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.

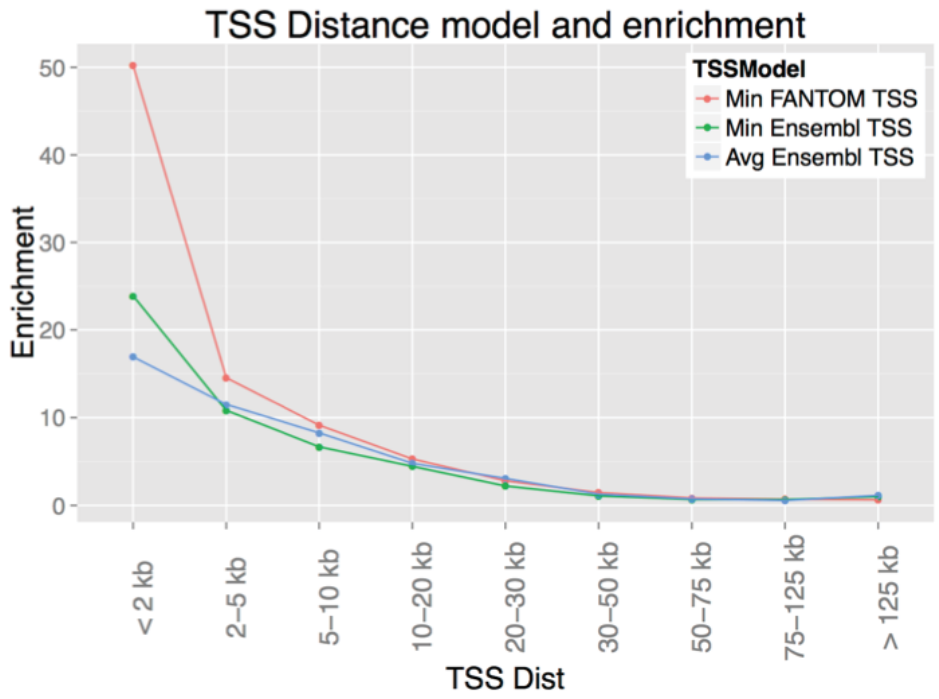
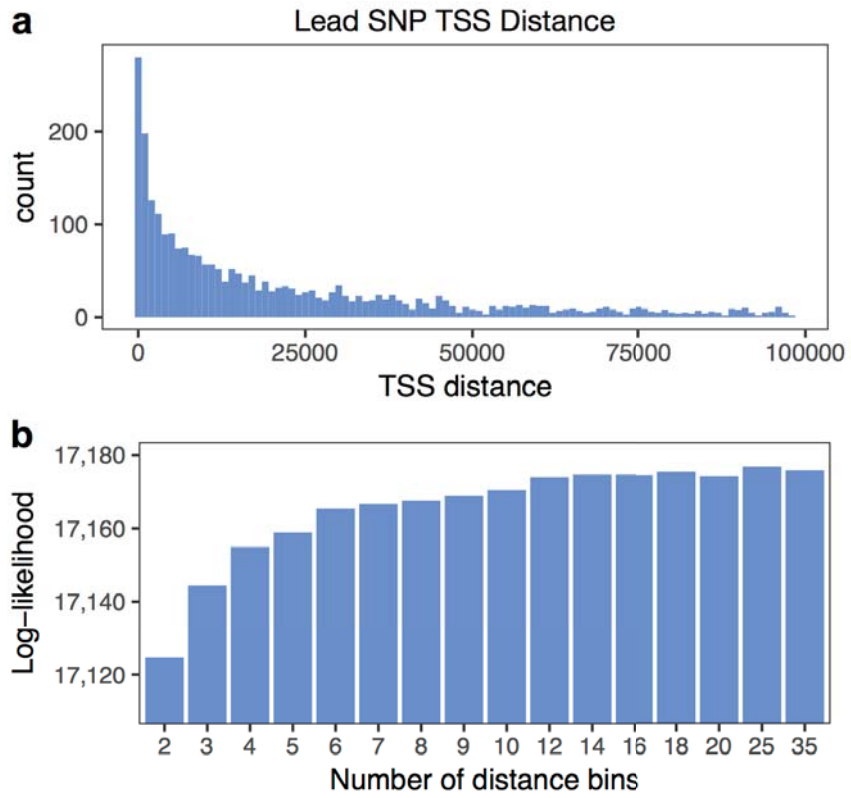


Figure 1: Enrichment of causal SNPs in fixed distance bins for the three TSS distance definitions described in the main text.

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.

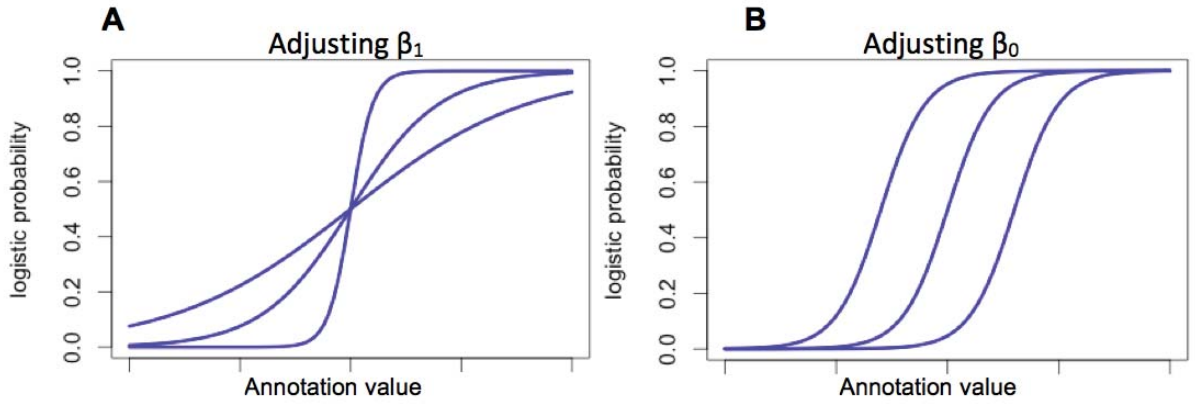


Figure 3: A standard logistic function enables controlling the slope parameter β_1 (**a**) which determines how quickly an annotation becomes informative, as well as the translation parameter β_0 (**b**) which determines at what absolute value the annotation begins to be informative.

