAS loci.

PRF score fine-mapping rests on an assumption that variants influencing complex traits have similar genomic properties to those influencing steady-state gene expression. This assumption is supported by the observation that both GWAS and eQTL associations are enriched in open chromatin, in enhancer regions, and near genes. However, it is possible that these enrichments differ quantitatively. Because eQTLs are tied to specific genes, we have a good estimate of the distribution of causal gene regulatory variants around genes; for complex traits, this is not the case. For example, eQTLs are highly enriched at gene promoters, and it may be that complex trait variants occur more often in elements distal to the regulated genes. It's also possible that the gene-regulatory effects of some complex trait variants occur only in response to specific stimuli or contexts, and would not be observed in steady-state eQTL studies such as the one PRF scores are based on. Studies are beginning to catalogue stimulation-specific eQTLs, and these have uncovered additional disease overlaps not seen in previous eQTL maps (Fairfax et al. 2014; M. N. Lee et al. 2014).

A limitation of our fine-mapping approach is that we assume there is a single causal variant at each locus. Some other methods that integrate functional annotation, such as PAINTOR (Kichaev et al. 2014) and RiVIERA-MT (Y. Li and Kellis 2016), allow for multiple causal variants, with an ensuing increased cost in computing time. Although we have not done so here, we note that PRF score fine-mapping could be applied separately to the p values from multiple independent signals determined by conditional analyses at a locus.

We have discussed four examples of PRF score fine-mapping for individual loci, highlighting both successes and failures. Each prioritized variant's PRF score can be broken down into the contribution from individual annotations. This allows investigators to evaluate the information provided by PRF scores in the context of prior biological knowledge, such as which genes are implicated at a locus, as well as variant features not included in the model, such as changes to coding sequence, chromosome conformation, or proximity to transcription factor binding sites, non-coding RNAs, or splice sites.

One metric that has been frequently referenced as an indicator of fine-mapping performance is the average reduction in credible set size. When evaluating this metric for PRF scores we compared our results with permuted scores. Surprisingly, fine-mapping with permuted scores resulted in a similar reduction in credible set size as when true scores were used. A potential explanation for this is that in some cases the PRF score Bayesian priors overwhelm the association statistics. In our example this is unlikely because we only selected GWAS loci with a lead SNP $p < 10^{-8}$; PRF scores tend to vary by at most 8 units of (natural) log-likelihood, whereas the statistical association at such loci varies over at least 18 units of log-likelihood. It should be kept in mind, however, that the weaker the statistical association, the greater an effect variant priors can have in general. Another plausible explanation for spurious credible set size reduction is that reweighted variants "crowd out" other credible set variants. When a locus has many variants in high LD, the credible set will be large. Applying any non-uniform prior will boost some variant posterior probabilities, and this will necessarily crowd out other variants from the credible set. The extent to which this crowding out occurs should depend on the distribution of priors; strongly peaked priors, even if completely random, will lead to individual variants being selected at the expense of others, reducing the credible set size. Because many fine-mapping methods integrate prior probabilities with statistical associations, much like we have done with PRF scores, we believe this reflects a general unreliability of credible set size as an indicator of fine-mapping performance.

How should fine-mapping performance then be determined? Simulations can provide a set of "known" causal and non-causal variants to assess performance against, but this may not reflect performance on real data. The ideal metric would be comparison against a set of known causal variants already fine-mapped from GWAS data, and which have experimental validation. Although the number of such cases to date is small, it is growing, and progress is likely to accelerate with the widespread use of CRISPR-Cas9 to demonstrate molecular effects of alleles in human cell lines. The application of gene editing to dissect complex trait associations depends on there being a small set of variants to consider, particularly if allelic replacement is used to provide the highest-quality evidence for individual causal variants.

With thousands of GWAS associations whose causal mechanisms remain to be discovered, statistical and epigenomic fine-mapping are essential to broaden the number of loci where experimental follow-up is feasible.

4.5 Methods

R source code for identifying cell type enrichments and doing PRF score fine-mapping are available at https://github.com/Jeremy37/prfcalc.

Differences in epigenome mean PRF score

To determine global differences in epigenome mean PRF scores, we computed PRF scores for 288,091 positions spaced every 10 kb along the human genome in each of the 119 epigenomes, and determined the mean of these for each epigenome. To identify factors driving the differences, we determined the annotation contributions for 2,881 PRF scores at positions spaced every 1 Mb along the human genome. For each annotation we determined the mean annotation contribution to PRF scores in each epigenome, and correlated these values with the mean PRF score across epigenomes. Five annotations showed correlation r^2 above 0.1; these were H3K4me3, H3K27ac, H3K36me3, H3K4me1, and 18.EnhAc. These five annotations also showed strong correlations amongst each other.

GWAS summary statistics and locus definitions

We downloaded summary statistics for six GWAS from the following URLs:

GWAS	File / URL
RA	File: RA_GWASmeta_European_v2.txt.gz URL: http://plaza.umin.ac.jp/~yokada/datasource/software.htm
Height	File: GIANT_HEIGHT_Wood_et_al_2014_publicrelease_HapMapCeuFreq.txt.gz URL: http://portals.broadinstitute.org/collaboration/giant/index.php
IBD	File:ftp.sanger.ac.uk/pub/consortia/ibdgenetics/iibdgc-trans-ancestry-filtered-summary-stats.tgz
HDL cholesterol	File: jointGwasMc_HDL.txt.gz URL: http://www.sph.umich.edu/csg/abecasis/public/lipids2013/
Schizophrenia	File: ckqny.scz2snpres.gz URL: http://www.med.unc.edu/pgc/results-and-downloads
Educational attainment	File: EduYears_Main.txt.gz URL: https://www.thessgac.org/data

For each GWAS we either downloaded a file listing the associated loci or extracted these details from supplementary tables. For consistency across GWAS, we defined the associated regions as a window of +/- 200 kb around the lead SNP position. Because some regions overlapped, and PRF scores assume a single causal variant, we removed one region from each pair of overlapping regions until there were no overlaps. This left the a total of 1,002 regions across the six GWAS.

Correlation of weightedPRF and maxPRF scores

To determine consistency between weightedPRF and maxPRF scores, for each GWAS we determined the weightedPRF and maxPRF for all epigenomes at each locus. For each GWAS, we then computed the correlation of maxPRF and weightedPRF scores across loci and epigenomes. The correlation R^2 of maxPRF and weighted PRF for different GWAS ranged from 0.54 for IBD to 0.77 for EduYears.

Comparing PRF scores in trait-relevant and non-relevant epigenomes

To evaluate the utility of using PRF scores from likely trait-relevant cell types, we used the 72 loci across the five GWAS (height excluded) where a single variant had naive PPA > 0.5. We arbitrarily chose three epigenomes (placenta, foreskin keratinocytes, and adipose-derived MSCs) that seemed unlikely to be relevant to any of the GWAS traits. For each locus we calculated relativePRF scores by subtracting the median score at the locus, and then determined the median relativePRF score among the likely causal variants (PPA > 0.5) and separately among the remaining variants. We plotted in Figure 8 the difference between these two values when a trait-relevant epigenome was used for each GWAS, or when each of the three less relevant epigenomes was used.

Fine-mapping with PRF scores

To perform fine-mapping, we first computed approximate Bayes Factors as described in Chapter 3, and used the method of fgwas to determine the PPA for each variant. We determined the 95% credible set of variants, and then removed variants with PPA < 0.001. We next computed functional PPAs using Ch. 3 Equation 5, where the PRF score is the x_i value from Ch. 3 Equation 2. For plotting, we determined the set of variants representing the top 5 by naive PPA and the top 5 by functional PPA; these are highlighted with coloured points, are numbered by their statistical association rank, and have annotation bar plots shown. We manually examined many plots across the five GWAS to select four loci that represented different scenarios encountered in fine-mapping.

Changes to implicated genes

For either naive lead SNPs or functional lead SNPs, we used bedtools closest to determine the distance of SNPs to the nearest gene, defining the gene body as the outermost positions of Gencode v19 annotations for coding regions or UTRs of the gene. Thereby, variants in an intron of a gene had a distance of zero.

Comparison with stratified LDSC

We extracted the cell type associations for stratified LDSC (Finucane et al. 2015) from their supplementary table 8, including log10 P value, cell type, and mark. We assigned cell types to epigenome groups in the same way as for PRF scores. We plotted the uncorrected log10 P values for each method in Figure 7 for SCZ, RA, and HDL, and labeled the top 6 cell types. Because stratified LDSC uses individual annotations, some cell types appear multiple times based on different histone marks or separate assays for the same mark.

4.6 Appendix - PRF fine-mapping at additional loci

REL locus - an RA regulatory variant that is not the lead SNP

At the *REL* locus associated with RA, high LD results in about ten SNPs having statistical PPAs in the range 0.05 - 0.17. However, PRF score fine-mapping using the epigenome E116-GM12878 LCLs strongly supports SNP #2 (Figure A1), boosting its PPA from 0.13 to 0.77. The lead SNP at the locus, rs34695944, is located in an intron of the gene *REL*, a strong candidate gene for RA due to its involvement in inflammation, immunity, and proliferation of B lymphocytes via a complex with NFκB. This SNP has low levels of histone modifications H3K36me3, H3K27ac, and H3K4me3, which each contribute modestly to its PRF score of 6.2; however, it is not in a distinct peak for any of these marks. In contrast, SNP #2, rs67574266, has a PRF score of 8.3 due to its location at a conserved nucleotide in the 5' UTR of *REL*, in peaks of DNase hypersensitivity, H3K4me3, and H3K27ac. This SNP is located in a CTCF peak, and alters an invariant position of the CTCF binding motif. Surprisingly, the SNP is not detected as an eQTL for *REL* in GTEx, although it is a weak eQTL for the nearby gene *PUS10* in a handful of tissues.

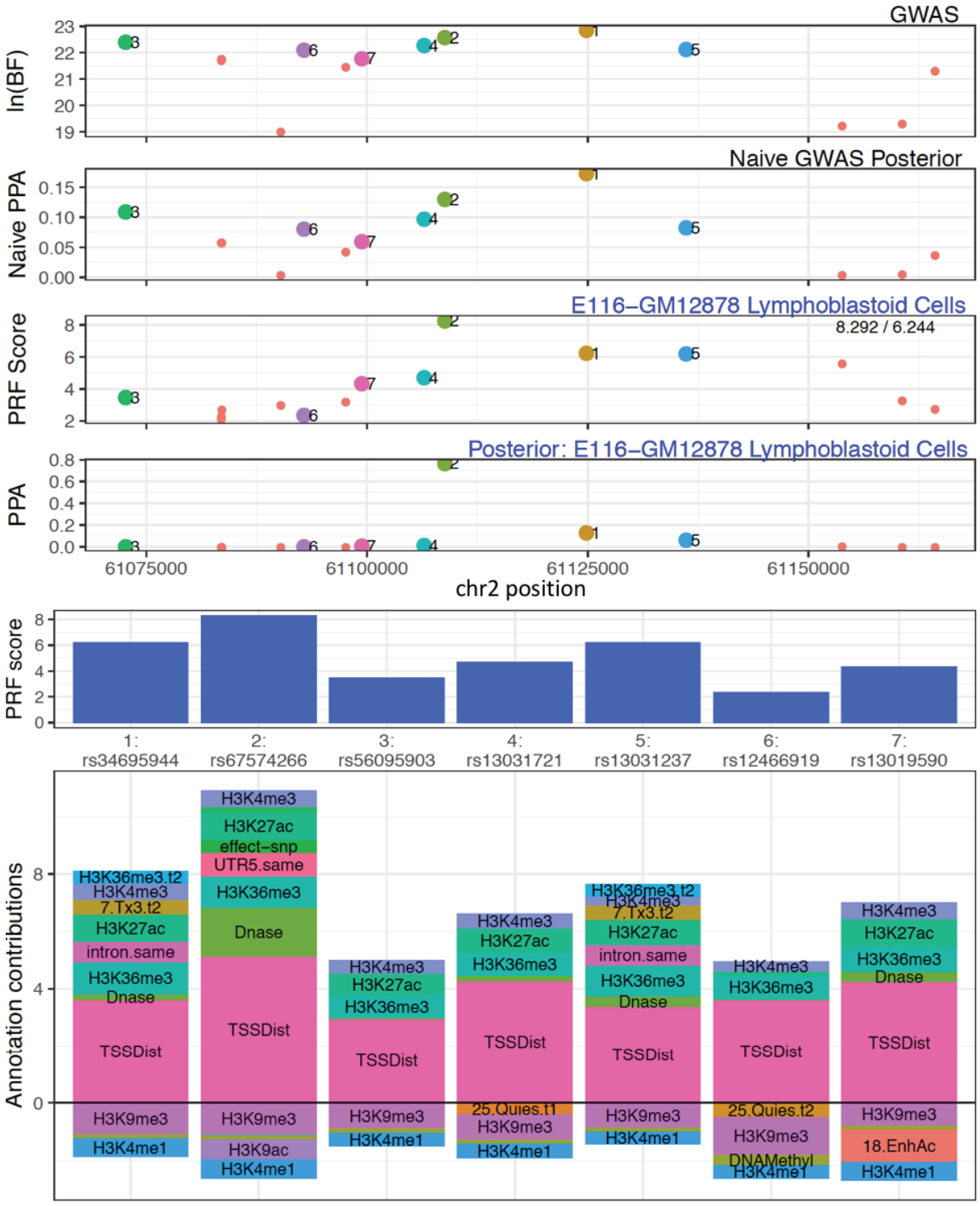


Figure A1: PRF score fine-mapping plots at the RA-associated *REL* locus.

SH2B3 locus - a nonsynonymous SNP for RA

A weak association for rheumatoid arthritis at 12q24 covers a large number of genes, with the association peak overlapping *SH2B3* and *ATXN2*. *SH2B3* is a likely gene to mediate the association, since it plays a critical role in hematopoiesis; in contrast, mutations of *ATXN2* are associated with spinocerebellar ataxia type 2, a neuromuscular and neurodegenerative disorder. The lead SNP is 10 kb upstream of *SH2B3*, but PRF scores strongly prioritize the second most-associated SNP, rs3184504, as likely causal, as it overlaps the coding portion of the gene with moderate levels of multiple histone modifications. Notably, this SNP is nonsynonymous, and therefore the change in coding sequence is more likely to be the cause for the association than changes to gene expression. This SNP is also the lead SNP for a large number of traits, including blood cell traits and other autoimmune disorders, further supporting it as a likely causal variant. This is one example among a number of others where coding variants receive a high PRF score, and are thereby prioritized as potentially causal. While PRF scores are not intended for use at loci with likely causal coding variants, they can potentially still provide useful information to investigators.

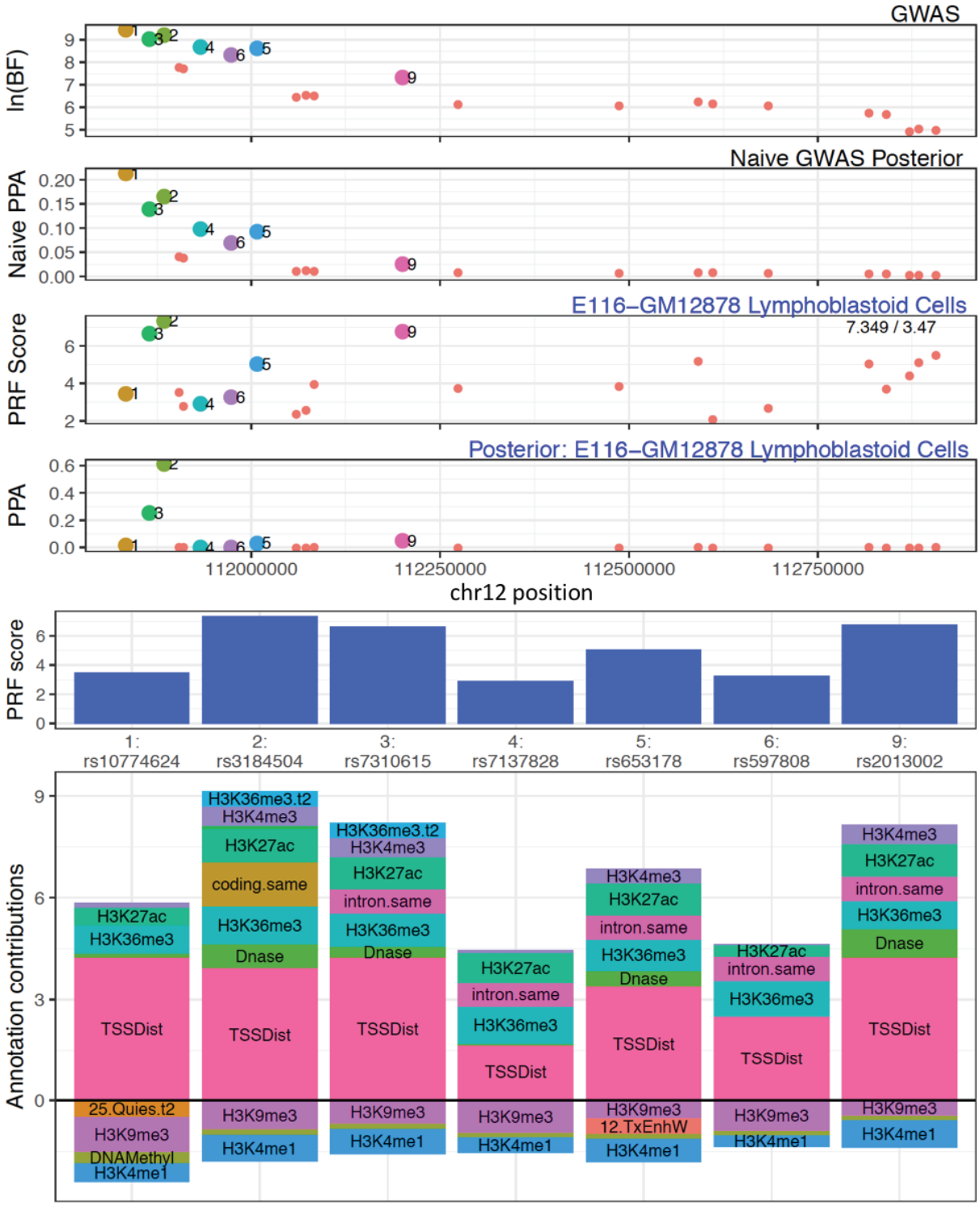


Figure A2: PRF score fine-mapping plots at the RA-associated *SH2B3* locus.

SLC22A4 locus - a possible regulatory SNP for IBD

At the *SLC22A4* locus associated with IBD, a number of SNPs in high LD receive modest PPAs when only statistical information is used. *SLC22A4* is a well-established risk gene for Crohn's disease (Peltekova et al. 2004). PRF scores highlight the lead SNP, rs35260072, as by far the most likely to be causal, as it occurs in many epigenetic marks in the first intron of *SLC22A4*. This SNP also overlaps a region of dense TF binding upstream of an alternative promoter of *SLC22A4*, and is predicted by centisnp to alter TF binding. Notably, there is a nonsynonymous SNP in *SLC22A4,* rs1050152, which is in high LD, and which should also be considered a strong candidate to be causal for the association.