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

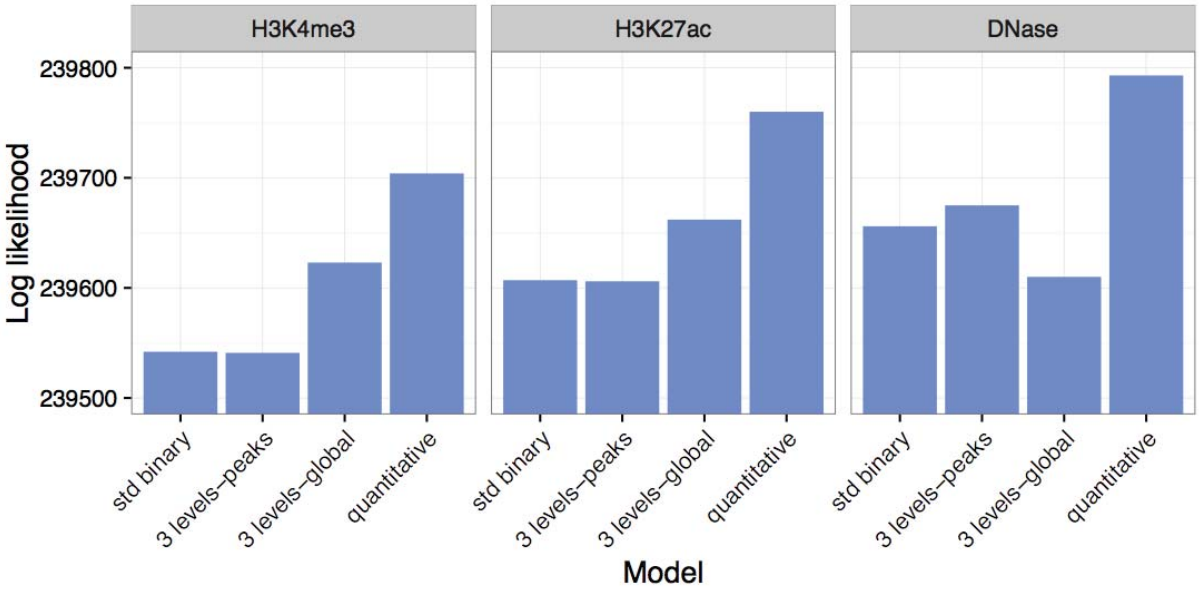


Figure 4: Cross-validation likelihood of quantitative and binary annotation models applied to three different annotations. 12-bin distance annotations were included in all models.

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.

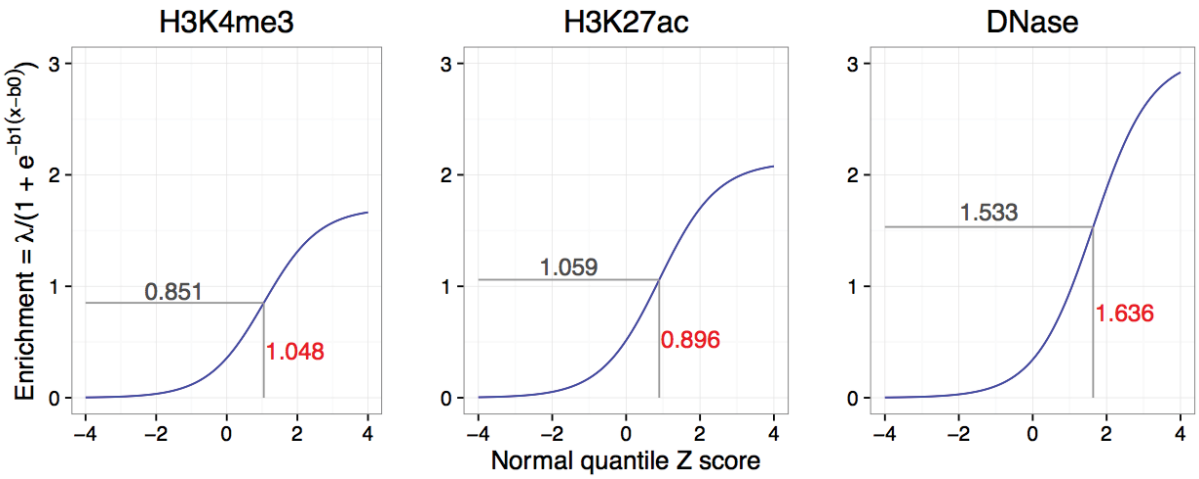


Figure 5: Parameters of the quantitative annotation model for 3 annotations. The x axis represents the normal quantile Z score across all SNPs; e.g. 95% of SNPs have annotation values between of +/- 1.96. The y axis represents the enrichment of SNPs with the highest scores relative to the lowest.

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.

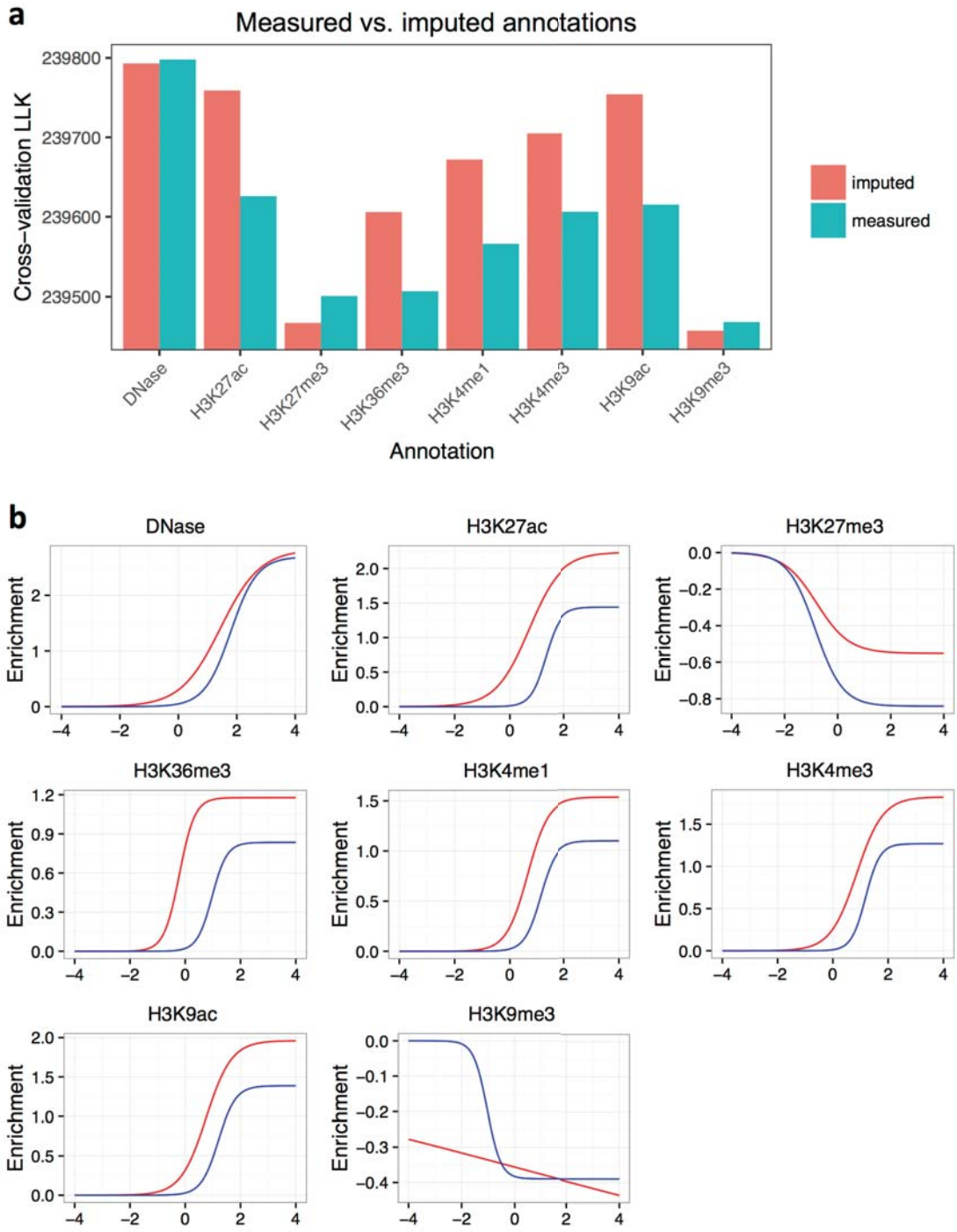


Figure 6: Imputed quantitative annotations from Roadmap Epigenomics outperform measured annotations. (**a**) Model log likelihoods, and (**b**) logistic curves defined by the optimal parameters for the same models; enrichments for imputed annotations are in red, measured annotations in blue. In all cases 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).

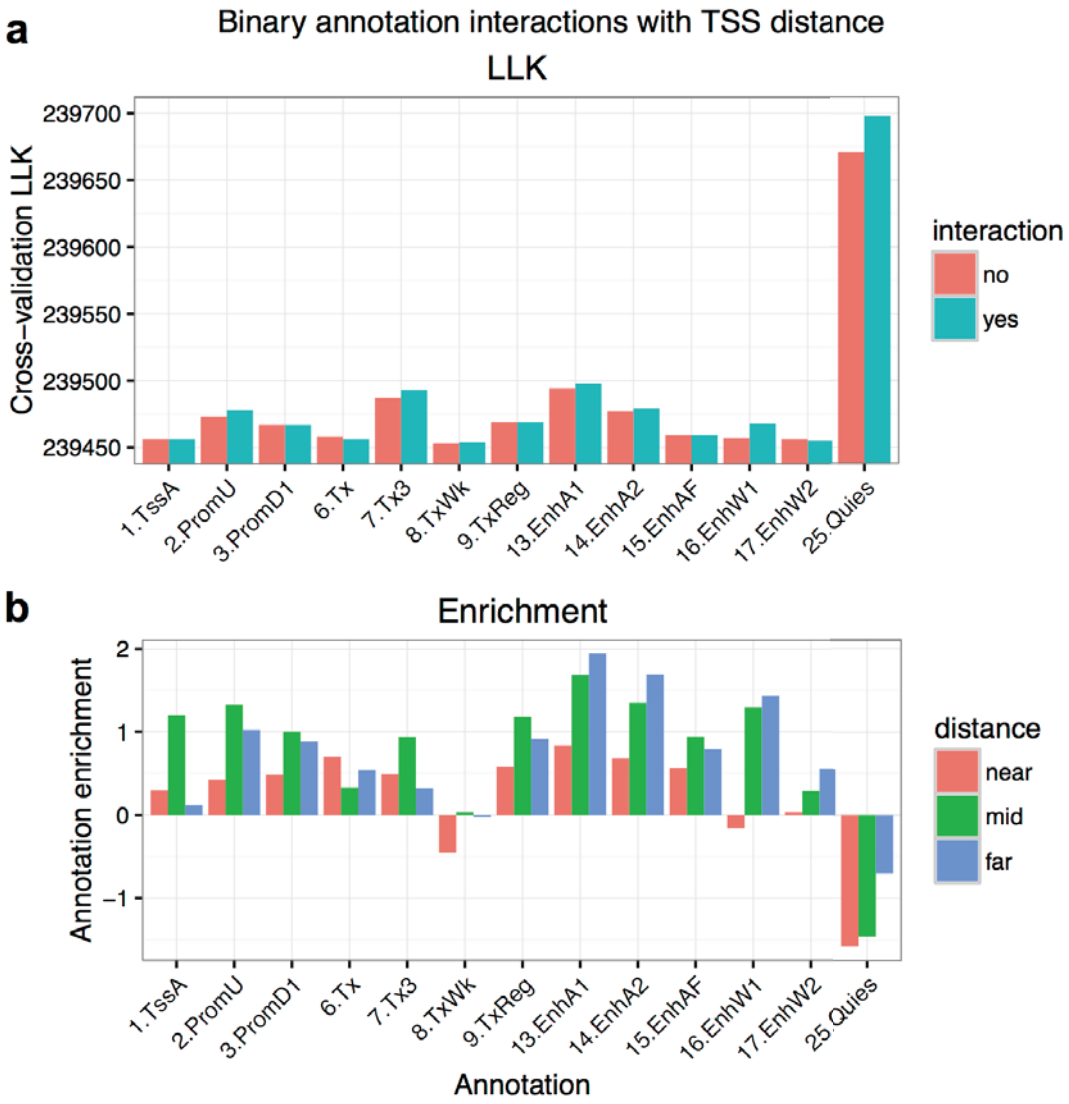


Figure 7: Modeling the interaction between binary annotations and distance to gene TSS slightly improved model performance in cross-validation. Shown are (**a**) model LLKs, and (**b**) the annotation enrichment in near/medium/far distance bins. A distance model is also included, so that the annotation*distance 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).

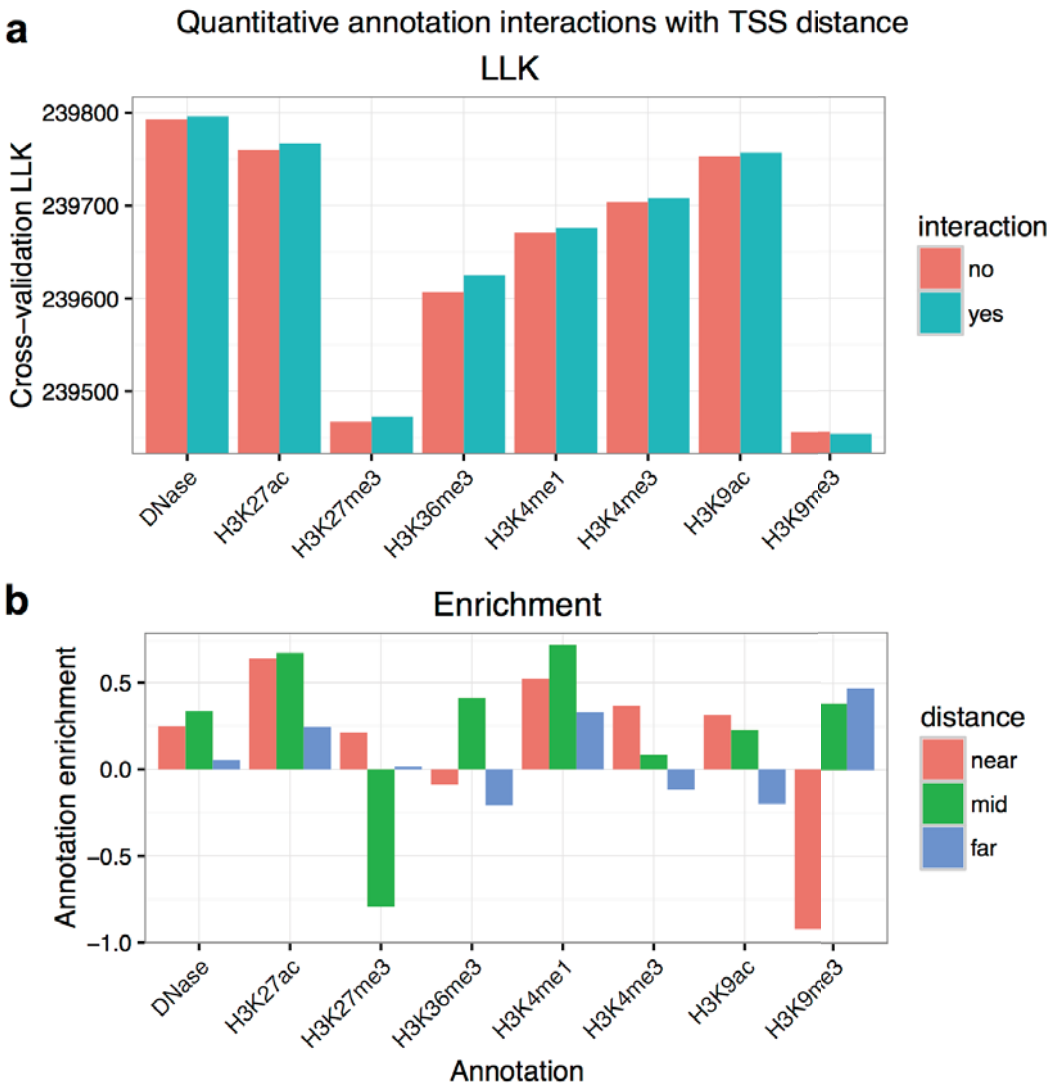


Figure 8: (**a**) For quantitative annotations, modeling the interaction between annotations and distance to gene TSS also slightly improves model performance in cross-validation. (**b**) Enrichments for causal SNPs in peaks for the quantitative annotations in (a), but split into three different distance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