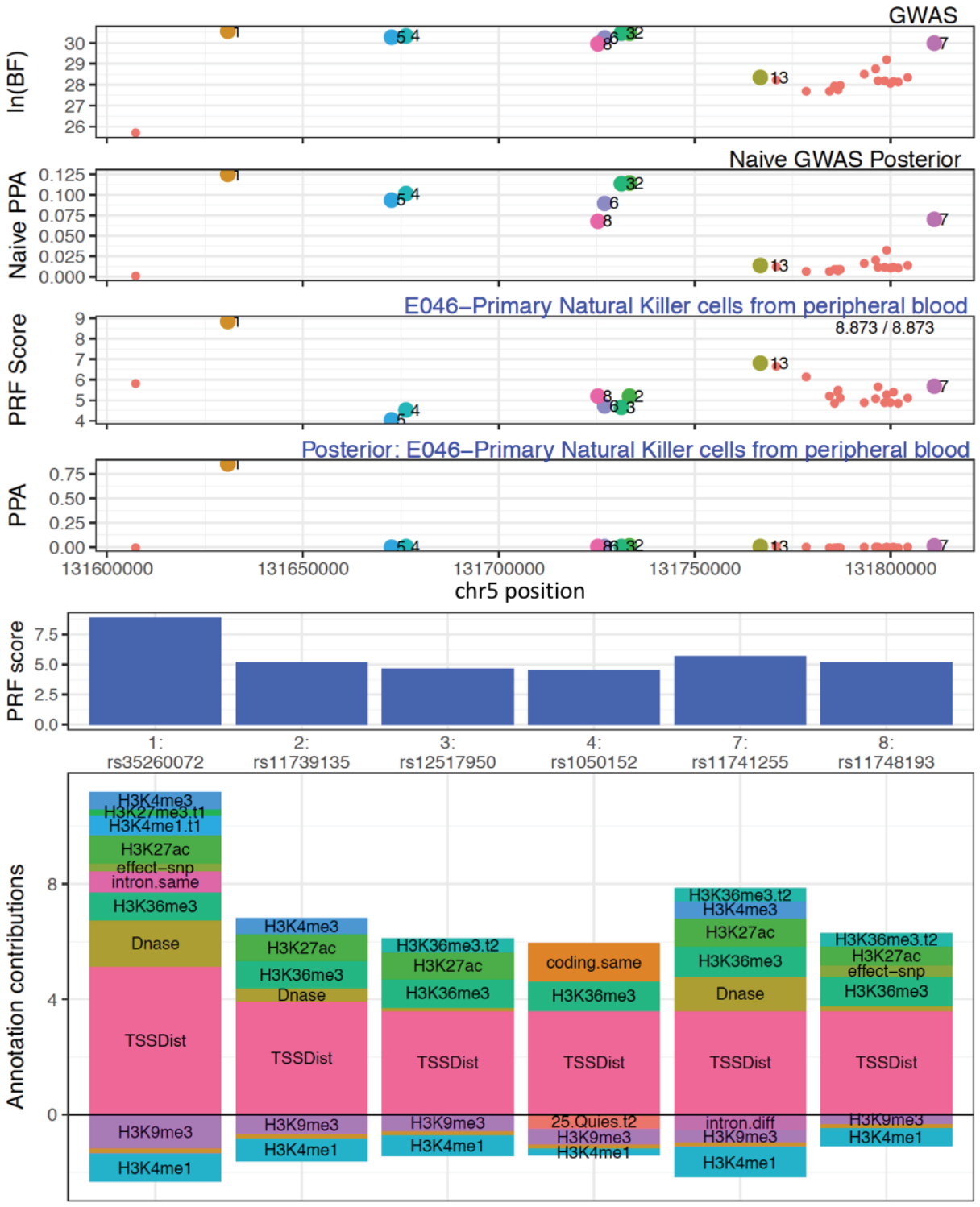


Figure A3: PRF score fine-mapping plot at the IBD-associated *SLC22A4* locus.

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.

6 References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Adzhubei, Ivan A., Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. 2010. “A Method and Server for Predicting Damaging Missense Mutations.” *Nature Methods* 7 (4): 248–49.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. “HTSeq--a Python Framework to Work with High-Throughput Sequencing Data.” *Bioinformatics* 31 (2): 166–69.

Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, et al. 2014. “An Atlas of Active Enhancers across Human Cell Types and Tissues.” *Nature* 507 (7493): 455–61.

Ashley, A. K., A. I. Hinds, W. H. Hanneman, R. B. Tjalkens, and M. E. Legare. 2016. “DJ-1 Mutation Decreases Astroglial Release of Inflammatory Mediators.” *Neurotoxicology* 52 (January): 198–203.

Barbeira, Alvaro, Scott P. Dickinson, Jason M. Torres, Rodrigo Bonazzola, Jiamao Zheng, Eric S. Torstenson, Heather E. Wheeler, et al. 2017. “Integrating Tissue Specific Mechanisms into GWAS Summary Results.” *bioRxiv*. doi:10.1101/045260.

Bardy, C., M. van den Hurk, B. Kakaradov, J. A. Erwin, B. N. Jaeger, R. V. Hernandez, T. Eames, et al. 2016. “Predicting the Functional States of Human iPSC-Derived Neurons with Single-Cell RNA-Seq and Electrophysiology.” *Molecular Psychiatry* 21 (11): 1573–88.

Bar-Nur, Ori, Holger A. Russ, Shimon Efrat, and Nissim Benvenisty. 2011. “Epigenetic Memory and Preferential Lineage-Specific Differentiation in Induced Pluripotent Stem Cells Derived from Human Pancreatic Islet Beta Cells.” *Cell Stem Cell* 9 (1): 17–23.

Baron, Maayan, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, et al. 2016. “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-Cell Population Structure.” *Cell Systems* 3 (4): 346–60.e4.

Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. 2007. “High-Resolution Profiling of Histone Methylations in the Human Genome.” *Cell* 129 (4): 823–37.

Battle, Alexis, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. 2015. “Genomic Variation. Impact of Regulatory Variation from RNA to Protein.” *Science* 347 (6222): 664–67.

Battle, Alexis, Sara Mostafavi, Xiaowei Zhu, James B. Potash, Myrna M. Weissman, Courtney McCormick, Christian D. Haudenschild, et al. 2014. “Characterizing the Genetic Basis of Transcriptome Diversity through RNA-Sequencing of 922 Individuals.” *Genome Research* 24 (1): 14–24.

Benjamini, Yuval, and Terence P. Speed. 2012. “Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing.” *Nucleic Acids Research* 40 (10): e72.

Benner, Christian, Chris C. A. Spencer, Aki S. Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. 2016. “FINEMAP: Efficient Variable Selection Using Summary Data from Genome-Wide Association Studies.” *Bioinformatics* 32 (10): 1493–1501.

Bielekova, B., M. Catalfamo, S. Reichert-Scrivner, A. Packer, M. Cerna, T. A. Waldmann, H. McFarland, P. A. Henkart, and R. Martin. 2006. “Regulatory CD56bright Natural Killer Cells Mediate Immunomodulatory Effects of IL-2R -Targeted Therapy (daclizumab) in Multiple Sclerosis.” *Proceedings of the National Academy of Sciences* 103 (15): 5941–46.

Bock, Christoph, Evangelos Kiskinis, Griet Verstappen, Hongcang Gu, Gabriella Boulting, Zachary D. Smith, Michael Ziller, et al. 2011. “Reference Maps of Human ES and iPS Cell Variation Enable High-Throughput Characterization of Pluripotent Cell Lines.” *Cell* 144 (3): 439–52.

Bojesen, Stig E., Karen A. Pooley, Sharon E. Johnatty, Jonathan Beesley, Kyriaki Michailidou, Jonathan P. Tyrer, Stacey L. Edwards, et al. 2013. “Multiple Independent Variants at the TERT Locus Are Associated with Telomere Length and Risks of Breast and Ovarian Cancer.” *Nature Genetics* 45 (4): 371–84, 384e1–2.

Boyle, Alan P., Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, et al. 2012. “Annotation of Functional Variation in Personal Genomes Using RegulomeDB.” *Genome Research* 22 (9): 1790–97.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10 (12): 1213–18.

Bunt, Martijn van de, Adrian Cortes, IGAS Consortium, Matthew A. Brown, Andrew P. Morris, and Mark I. McCarthy. 2015. “Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci.” *PLoS Genetics* 11 (9): e1005535.

Burrows, Courtney K., Nicholas E. Banovich, Bryan J. Pavlovic, Kristen Patterson, Irene Gallego Romero, Jonathan K. Pritchard, and Yoav Gilad. 2016. “Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in iPSCs.” *PLoS Genetics* 12 (1): e1005793.

Cacchiarelli, Davide, Xiaojie Qiu, Sanjay Srivatsan, Michael Ziller, Eliah Overbey, Jonna Grimsby, Prapti Pokharel, et al. 2017. “Aligning Single-Cell Developmental and Reprogramming Trajectories Identifies Molecular Determinants of Reprogramming Outcome.” *bioRxiv*. doi:10.1101/122531.

Cao, Lishuang, Aoibhinn McDonnell, Anja Nitzsche, Aristos Alexandrou, Pierre-Philippe Saintot, Alexandre J. C. Loucif, Adam R. Brown, et al. 2016. “Pharmacological Reversal of a Pain Phenotype in iPSC-Derived Sensory Neurons and Patients with Inherited Erythromelalgia.” *Science Translational Medicine* 8 (335): 335ra56.

Carbone, Fortunata, Veronica De Rosa, Pietro B. Carrieri, Silvana Montella, Dario Bruzzese, Antonio Porcellini, Claudio Procaccini, Antonio La Cava, and Giuseppe Matarese. 2014. “Regulatory T Cell Proliferative Potential Is Impaired in Human Autoimmune Disease.” *Nature Medicine* 20 (1): 69–74.

Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. 2015. “Tools and Best Practices for Data Processing in Allelic Expression Analysis.” *Genome Biology* 16 (September): 195.

Chambers, Stuart M., Yuchen Qi, Yvonne Mica, Gabsang Lee, Xin-Jun Zhang, Lei Niu, James Bilsland, et al. 2012. “Combined Small-Molecule Inhibition Accelerates Developmental Timing and Converts Human Pluripotent Stem Cells into Nociceptors.” *Nature Biotechnology* 30 (7): 715–20.

Chen, Wenan, Beth R. Larrabee, Inna G. Ovsyannikova, Richard B. Kennedy, Iana H. Haralambieva, Gregory A. Poland, and Daniel J. Schaid. 2015. “Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics.” *Genetics* 200 (3): 719–36.

Chong, Jessica X., Kati J. Buckingham, Shalini N. Jhangiani, Corinne Boehm, Nara Sobreira, Joshua D. Smith, Tanya M. Harrell, et al. 2015. “The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities.” *American Journal of Human Genetics* 97 (2): 199–215.

Chronis, Constantinos, Petko Fiziev, Bernadett Papp, Stefan Butz, Giancarlo Bonora, Shan Sabri, Jason Ernst, and Kathrin Plath. 2017. “Cooperative Binding of Transcription Factors Orchestrates Reprogramming.” *Cell* 168 (3): 442–59.e20.

Chun, Sung, Alexandra Casparino, Nikolaos A. Patsopoulos, Damien C. Croteau-Chonka, Benjamin A. Raby, Philip L. De Jager, Shamil R. Sunyaev, and Chris Cotsapas. 2017. “Limited Statistical Evidence for Shared Genetic Effects of eQTLs and Autoimmune-Disease-Associated Loci in Three Major Immune-Cell Types.” *Nature Genetics* 49 (4): 600–605.

Church, Chris, Lee Moir, Fiona McMurray, Christophe Girard, Gareth T. Banks, Lydia Teboul, Sara Wells, et al. 2010. “Overexpression of Fto Leads to Increased Food Intake and Results in Obesity.” *Nature Genetics* 42 (12): 1086–92.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w1118; Iso-2; Iso-3.” *Fly* 6 (2): 80–92.

Claussnitzer, Melina, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, et al. 2015. “FTO Obesity Variant Circuitry and Adipocyte Browning in Humans.” *The New England Journal of Medicine* 373 (10): 895–907.

Creyghton, Menno P., Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, et al. 2010. “Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (50): 21931–36.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58.

D’Antonio, Matteo, Grace Woodruff, Jason L. Nathanson, Agnieszka D’Antonio-Chronowska, Angelo Arias, Hiroko Matsui, Roy Williams, et al. 2017. “High-Throughput and Cost-Effective Characterization of Induced Pluripotent Stem Cells.” *Stem Cell Reports*, April. doi:10.1016/j.stemcr.2017.03.011.

Davis, Joe R., Laure Fresard, David A. Knowles, Mauro Pala, Carlos D. Bustamante, Alexis Battle, and Stephen B. Montgomery. 2016. “An Efficient Multiple-Testing Adjustment for eQTL Studies That Accounts for Linkage Disequilibrium between Variants.” *American Journal of Human Genetics* 98 (1): 216–24.

Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. “Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP.” *PLoS Computational Biology* 6 (12): e1001025.

Degner, Jacob F., Athma A. Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J. Gaffney, Joseph K. Pickrell, Sherryl De Leon, et al. 2012. “DNase I Sensitivity QTLs Are a Major Determinant of Human Expression Variation.” *Nature* 482 (7385): 390–94.

DeLuca, David S., Joshua Z. Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. 2012. “RNA-SeQC: RNA-Seq Metrics for Quality Control and Process Optimization.” *Bioinformatics* 28 (11): 1530–32.

Dianat, Noushin, Clara Steichen, Ludovic Vallier, Anne Weber, and Anne Dubart-Kupperschmitt. 2013. “Human Pluripotent Stem Cells for Modelling Human Liver Diseases and Cell Therapy.” *Current Gene Therapy* 13 (2): 120–32.

Dimas, A. S., S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, et al. 2009. “Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner.” *Science* 325 (5945): 1246–50.

Ding, Jun, Johann E. Gudjonsson, Liming Liang, Philip E. Stuart, Yun Li, Wei Chen, Michael Weichenthal, et al. 2010. “Gene Expression in Skin and Lymphoblastoid Cells: Refined Statistical Method Reveals Extensive Overlap in Cis-eQTL Signals.” *American Journal of Human Genetics* 87 (6): 779–89.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature* 485 (7398): 376–80.

Doniger, Scott W., and Justin C. Fay. 2007. “Frequent Gain and Loss of Functional Transcription Factor Binding Sites.” *PLoS Computational Biology* 3 (5): e99.

Dunham, Ian, Eugene Kulesha, Valentina Iotchkova, Sandro Morganella, and Ewan Birney. 2015. “FORGE: A Tool to Discover Cell Specific Enrichments of GWAS Associated SNPs in Regulatory Regions.” *F1000Research*. doi:10.12688/f1000research.6032.1.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

Ernst, Jason, and Manolis Kellis. 2012. “ChromHMM: Automating Chromatin-State Discovery and Characterization.” *Nature Methods* 9 (3): 215–16.

Ernst, Jason, and Manolis Kellis. 2015. “Large-Scale Imputation of Epigenomic Datasets for Systematic Annotation of Diverse Human Tissues.” *Nature Biotechnology* 33 (4): 364–76.

Evans, David M., Chris C. A. Spencer, Jennifer J. Pointon, Zhan Su, David Harvey, Grazyna Kochan, Udo Oppermann, et al. 2011. “Interaction between ERAP1 and HLA-B27 in Ankylosing Spondylitis Implicates Peptide Handling in the Mechanism for HLA-B27 in Disease Susceptibility.” *Nature Genetics* 43 (8): 761–67.

Fairfax, Benjamin P., Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, et al. 2014. “Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression.” *Science* 343 (6175): 1246949.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Alistair R. R. Forrest, Hideya Kawaji, Michael Rehli, J. Kenneth Baillie, Michiel J. L. de Hoon, Vanja Haberle, et al. 2014. “A Promoter-Level Mammalian Expression Atlas.” *Nature* 507 (7493): 462–70.

Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, et al. 2015. “Genetic and Epigenetic Fine Mapping of Causal Autoimmune Disease Variants.” *Nature* 518 (7539): 337–43.

Ferraro, A., A. M. D’Alise, T. Raj, N. Asinovski, R. Phillips, A. Ergun, J. M. Replogle, et al. 2014. “Interindividual Variation in Human T Regulatory Cells.” *Proceedings of the National Academy of Sciences* 111 (12): E1111–20.

Finucane, Hilary K., Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, et al. 2015. “Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics.” *Nature Genetics* 47 (11): 1228–35.

Fischer, Julia, Linda Koch, Christian Emmerling, Jeanette Vierkotten, Thomas Peters, Jens C. Brüning, and Ulrich Rüther. 2009. “Inactivation of the Fto Gene Protects from Obesity.” *Nature* 458 (7240): 894–98.

Fuchsberger, Christian, Jason Flannick, Tanya M. Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J. Gaulton, Clement Ma, et al. 2016. “The Genetic Architecture of Type 2 Diabetes.” *Nature* 536 (7614): 41–47.

Gaffney, Daniel J., Jean-Baptiste Veyrieras, Jacob F. Degner, Roger Pique-Regi, Athma A. Pai, Gregory E. Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. 2012. “Dissecting the Regulatory Architecture of Gene Expression QTLs.” *Genome Biology* 13 (1): R7.

Gagliano, Sarah A., Michael R. Barnes, Michael E. Weale, and Jo Knight. 2014. “A Bayesian Method to Incorporate Hundreds of Functional Characteristics with Association Evidence to Improve Variant Prioritization.” *PloS One* 9 (5): e98122.

Gamazon, Eric R., Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, et al. 2015. “A Gene-Based Association Method for Mapping Traits Using Reference Transcriptome Data.” *Nature Genetics* 47 (9): 1091–98.

Gates, Leah A., Jiejun Shi, Aarti D. Rohira, Qin Feng, Bokai Zhu, Mark T. Bedford, Cari A. Sagum, et al. 2017. “Acetylation on Histone H3 Lysine 9 Mediates a Switch from Transcription Initiation to Elongation.” *The Journal of Biological Chemistry* 292 (35): 14456–72.

Ghanbari, Mohsen, M. Arfan Ikram, Hans W. J. de Looper, Albert Hofman, Stefan J. Erkeland, Oscar H. Franco, and Abbas Dehghan. 2016. “Genome-Wide Identification of microRNA-Related Variants Associated with Risk of Alzheimer’s Disease.” *Scientific Reports* 6 (1). doi:10.1038/srep28387.

Giambartolomei, Claudia, Damjan Vukcevic, Eric E. Schadt, Lude Franke, Aroon D. Hingorani, Chris Wallace, and Vincent Plagnol. 2014. “Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics.” *PLoS Genetics* 10 (5): e1004383.

Global Lipids Genetics Consortium, Cristen J. Willer, Ellen M. Schmidt, Sebanti Sengupta, Gina M. Peloso, Stefan Gustafsson, Stavroula Kanoni, et al. 2013. “Discovery and Refinement of Loci Associated with Lipid Levels.” *Nature Genetics* 45 (11): 1274–83.

Glubb, Dylan M., Mel J. Maranian, Kyriaki Michailidou, Karen A. Pooley, Kerstin B. Meyer, Siddhartha Kar, Saskia Carlebur, et al. 2015. “Fine-Scale Mapping of the 5q11.2 Breast Cancer Locus Reveals at Least Three Independent Risk Variants Regulating MAP3K1.” *American Journal of Human Genetics* 96 (1): 5–20.