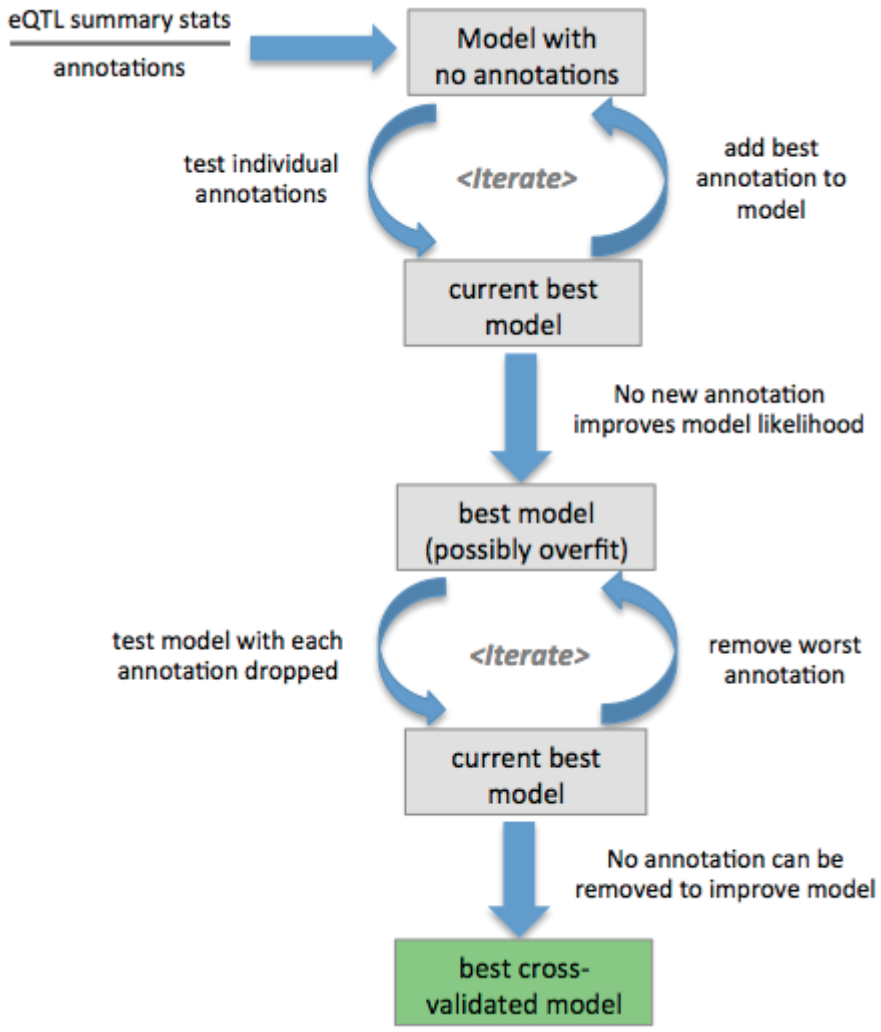


Figure 9: Schematic of model building process: forward stepwise selection to add annotations, followed by removal of annotations using cross-validation.

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.

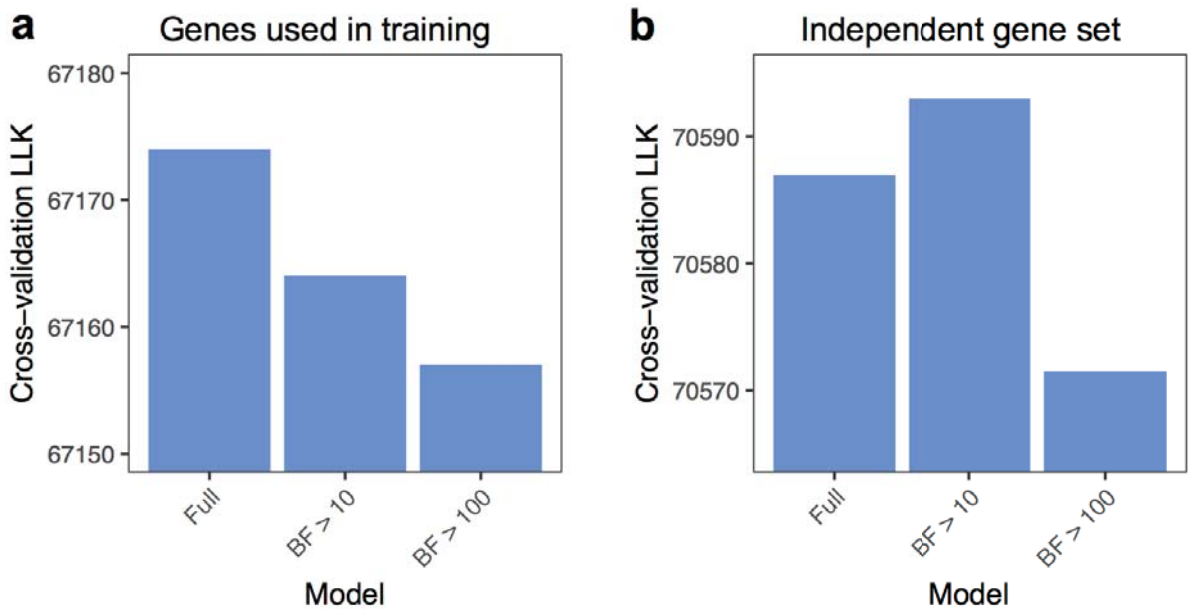


Figure 10: Cross-validation LLK of models shown in Table 4, for either (**a**) the set of 1,000 genes used in training, or (**b**) a separate 1,000 genes.

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).