Gonzàlez-Porta, Mar, Adam Frankish, Johan Rung, Jennifer Harrow, and Alvis Brazma. 2013. “Transcriptome Analysis of Human Tissues and Cell Lines Reveals One Dominant Transcript per Gene.” *Genome Biology* 14 (7): R70.

Gregory, Adam P., Calliope A. Dendrou, Kathrine E. Attfield, Aiden Haghikia, Dionysia K. Xifara, Falk Butter, Gereon Poschmann, et al. 2012. “TNF Receptor 1 Genetic Risk Mirrors Outcome of Anti-TNF Therapy in Multiple Sclerosis.” *Nature* 488 (7412): 508–11.

Grzybek, Maciej, Aleksandra Golonko, Marta Walczak, and Pawel Lisowski. 2017. “Epigenetics of Cell Fate Reprogramming and Its Implications for Neurological Disorders Modelling.” *Neurobiology of Disease* 99 (March): 84–120.

GTEx Consortium. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45 (6): 580–85.

GTEx Consortium. 2015. “Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans.” *Science* 348 (6235): 648–60.

GTEx Consortium et al. 2017. “Genetic Effects on Gene Expression across Human Tissues.” *Nature* 550 (7675): 204–13.

Guenther, Catherine A., Bosiljka Tasic, Liqun Luo, Mary A. Bedell, and David M. Kingsley. 2014. “A Molecular Basis for Classic Blond Hair Color in Europeans.” *Nature Genetics* 46 (7): 748–52.

Gulko, Brad, Melissa J. Hubisz, Ilan Gronau, and Adam Siepel. 2015. “A Method for Calculating Probabilities of Fitness Consequences for Point Mutations across the Human Genome.” *Nature Genetics* 47 (3): 276–83.

Gupta, Rajat M., Joseph Hadaya, Aditi Trehan, Seyedeh M. Zekavat, Carolina Roselli, Derek Klarin, Connor A. Emdin, et al. 2017. “A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression.” *Cell* 170 (3): 522–33.e15.

Gusev, Alexander, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W. J. H.

Penninx, Rick Jansen, et al. 2016. “Integrative Approaches for Large-Scale Transcriptome-Wide Association Studies.” *Nature Genetics* 48 (3): 245–52.

Hait, Tom Aharon, David Amar, Ron Shamir, and Ran Elkon. 2017. “An Extensive Enhancer-Promoter Map Generated by Genome-Scale Analysis of Enhancer and Gene Activity Patterns.” *bioRxiv*. doi:10.1101/190231.

Handel, Adam E., Satyan Chintawar, Tatjana Lalic, Emma Whiteley, Jane Vowles, Alice Giustacchini, Karene Argoud, et al. 2016. “Assessing Similarity to Primary Tissue and Cortical Layer Identity in Induced Pluripotent Stem Cell-Derived Cortical Neurons through Single-Cell Transcriptomics.” *Human Molecular Genetics* 25 (5): 989–1000.

Hansen, Kasper D., Rafael A. Irizarry, and Zhijin Wu. 2012. “Removing Technical Variability in RNA-Seq Data Using Conditional Quantile Normalization.” *Biostatistics* 13 (2): 204–16.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. “GENCODE: The Reference Human Genome Annotation for The ENCODE Project.” *Genome Research* 22 (9): 1760–74.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. “Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities.” *Molecular Cell* 38 (4): 576–89.

Hindorff, Lucia A., Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio. 2009. “Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (23): 9362–67.

Hoffman, Michael M., Orion J. Buske, Jie Wang, Zhiping Weng, Jeff A. Bilmes, and William Stafford Noble. 2012. “Unsupervised Pattern Discovery in Human Chromatin Structure through Genomic Segmentation.” *Nature Methods* 9 (5): 473–76.

Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R. Abecasis. 2012. “Fast and Accurate Genotype Imputation in Genome-Wide Association Studies through Pre-Phasing.” *Nature Genetics* 44 (8): 955–59.

Huang, Hailiang, Ming Fang, Luke Jostins, Maša Umićević Mirkov, Gabrielle Boucher, Carl A. Anderson, Vibeke Andersen, et al. 2017. “Fine-Mapping Inflammatory Bowel Disease Loci to Single-Variant Resolution.” *Nature* 547 (7662): 173–78.

Huang, Yi-Fei, Brad Gulko, and Adam Siepel. 2017. “Fast, Scalable Prediction of Deleterious Noncoding Variants from Functional and Population Genomic Data.” *Nature Genetics* 49 (4): 618–24.

Hu, Bao-Yang, Jason P. Weick, Junying Yu, Li-Xiang Ma, Xiao-Qing Zhang, James A. Thomson, and Su-Chun Zhang. 2010. “Neural Differentiation of Human Induced Pluripotent Stem Cells Follows Developmental Principles but with Variable Potency.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (9): 4335–40.

Hunt, S. P., A. Pini, and G. Evan. 1987. “Induction of c-Fos-like Protein in Spinal Cord Neurons Following Sensory Stimulation.” *Nature* 328 (6131): 632–34.

Inoue, Fumitaka, Martin Kircher, Beth Martin, Gregory M. Cooper, Daniela M. Witten, Michael T. McManus, Nadav Ahituv, and Jay Shendure. 2016. “A Systematic Comparison Reveals Substantial Differences in Chromosomal versus Episomal Encoding of Enhancer Activity.” *Genome Research*. doi:10.1101/gr.212092.116.

Ioannidis, Nilah M., Joe R. Davis, Marianne K. DeGorter, Nicholas B. Larson, Shannon K. McDonnell, Amy J. French, Alexis J. Battle, et al. 2017. “FIRE: Functional Inference of

Genetic Variants That Regulate Gene Expression." *Bioinformatics* , August. doi:10.1093/bioinformatics/btx534.

Iotchkova, Valentina, Graham R. S. Ritchie, Matthias Geihs, Sandro Morganella, Josine L. Min, Klaudia Walter, Nicholas J. Timpson, et al. 2016. "GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD Correction." *bioRxiv*. doi:10.1101/085738.

Itzhaki, Ilanit, Leonid Maizels, Irit Huber, Limor Zwi-Dantsis, Oren Caspi, Aaron Winterstern, Oren Feldman, et al. 2011. "Modelling the Long QT Syndrome with Induced Pluripotent Stem Cells." *Nature* 471 (7337): 225–29.

Iwafuchi-Doi, Makiko, Greg Donahue, Akshay Kakumanu, Jason A. Watts, Shaun Mahony, B. Franklin Pugh, Dolim Lee, Klaus H. Kaestner, and Kenneth S. Zaret. 2016. "The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation." *Molecular Cell* 62 (1): 79–91.

Iyer, Matthew K., Yashar S. Niknafs, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R. Barrette, et al. 2015. "The Landscape of Long Noncoding RNAs in the Human Transcriptome." *Nature Genetics* 47 (3): 199–208.

Javierre, Biola M., Oliver S. Burren, Steven P. Wilder, Roman Kreuzhuber, Steven M. Hill, Sven Sewitz, Jonathan Cairns, et al. 2016. "Lineage-Specific Genome Architecture Links Enhancers and Non-Coding Disease Variants to Target Gene Promoters." *Cell* 167 (5): 1369–84.e19.

Jordan, J. Dedrick, John Cijiang He, Narat J. Eungdamrong, Ivone Gomes, Wasif Ali, Tracy Nguyen, Trever G. Bivona, Mark R. Philips, Lakshmi A. Devi, and Ravi Iyengar. 2005. "Cannabinoid Receptor-Induced Neurite Outgrowth Is Mediated by Rap1 Activation through G(alpha)o/i-Triggered Proteasomal Degradation of Rap1GAPII." *The Journal of Biological Chemistry* 280 (12): 11413–21.

Jun, Goo, Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Gonçalo R. Abecasis, Michael Boehnke, and Hyun Min Kang. 2012. "Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data." *American Journal of Human Genetics* 91 (5): 839–48.

Jung, Seung Hyo, Kyung Jong Won, Kang Pa Lee, Dong Hyun Lee, Suyeol Yu, Dong-Youb Lee, Eun-Hye Seo, et al. 2014. "DJ-1 Protein Regulates CD3+ T Cell Migration via Overexpression of CXCR4 Receptor." *Atherosclerosis* 235 (2): 503–9.

Kathiresan, Sekar, Olle Melander, Candace Guiducci, Aarti Surti, Noël P. Burtt, Mark J. Rieder, Gregory M. Cooper, et al. 2008. "Six New Loci Associated with Blood Low-Density Lipoprotein Cholesterol, High-Density Lipoprotein Cholesterol or Triglycerides in Humans." *Nature Genetics* 40 (2): 189–97.

Kelley, David R., Jasper Snoek, and John L. Rinn. 2016. "Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks." *Genome Research* 26 (7): 990–99.

Kichaev, Gleb, Megan Roytman, Ruth Johnson, Eleazar Eskin, Sara Lindström, Peter Kraft, and Bogdan Pasaniuc. 2017. "Improved Methods for Multi-Trait Fine Mapping of Pleiotropic Risk Loci." *Bioinformatics* 33 (2): 248–55.

Kichaev, Gleb, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L. Price, Peter Kraft, and Bogdan Pasaniuc. 2014. "Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies." *PLoS Genetics* 10 (10): e1004722.

Kilpinen, Helena, Angela Goncalves, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, et al. 2017. "Common Genetic Variation Drives Molecular Heterogeneity in Human iPSCs." *Nature* 546 (7658): 370–75.

Kim, K., A. Doi, B. Wen, K. Ng, R. Zhao, P. Cahan, J. Kim, et al. 2010. "Epigenetic Memory

in Induced Pluripotent Stem Cells." *Nature* 467 (7313): 285–90.