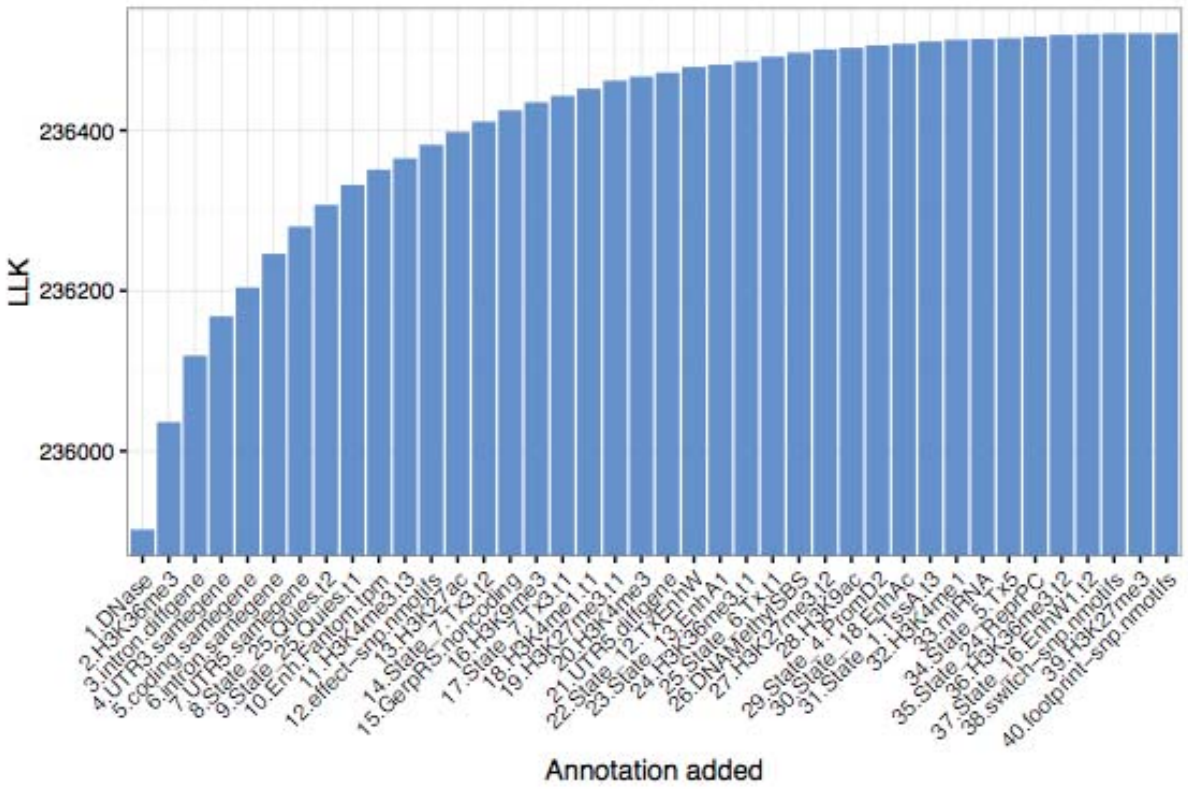


Figure 11: Model LLK with each annotation that was added over 40 iterations. Each model includes 12 distance bins, as well as all the annotations added prior to it.

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.

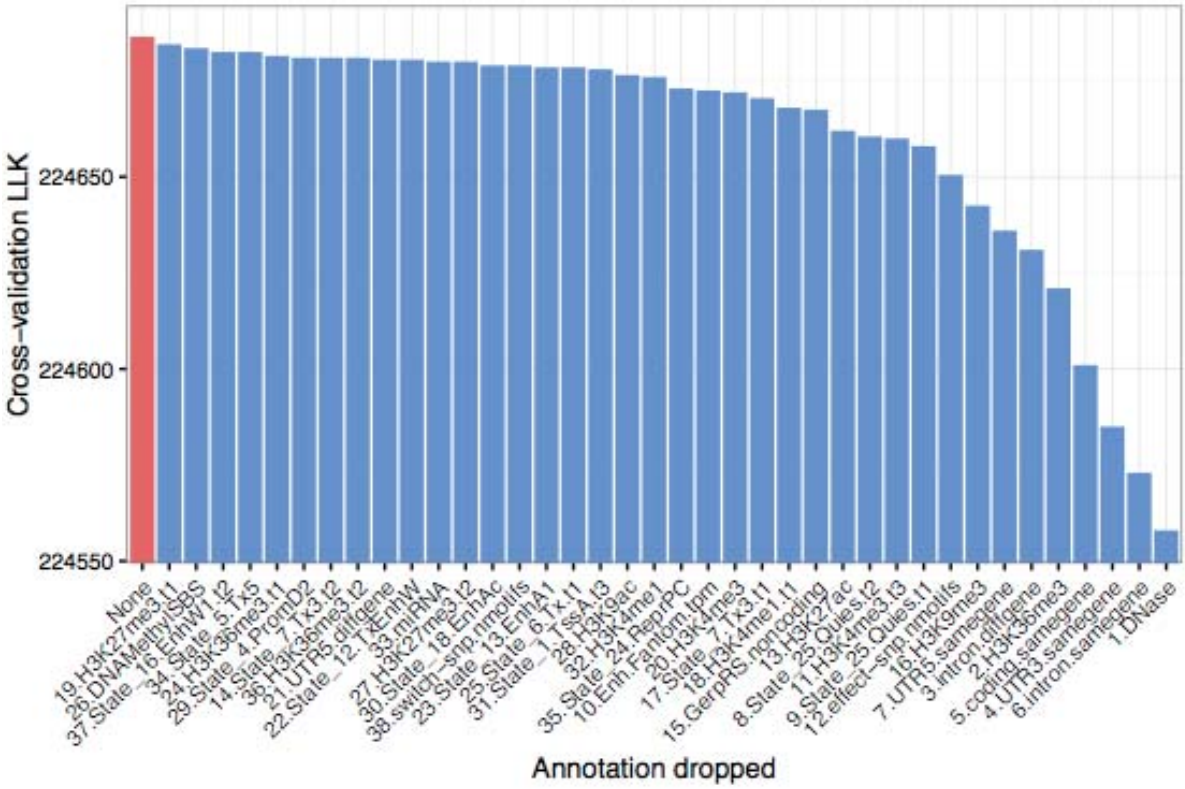


Figure 12: Model cross-validation LLK when each annotation is individually dropped. The full model with no annotations dropped is on the left, and no annotation can be dropped without reducing the LLK. For the x-axis labels, the number preceding each annotation name is the order in which it was added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

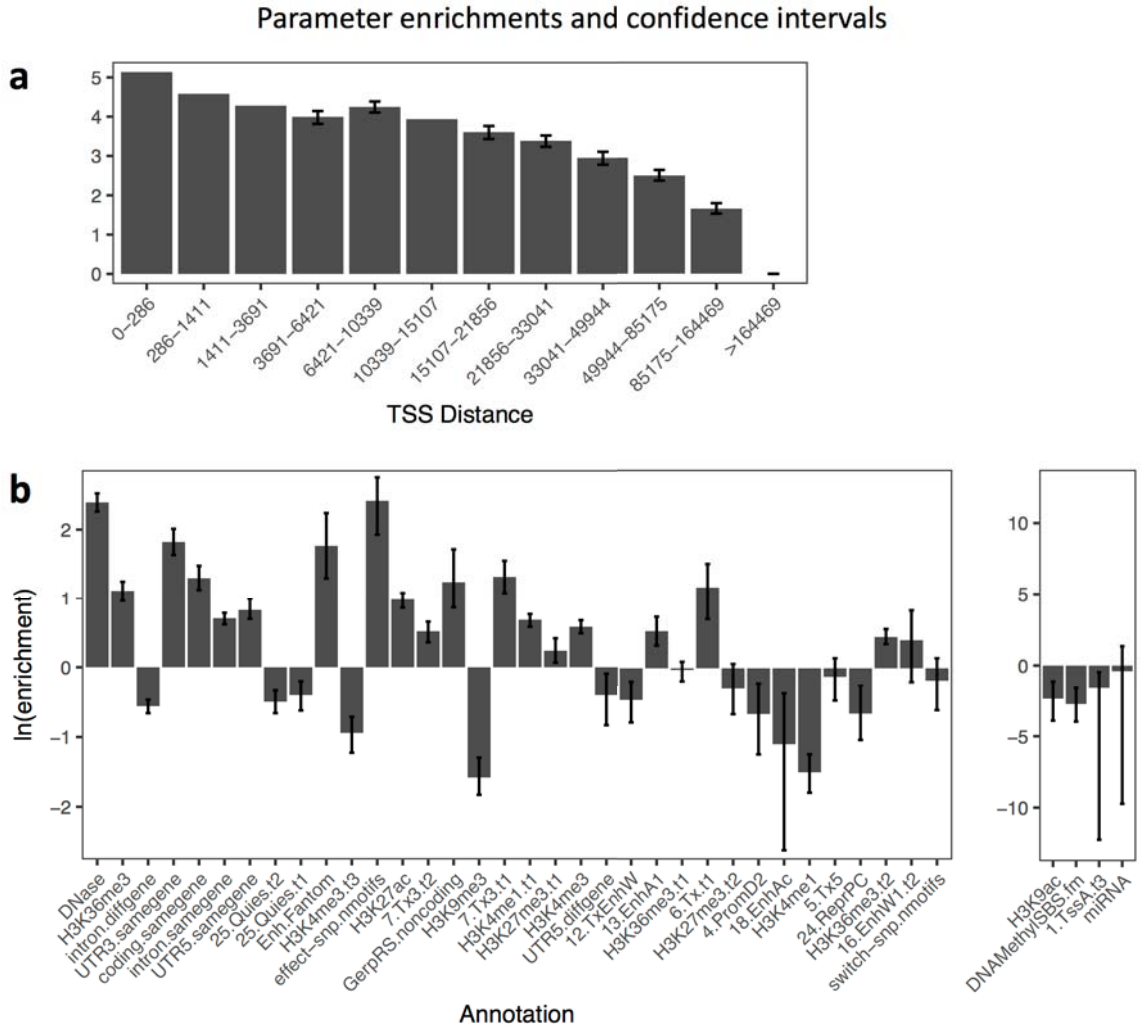


Figure 13: Enrichments for (**a**) TSS distance annotations, and (**b**) the 38 other annotations included in the final model. Confidence intervals were determined by individually adjusting annotation enrichments until the model LLK is decreased by 2 units. Fgwas failed when computing confidence intervals 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.

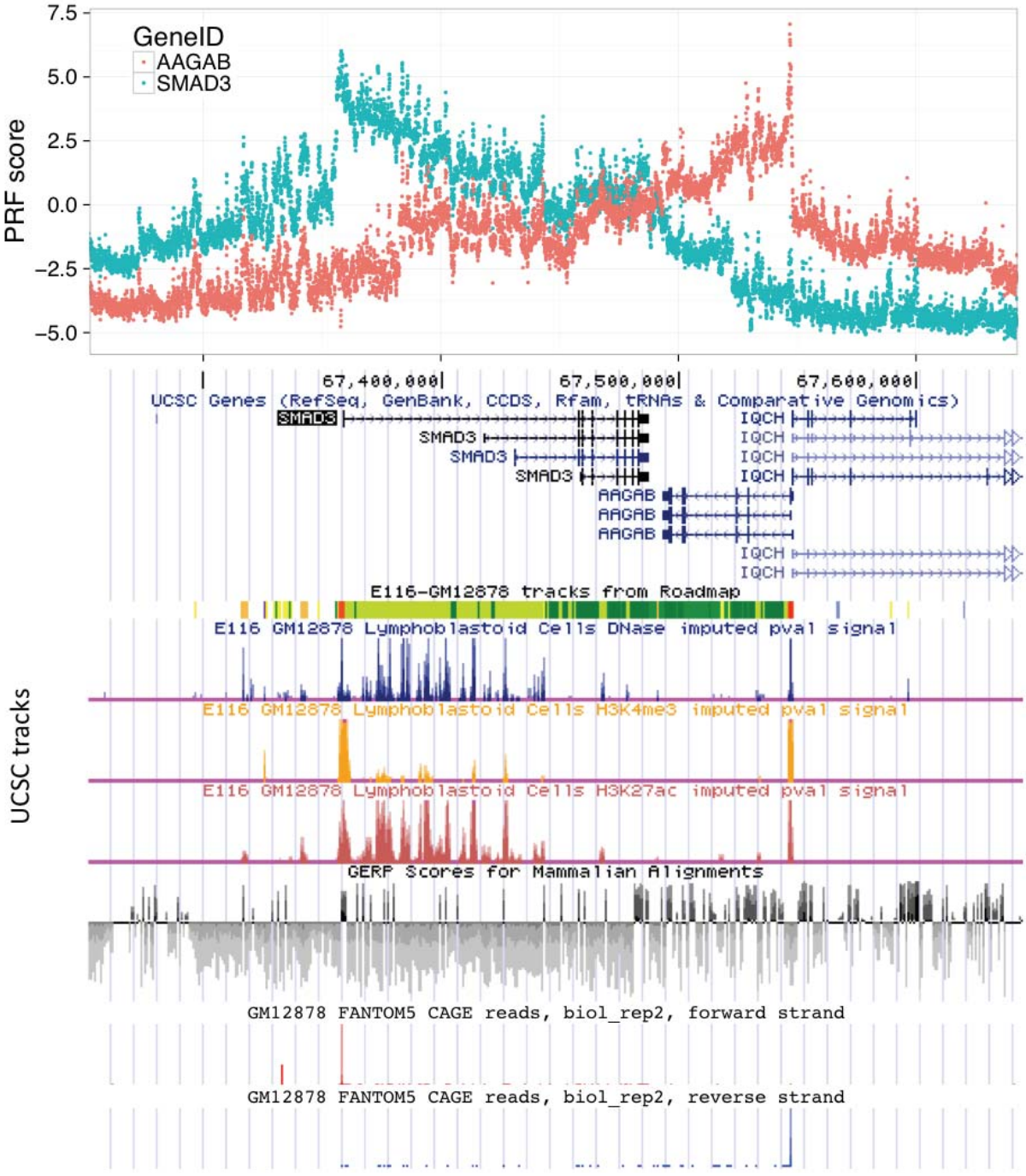


Figure 14: PRF scores for positions in a 400 kb window around *SMAD3*. Scores for two genes are shown, but multiple other genes within 1 Mb also have scores in the region. PRF scores peak towards the TSS of genes. Here, only two TSSes are used, which are visible in the FANTOM5 CAGE reads track.