Kindt, Alida S. D., Pau Navarro, Colin A. M. Semple, and Chris S. Haley. 2013. "The Genomic Signature of Trait-Associated Variants." *BMC Genomics* 14 (February): 108.

Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46 (3): 310–15.

Kiselev, Vladimir Yu, Kristina Kirschner, Michael T. Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N. Natarajan, et al. 2017. "SC3: Consensus Clustering of Single-Cell RNA-Seq Data." *Nature Methods* 14 (5): 483–86.

Kléber, Maurice, Hye-Youn Lee, Heiko Wurdak, Johanna Buchstaller, Martin M. Riccomagno, Lars M. Ittner, Ueli Suter, Douglas J. Epstein, and Lukas Sommer. 2005. "Neural Crest Stem Cell Maintenance by Combinatorial Wnt and BMP Signaling." *The Journal of Cell Biology* 169 (2): 309–20.

Kohno, Tatsuro, Kimberly A. Moore, Hiroshi Baba, and Clifford J. Woolf. 2003. "Peripheral Nerve Injury Alters Excitatory Synaptic Transmission in Lamina II of the Rat Dorsal Horn." *The Journal of Physiology* 548 (Pt 1): 131–38.

Kolde, Raivo, Sven Laur, Priit Adler, and Jaak Vilo. 2012. "Robust Rank Aggregation for Gene List Integration and Meta-Analysis." *Bioinformatics* 28 (4): 573–80.

Kumar, Prateek, Steven Henikoff, and Pauline C. Ng. 2009. "Predicting the Effects of Coding Non-Synonymous Variants on Protein Function Using the SIFT Algorithm." *Nature Protocols* 4 (7): 1073–81.

Kumasaka, Natsuhiko, Andrew J. Knights, and Daniel J. Gaffney. 2016. "Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq." *Nature Genetics* 48 (2): 206–13.

Lahens, Nicholas F., Ibrahim Halil Kavakli, Ray Zhang, Katharina Hayer, Michael B. Black, Hannah Dueck, Angel Pizarro, et al. 2014. "IVT-Seq Reveals Extreme Bias in RNA Sequencing." *Genome Biology* 15 (6): R86.

Lange, Katrina M. de, Loukas Moutsianas, James C. Lee, Christopher A. Lamb, Yang Luo, Nicholas A. Kennedy, Luke Jostins, et al. 2017. "Genome-Wide Association Study Implicates Immune Activation of Multiple Integrin Genes in Inflammatory Bowel Disease." *Nature Genetics* 49 (2): 256–61.

Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013. "Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans." *Nature* 501 (7468): 506–11.

Lee, Dongwon, David U. Gorkin, Maggie Baker, Benjamin J. Strober, Alessandro L. Asoni, Andrew S. McCallion, and Michael A. Beer. 2015. "A Method to Predict the Impact of Regulatory Variants from DNA Sequence." *Nature Genetics* 47 (8): 955–61.

Lee, Gabsang, Lee Gabsang, Eirini P. Papapetrou, Kim Hyesoo, Stuart M. Chambers, Mark J. Tomishima, Christopher A. Fasano, et al. 2009. "Modelling Pathogenesis and Treatment of Familial Dysautonomia Using Patient-Specific iPSCs." *Nature* 461 (7262): 402–6.

Lee, Mark N., Chun Ye, Alexandra-Chloé Villani, Towfique Raj, Weibo Li, Thomas M. Eisenhaure, Selina H. Imboywa, et al. 2014. "Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells." *Science* 343 (6175): 1246980.

Lee, Seunggeung, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin. 2014. "Rare-Variant Association Analysis: Study Designs and Statistical Tests." *American Journal of Human Genetics* 95 (1): 5–23.

Lessard, Julie, Jiang I. Wu, Jeffrey A. Ranish, Mimi Wan, Monte M. Winslow, Brett T. Staahl, Hai Wu, Ruedi Aebersold, Isabella A. Graef, and Gerald R. Crabtree. 2007. "An

Essential Switch in Subunit Composition of a Chromatin Remodeling Complex during Neural Development." *Neuron* 55 (2): 201–15.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7): 923–30.

Ling, Hui, Riccardo Spizzo, Yaser Atlasi, Milena Nicoloso, Masayoshi Shimizu, Roxana S. Redis, Naohiro Nishida, et al. 2013. "CCAT2, a Novel Noncoding RNA Mapping to 8q24, Underlies Metastatic Progression and Chromosomal Instability in Colon Cancer." *Genome Research* 23 (9): 1446–61.

Liu, Guang-Hui, Basam Z. Barkho, Sergio Ruiz, Dinh Diep, Jing Qu, Sheng-Lian Yang, Athanasia D. Panopoulos, et al. 2011. "Recapitulation of Premature Ageing with iPSCs from Hutchinson-Gilford Progeria Syndrome." *Nature* 472 (7342): 221–25.

Liu, Jimmy Z., Suzanne van Sommeren, Hailiang Huang, Siew C. Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, et al. 2015. "Association Analyses Identify 38 Susceptibility Loci for Inflammatory Bowel Disease and Highlight Shared Genetic Risk across Populations." *Nature Genetics* 47 (9): 979–86.

Liu, Wenjun, Hailong Wu, Lili Chen, Yankai Wen, Xiaoni Kong, and Wei-Qiang Gao. 2015. "Park7 Interacts with p47(phox) to Direct NADPH Oxidase-Dependent ROS Production and Protect against Sepsis." *Cell Research* 25 (6): 691–706.

Li, Yang I., Bryce van de Geijn, Anil Raj, David A. Knowles, Allegra A. Petti, David Golan, Yoav Gilad, and Jonathan K. Pritchard. 2016. "RNA Splicing Is a Primary Link between Genetic Variation and Disease." *Science* 352 (6285): 600–604.

Li, Yang I., David A. Knowles, and Jonathan K. Pritchard. 2016. "LeafCutter: Annotation-Free Quantification of RNA Splicing." *bioRxiv*. doi:10.1101/044107.

Li, Yue, and Manolis Kellis. 2016. "RiVIERA-MT: A Bayesian Model to Infer Risk Variants in Related Traits Using Summary Statistics and Functional Genomic Annotations." *bioRxiv*. doi:10.1101/059345.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." doi:10.1101/002832.

Lu, Qiongshi, Xinwei Yao, Yiming Hu, and Hongyu Zhao. 2016. "GenoWAP: GWAS Signal Prioritization through Integrated Analysis of Genomic Functional Annotation." *Bioinformatics* 32 (4): 542–48.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–53.

Mathelier, Anthony, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W. Wasserman. 2016. "DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo." *Cell Systems* 3 (3): 278–86.e4.

McCarthy, Davis J., Peter Humburg, Alexander Kanapin, Manuel A. Rivas, Kyle Gaulton, Jean-Baptiste Cazier, Peter Donnelly, and asds. 2014. "Choice of Transcripts and Software Has a Large Effect on Variant Annotation." *Genome Medicine* 6 (3): 26.

McCarthy, Shane, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R. Wood, Alexander Teumer, Hyun Min Kang, et al. 2016. "A Reference Panel of 64,976 Haplotypes for Genotype Imputation." *Nature Genetics* 48 (10): 1279–83.

McClellan, Jon, and Mary-Claire King. 2010. "Genetic Heterogeneity in Human Disease." *Cell* 141 (2): 210–17.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." *Genome Biology* 17 (1): 122.

Melchionda, Laura, Tobias B. Haack, Steven Hardy, Truus E. M. Abbink, Erika Fernandez-Vizarra, Eleonora Lamantea, Silvia Marchet, et al. 2014. “Mutations in APOPT1, Encoding a Mitochondrial Protein, Cause Cavitating Leukoencephalopathy with Cytochrome c Oxidase Deficiency.” *American Journal of Human Genetics* 95 (3): 315–25.

Mele, M., P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, et al. 2015. “The Human Transcriptome across Tissues and Individuals.” *Science* 348 (6235): 660–65.

Melzer, David, John R. B. Perry, Dena Hernandez, Anna-Maria Corsi, Kara Stevens, Ian Rafferty, Fulvio Lauretani, et al. 2008. “A Genome-Wide Association Study Identifies Protein Quantitative Trait Loci (pQTLs).” *PLoS Genetics* 4 (5): e1000072.

Moyerbrailean, Gregory A., Cynthia A. Kalita, Chris T. Harvey, Xiaoquan Wen, Francesca Luca, and Roger Pique-Regi. 2016. “Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding?” *PLoS Genetics* 12 (2): e1005875.

Mumbach, Maxwell R., Ansuman T. Satpathy, Evan A. Boyle, Chao Dai, Benjamin G. Gowen, Seung Woo Cho, Michelle L. Nguyen, et al. 2017. “Enhancer Connectome in Primary Human Cells Identifies Target Genes of Disease-Associated DNA Elements.” *Nature Genetics*, September. doi:10.1038/ng.3963.

Musunuru, Kiran, Alanna Strong, Maria Frank-Kamenetsky, Noemi E. Lee, Tim Ahfeldt, Katherine V. Sachs, Xiaoyu Li, et al. 2010. “From Noncoding Variant to Phenotype via SORT1 at the 1p13 Cholesterol Locus.” *Nature* 466 (7307): 714–19.

Newman, Aaron M., Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu, Chuong D. Hoang, Maximilian Diehn, and Ash A. Alizadeh. 2015. “Robust Enumeration of Cell Subsets from Tissue Expression Profiles.” *Nature Methods* 12 (5): 453–57.

Nica, Alexandra C., Leopold Parts, Daniel Glass, James Nisbet, Amy Barrett, Magdalena Sekowska, Mary Travers, et al. 2011. “The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study.” *PLoS Genetics* 7 (2): e1002003.