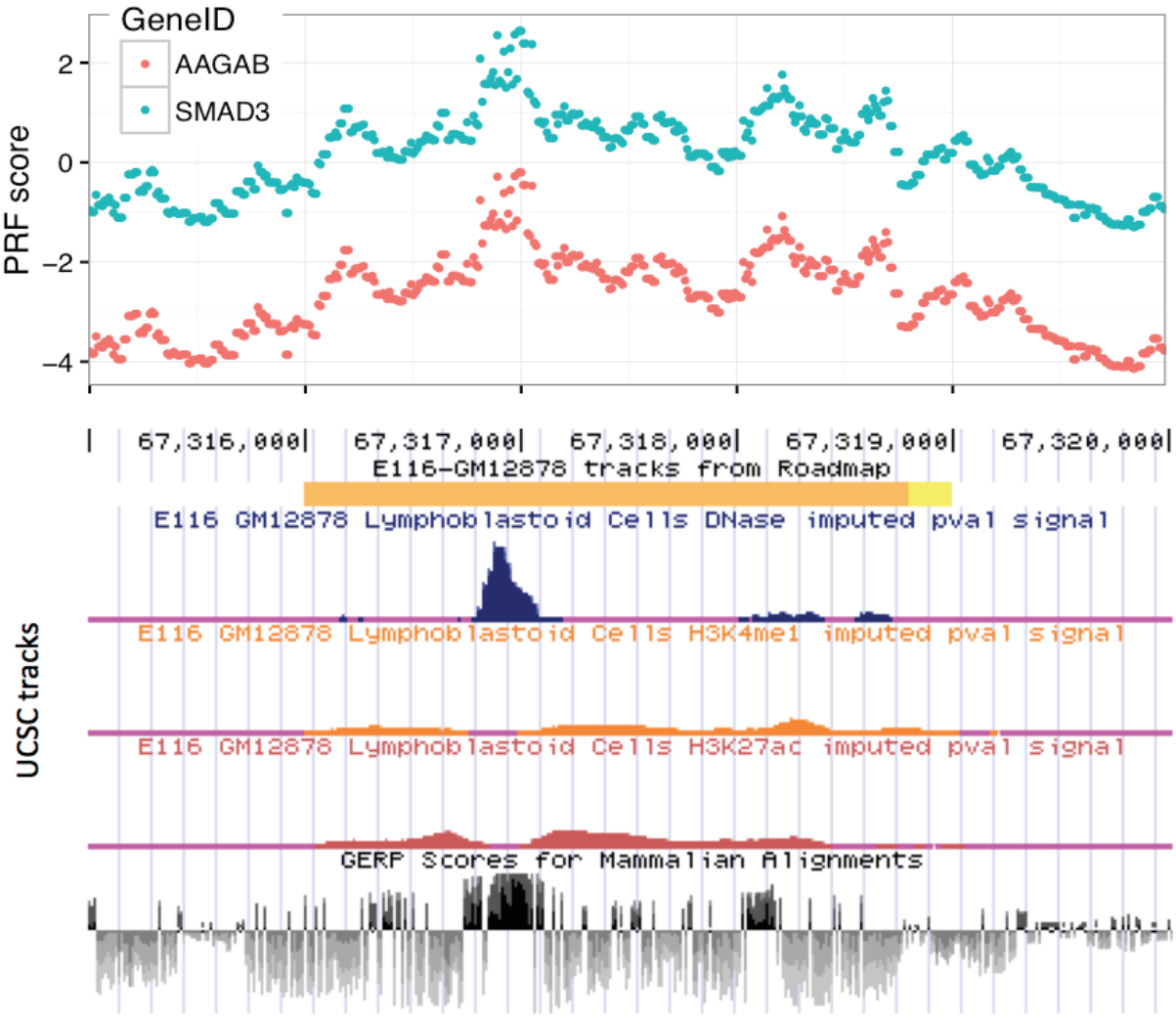


Figure 15: PRF scores in a 5 kb region upstream of *SMAD3*.

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.

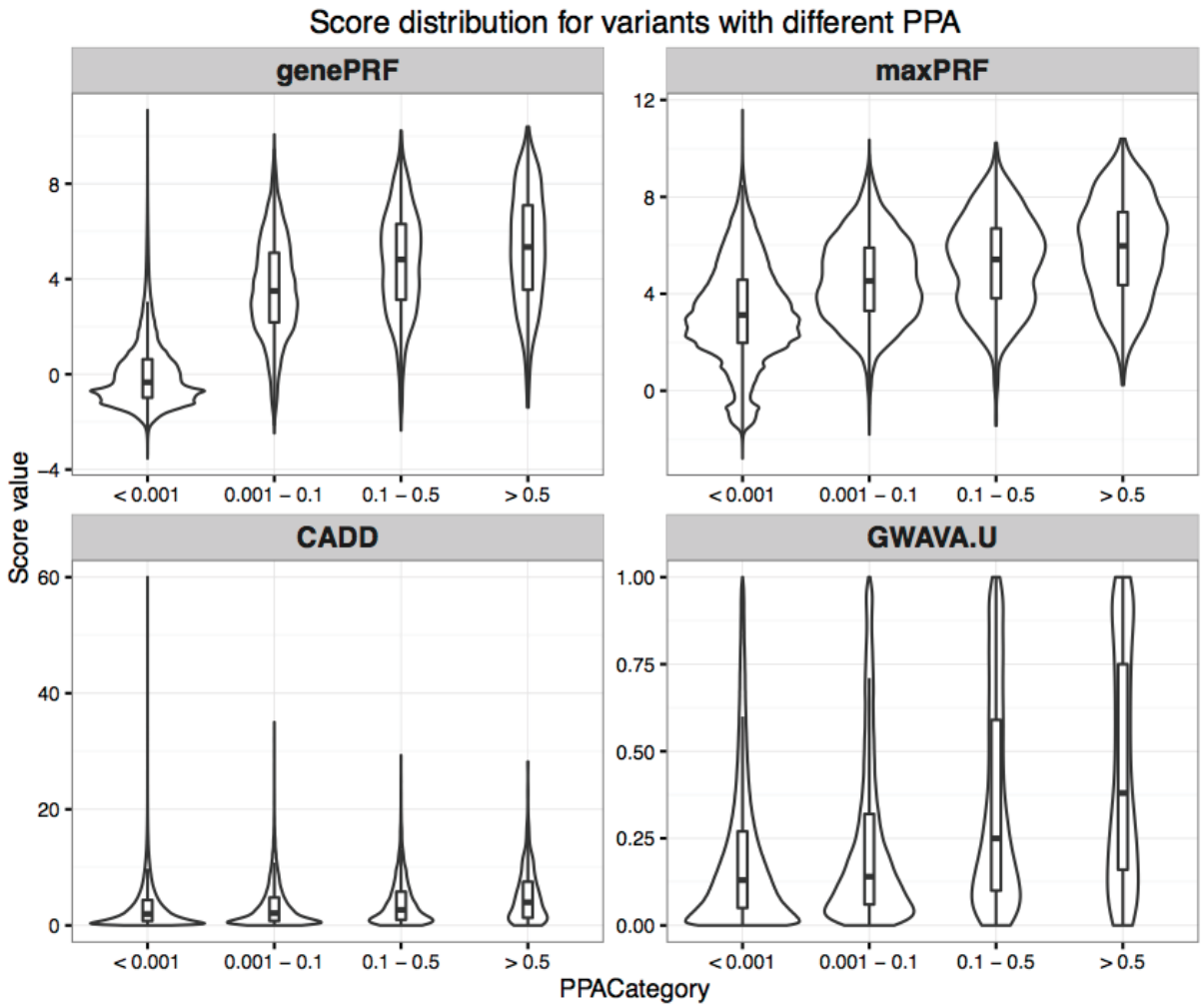


Figure 16: Scores for variants tested for association with gene expression in GTEx subcutaneous adipose tissue, stratified by posterior probability of association, for 4 different scoring methods: genePRF, maxPRF, CADD, and GWAVA.U. PRF scores were calculated using the E063 adipose nuclei epigenome. 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.