Nicolae, Dan L., Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, and Nancy J. Cox. 2010. “Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS.” *PLoS Genetics* 6 (4): e1000888.

Okada, Yukinori, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, et al. 2014. “Genetics of Rheumatoid Arthritis Contributes to Biology and Drug Discovery.” *Nature* 506 (7488): 376–81.

Okbay, Aysu, Jonathan P. Beauchamp, Mark Alan Fontana, James J. Lee, Tune H. Pers, Cornelius A. Rietveld, Patrick Turley, et al. 2016. “Genome-Wide Association Study Identifies 74 Loci Associated with Educational Attainment.” *Nature* 533 (7604): 539–42.

Ongen, Halit, Alfonso Buil, Andrew Anand Brown, Emmanouil T. Dermitzakis, and Olivier Delaneau. 2016. “Fast and Efficient QTL Mapper for Thousands of Molecular Phenotypes.” *Bioinformatics* 32 (10): 1479–85.

Palazzo, Alexander F., and Eliza S. Lee. 2015. “Non-Coding RNA: What Is Functional and What Is Junk?” *Frontiers in Genetics* 6 (January): 2.

Panopoulos, Athanasia D., Matteo D’Antonio, Paola Benaglio, Roy Williams, Sherin I. Hashem, Bernhard M. Schuldt, Christopher DeBoever, et al. 2017. “iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types.” *Stem Cell Reports*, April. doi:10.1016/j.stemcr.2017.03.012.

Papalexi, Efthymia, and Rahul Satija. 2017. “Single-Cell RNA Sequencing to Explore Immune Cell Heterogeneity.” *Nature Reviews. Immunology*, August. doi:10.1038/nri.2017.76.

Parkes, Miles, Adrian Cortes, David A. van Heel, and Matthew A. Brown. 2013. “Genetic Insights into Common Pathways and Complex Relationships among Immune-Mediated Diseases.” *Nature Reviews. Genetics* 14 (9): 661–73.

Pashos, Evanthia E., Yoson Park, Xiao Wang, Avanthi Raghavan, Wenli Yang, Deepti Abbey, Derek T. Peters, et al. 2017. “Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci.” *Cell Stem Cell* 20 (4): 558–70.e10.

Pasquali, Lorenzo, Kyle J. Gaulton, Santiago A. Rodríguez-Seguí, Loris Mularoni, Irene Miguel-Escalada, İldem Akerman, Juan J. Tena, et al. 2014. “Pancreatic Islet Enhancer Clusters Enriched in Type 2 Diabetes Risk-Associated Variants.” *Nature Genetics* 46 (2): 136–43.

Passier, Robert, Valeria Orlova, and Christine Mummery. 2016. “Complex Tissue and Disease Modeling Using hiPSCs.” *Cell Stem Cell* 18 (3): 309–21.

Peltekova, Vanya D., Richard F. Wintle, Laurence A. Rubin, Christopher I. Amos, Qiqing Huang, Xiangjun Gu, Bill Newman, et al. 2004. “Functional Variants of OCTN Cation Transporter Genes Are Associated with Crohn Disease.” *Nature Genetics* 36 (5): 471–75.

Pers, Tune H., Pascal Timshel, and Joel N. Hirschhorn. 2015. “SNPsnap: A Web-Based Tool for Identification and Annotation of Matched SNPs.” *Bioinformatics* 31 (3): 418–20.

Peters, Marjolein J., Linda Broer, Hanneke L. D. M. Willemen, Gudny Eiriksdottir, Lynne J. Hocking, Kate L. Holliday, Michael A. Horan, et al. 2013. “Genome-Wide Association Study Meta-Analysis of Chronic Widespread Pain: Evidence for Involvement of the 5p15.2 Region.” *Annals of the Rheumatic Diseases* 72 (3): 427–36.

Pickrell, Joseph K. 2014. “Joint Analysis of Functional Genomic Data and Genome-Wide Association Studies of 18 Human Traits.” *American Journal of Human Genetics* 94 (4): 559–73.

Pickrell, Joseph K., John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. 2010. “Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing.” *Nature* 464 (7289): 768–72.

Polo, Jose M., Susanna Liu, Maria Eugenia Figueroa, Warakorn Kulalert, Sarah Eminli, Kah Yong Tan, Effie Apostolou, et al. 2010. “Cell Type of Origin Influences the Molecular and Functional Properties of Mouse Induced Pluripotent Stem Cells.” *Nature Biotechnology* 28 (8): 848–55.

Praetorius, Christian, Christine Grill, Simon N. Stacey, Alexander M. Metcalf, David U. Gorkin, Kathleen C. Robinson, Eric Van Otterloo, et al. 2013. “A Polymorphism in IRF4 Affects Human Pigmentation through a Tyrosinase-Dependent MITF/TFAP2A Pathway.” *Cell* 155 (5): 1022–33.

Raghavan, Avanthi, Raghavan Avanthi, Wang Xiao, Rogov Peter, Wang Li, Zhang Xiaolan, Tarjei S. Mikkelsen, and Musunuru Kiran. 2016. “High-Throughput Screening and CRISPR-Cas9 Modeling of Causal Lipid-Associated Expression Quantitative Trait Locus Variants.” doi:10.1101/056820.

Ramasamy, Adaikalavan, Daniah Trabzuni, Sebastian Guelfi, Vibin Varghese, Colin Smith, Robert Walker, Tisham De, et al. 2014. “Genetic Variability in the Regulation of Gene Expression in Ten Regions of the Human Brain.” *Nature Neuroscience* 17 (10): 1418–28.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80.

Ritchie, Graham R. S., Ian Dunham, Eleftheria Zeggini, and Paul Flicek. 2014. “Functional

Annotation of Noncoding Sequence Variants." *Nature Methods* 11 (3): 294–96.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. "Integrative Analysis of 111 Reference Human Epigenomes." *Nature* 518 (7539): 317–30.

Robinton, Daisy A., and George Q. Daley. 2012. "The Promise of Induced Pluripotent Stem Cells in Research and Therapy." *Nature* 481 (7381): 295–305.

Ruffell, Daniela, Foteini Mourkioti, Adriana Gambardella, Peggy Kirstetter, Rodolphe G. Lopez, Nadia Rosenthal, and Claus Nerlov. 2009. "A CREB-C/EBPbeta Cascade Induces M2 Macrophage-Specific Gene Expression and Promotes Muscle Injury Repair." *Proceedings of the National Academy of Sciences of the United States of America* 106 (41): 17475–80.

Ryan, Niamh M., Stewart W. Morris, David J. Porteous, Martin S. Taylor, and Kathryn L. Evans. 2014. "SuRFing the Genomics Wave: An R Package for Prioritising SNPs by Functionality." *Genome Medicine* 6 (10): 79.

Sala, Luca, Milena Bellin, and Christine L. Mummery. 2016. "Integrating Cardiomyocytes from Human Pluripotent Stem Cells in Safety Pharmacology: Has the Time Come?" *British Journal of Pharmacology*, September. doi:10.1111/bph.13577.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. "Biological Insights from 108 Schizophrenia-Associated Genetic Loci." *Nature* 511 (7510): 421–27.

Schork, Andrew J., Wesley K. Thompson, Phillip Pham, Ali Torkamani, J. Cooper Roddey, Patrick F. Sullivan, John R. Kelsoe, et al. 2013. "All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs." *PLoS Genetics* 9 (4): e1003449.

Sheffield, Nathan C., and Bock Christoph. 2015. "LOLA: Enrichment Analysis for Genomic Region Sets and Regulatory Elements in R and Bioconductor." *Bioinformatics* 32 (4): 587–89.

Sheffield, Nathan C., Robert E. Thurman, Lingyun Song, Alexias Safi, John A. Stamatoyannopoulos, Boris Lenhard, Gregory E. Crawford, and Terrence S. Furey. 2013. "Patterns of Regulatory Activity across Diverse Human Cell Types Predict Tissue Identity, Transcription Factor Binding, and Long-Range Interactions." *Genome Research* 23 (5): 777–88.

Shihab, Hashem A., Mark F. Rogers, Julian Gough, Matthew Mort, David N. Cooper, Ian N. M. Day, Tom R. Gaunt, and Colin Campbell. 2015. "An Integrative Approach to Predicting the Functional Effects of Non-Coding and Coding Sequence Variation." *Bioinformatics* 31 (10): 1536–43.

Siepel, Adam, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, et al. 2005. "Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes." *Genome Research* 15 (8): 1034–50.

Singh, Tarjinder, Adam P. Levine, Philip J. Smith, Andrew M. Smith, Anthony W. Segal, and Jeffrey C. Barrett. 2015. "Characterization of Expression Quantitative Trait Loci in the Human Colon." *Inflammatory Bowel Diseases* 21 (2): 251–56.

Smemo, Scott, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gómez-Marín, Ivy Aneas, et al. 2014. "Obesity-Associated Variants within FTO Form Long-Range Functional Connections with IRX3." *Nature* 507 (7492): 371–75.

Smith, Brenden W., Sarah S. Rozelle, Amy Leung, Jessalyn Ubellacker, Ashley Parks, Shirley K. Nah, Deborah French, et al. 2013. "The Aryl Hydrocarbon Receptor Directs Hematopoietic Progenitor Cell Expansion and Differentiation." *Blood* 122 (3): 376–85.

Soldner, Frank, Yonatan Stelzer, Chikdu S. Shivalila, Brian J. Abraham, Jeanne C. Latourelle, M. Inmaculada Barrasa, Johanna Goldmann, Richard H. Myers, Richard A. Young, and Rudolf Jaenisch. 2016. "Parkinson-Associated Risk Variant in Distal Enhancer of α-

Synuclein Modulates Target Gene Expression." *Nature* 533 (7601): 95–99.

Spain, Sarah L., and Jeffrey C. Barrett. 2015. "Strategies for Fine-Mapping Complex Traits." *Human Molecular Genetics* 24 (R1): R111–19.

Speed, Doug, Na Cai, UCLEB Consortium, Michael R. Johnson, Sergey Nejentsev, and David J. Balding. 2017. "Reevaluation of SNP Heritability in Complex Human Traits." *Nature Genetics* 49 (7): 986–92.

Spilker, Christina, Spilker Christina, and Michael R. Kreutz. 2010. "RapGAPs in Brain: Multipurpose Players in Neuronal Rap Signalling." *The European Journal of Neuroscience* 32 (1): 1–9.

Spitz, François. 2016. "Gene Regulation at a Distance: From Remote Enhancers to 3D Regulatory Ensembles." *Seminars in Cell & Developmental Biology* 57: 57–67.

Stark, Amy L., Ronald J. Hause Jr, Lidija K. Gorsic, Nirav N. Antao, Shan S. Wong, Sophie H. Chung, Daniel F. Gill, et al. 2014. "Protein Quantitative Trait Loci Identify Novel Candidates Modulating Cellular Response to Chemotherapy." *PLoS Genetics* 10 (4): e1004192.

Stenson, Peter D., Matthew Mort, Edward V. Ball, Katy Shaw, Andrew Phillips, and David N. Cooper. 2014. "The Human Gene Mutation Database: Building a Comprehensive Mutation Repository for Clinical and Molecular Genetics, Diagnostic Testing and Personalized Genomic Medicine." *Human Genetics* 133 (1): 1–9.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. "An Integrated Map of Structural Variation in 2,504 Human Genomes." *Nature* 526 (7571): 75–81.

Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors." *Cell* 126 (4): 663–76.

Tan, Minjia, Hao Luo, Sangkyu Lee, Fulai Jin, Jeong Soo Yang, Emilie Montellier, Thierry Buchou, et al. 2011. "Identification of 67 Histone Marks and Histone Lysine Crotonylation as a New Type of Histone Modification." *Cell* 146 (6): 1016–28.

Teng, Mingxiang, Michael I. Love, Carrie A. Davis, Sarah Djebali, Alexander Dobin, Brenton R. Graveley, Sheng Li, et al. 2016. "A Benchmark for RNA-Seq Quantification Pipelines." *Genome Biology* 17 (April): 74.

Tewhey, Ryan, Dylan Kotliar, Daniel S. Park, Brandon Liu, Sarah Winnicki, Steven K. Reilly, Kristian G. Andersen, et al. 2016. "Direct Identification of Hundreds of Expression-Modulating Variants Using a Multiplexed Reporter Assay." *Cell* 165 (6): 1519–29.

Thurner, Matthias, Martijn van de Bunt, Kyle Gaulton, Amy Barrett, Amanda J. Bennett, Jason M. Torres, Vibe Nylander, et al. 2017. "Integration of Human Pancreatic Islet Genomic Data Refines Regulatory Mechanisms at Type 2 Diabetes Susceptibility Loci." doi:10.1101/190892.

Trynka, Gosia, Cynthia Sandor, Buhm Han, Han Xu, Barbara E. Stranger, X. Shirley Liu, and Soumya Raychaudhuri. 2013. "Chromatin Marks Identify Critical Cell Types for Fine Mapping Complex Trait Variants." *Nature Genetics* 45 (2): 124–30.

Trynka, Gosia, Harm-Jan Westra, Kamil Slowikowski, Xinli Hu, Han Xu, Barbara E. Stranger, Robert J. Klein, Buhm Han, and Soumya Raychaudhuri. 2015. "Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-Coding Variants within Complex-Trait Loci." *American Journal of Human Genetics* 97 (1): 139–52.

Turner, Adam W., Amy Martinuk, Anada Silva, Paulina Lau, Majid Nikpay, Per Eriksson, Lasse Folkersen, et al. 2016. "Functional Analysis of a Novel Genome-Wide Association Study Signal in SMAD3 That Confers Protection From Coronary Artery Disease." *Arteriosclerosis, Thrombosis, and Vascular Biology* 36 (5): 972–83.

Vaquerizas, Juan M., Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. 2009. “A Census of Human Transcription Factors: Function, Expression and Evolution.” *Nature Reviews. Genetics* 10 (4): 252–63.

Veerman, Christiaan C., Georgios Kosmidis, Christine L. Mummery, Simona Casini, Arie O. Verkerk, and Milena Bellin. 2015. “Immaturity of Human Stem-Cell-Derived Cardiomyocytes in Culture: Fatal Flaw or Soluble Problem?” *Stem Cells and Development* 24 (9): 1035–52.

Wainger, Brian J., Evangelos Kiskinis, Cassidy Mellin, Ole Wiskow, Steve S. W. Han, Jackson Sandoe, Numa P. Perez, et al. 2014. “Intrinsic Membrane Hyperexcitability of Amyotrophic Lateral Sclerosis Patient-Derived Motor Neurons.” *Cell Reports* 7 (1): 1–11.

Wakefield, Jon. 2009. “Bayes Factors for Genome-Wide Association Studies: Comparison with P-Values.” *Genetic Epidemiology* 33 (1): 79–86.

Wallace, Chris, Antony J. Cutler, Nikolas Pontikos, Marcin L. Pekalski, Oliver S. Burren, Jason D. Cooper, Arcadio Rubio García, et al. 2015. “Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping.” *PLoS Genetics* 11 (6): e1005272.

Wang, Xinchen, Liang He, Sarah Goggin, Alham Saadat, Li Wang, Melina Claussnitzer, and Manolis Kellis. 2017. “High-Resolution Genome-Wide Functional Dissection of Transcriptional Regulatory Regions in Human.” *bioRxiv*. doi:10.1101/193136.

Ward, Lucas D., and Manolis Kellis. 2012. “HaploReg: A Resource for Exploring Chromatin States, Conservation, and Regulatory Motif Alterations within Sets of Genetically Linked Variants.” *Nucleic Acids Research* 40 (Database issue): D930–34.

Warren, Curtis R., Cashell E. Jaquish, and Chad A. Cowan. 2017. “The NextGen Genetic Association Studies Consortium: A Foray into In Vitro Population Genetics.” *Cell Stem Cell* 20 (4): 431–33.

Warren, Curtis R., John F. O’Sullivan, Max Friesen, Caroline E. Becker, Xiaoling Zhang, Poching Liu, Yoshiyuki Wakabayashi, et al. 2017. “Induced Pluripotent Stem Cell Differentiation Enables Functional Validation of GWAS Variants in Metabolic Disease.” *Cell Stem Cell* 20 (4): 547–57.e7.

Wellcome Trust Case Control Consortium. 2007. “Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls.” *Nature* 447 (7145): 661–78.

Wellcome Trust Case Control Consortium, Julian B. Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, et al. 2012. “Bayesian Refinement of Association Signals for 14 Loci in 3 Common Diseases.” *Nature Genetics* 44 (12): 1294–1301.

Wernig, Marius, Jian-Ping Zhao, Jan Pruszak, Eva Hedlund, Dongdong Fu, Frank Soldner, Vania Broccoli, Martha Constantine-Paton, Ole Isacson, and Rudolf Jaenisch. 2008. “Neurons Derived from Reprogrammed Fibroblasts Functionally Integrate into the Fetal Brain and Improve Symptoms of Rats with Parkinson’s Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (15): 5856–61.

Westra, Harm-Jan, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, et al. 2013. “Systematic Identification of Trans eQTLs as Putative Drivers of Known Disease Associations.” *Nature Genetics* 45 (10): 1238–43.

Willis, Dianna E., Meng Wang, Elizabeth Brown, Lilah Fones, and John W. Cave. 2016. “Selective Repression of Gene Expression in Neuropathic Pain by the Neuron-Restrictive Silencing Factor/repressor Element-1 Silencing Transcription (NRSF/REST).” *Neuroscience Letters* 625 (June): 20–25.

Yang, Jian, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna A. E. Vinkhuyzen, Sang

Hong Lee, Matthew R. Robinson, et al. 2015. “Genetic Variance Estimation with Imputed Variants Finds Negligible Missing Heritability for Human Height and Body Mass Index.” *Nature Genetics* 47 (10): 1114–20.

Young, Gareth T., Gutteridge Alex, Heather DE Fox, Anna L. Wilbrey, Cao Lishuang, Lily T. Cho, Adam R. Brown, et al. 2014. “Characterizing Human Stem Cell–derived Sensory Neurons at the Single-Cell Level Reveals Their Ion Channel Expression and Utility in Pain Research.” *Molecular Therapy: The Journal of the American Society of Gene Therapy* 22 (8): 1530–43.

Zhang, Xiaoyang, Richard Cowper-Sal lari, Swneke D. Bailey, Jason H. Moore, and Mathieu Lupien. 2012. “Integrative Functional Genomics Identifies an Enhancer Looping to the SOX9 Gene Disrupted by the 17q24.3 Prostate Cancer Risk Locus.” *Genome Research* 22 (8): 1437–46.

Zhao, Hao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. 2014. “CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies.” *Bioinformatics* 30 (7): 1006–7.

Zhou, Jian, and Olga G. Troyanskaya. 2015. “Predicting Effects of Noncoding Variants with Deep Learning–based Sequence Model.” *Nature Methods* 12 (10): 931–34.

Zhou, Pingzhu, Aibin He, and William T. Pu. 2012. “Regulation of GATA4 Transcriptional Activity in Cardiovascular Development and Disease.” *Current Topics in Developmental Biology* 100: 143–69.

Zhu, Zhihong, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R. Robinson, Joseph E. Powell, Grant W. Montgomery, et al. 2016. “Integration of Summary Data from GWAS and eQTL Studies Predicts Complex Trait Gene Targets.” *Nature Genetics* 48 (5): 481–87.