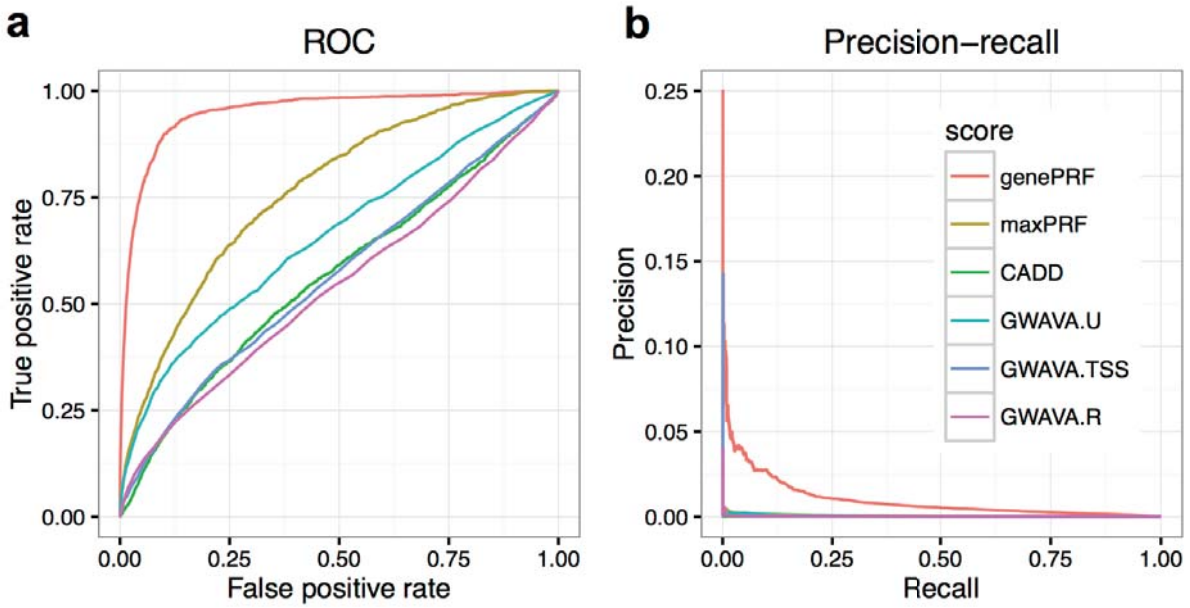


Figure 17: (**a**) ROC curves, and (**b**) precision-recall curves for identifying lead variants (with $p < 10^{-12}$) for eGenes in GTEx subcutaneous adipose tissue from among all variants within 1 Mb.

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.

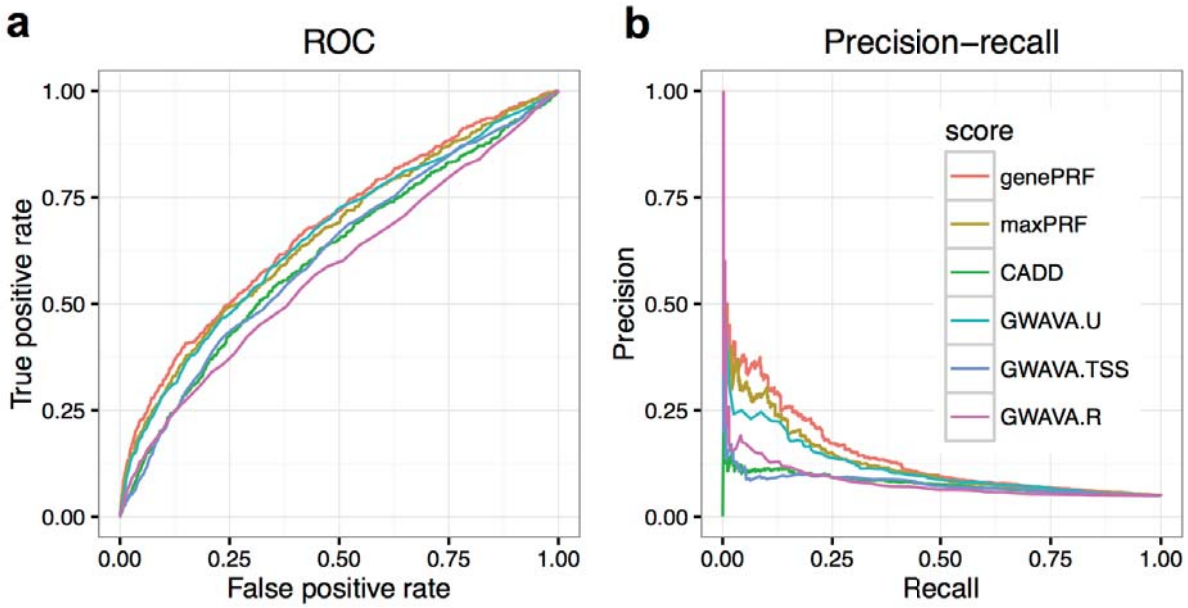


Figure 18: (**a**) ROC curves, and (**b**) precision-recall curves for identifying lead variants in GTEx subcutaneous adipose tissue for "confident eGenes" where the lead variant has a statistical PPA > 0.5. To represent a fine-mapping scenario, only the top 20 variants by statistical association are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

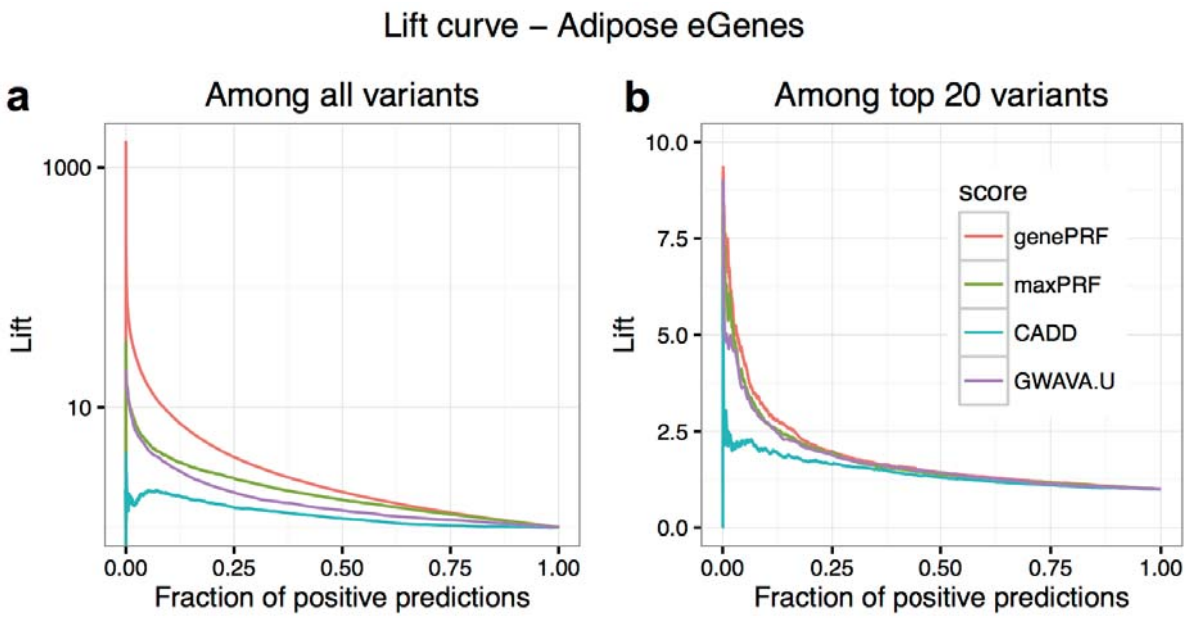


Figure 19: Lift curves for identifying lead eQTL variants in GTEx subcutaneous adipose from among (**a**) all variants in the 1 Mb cis-window or (**b**) the top 20 statistically associated variants. In panel (**a**) the lift values are plotted on a logged axis because they vary over orders of magnitude.

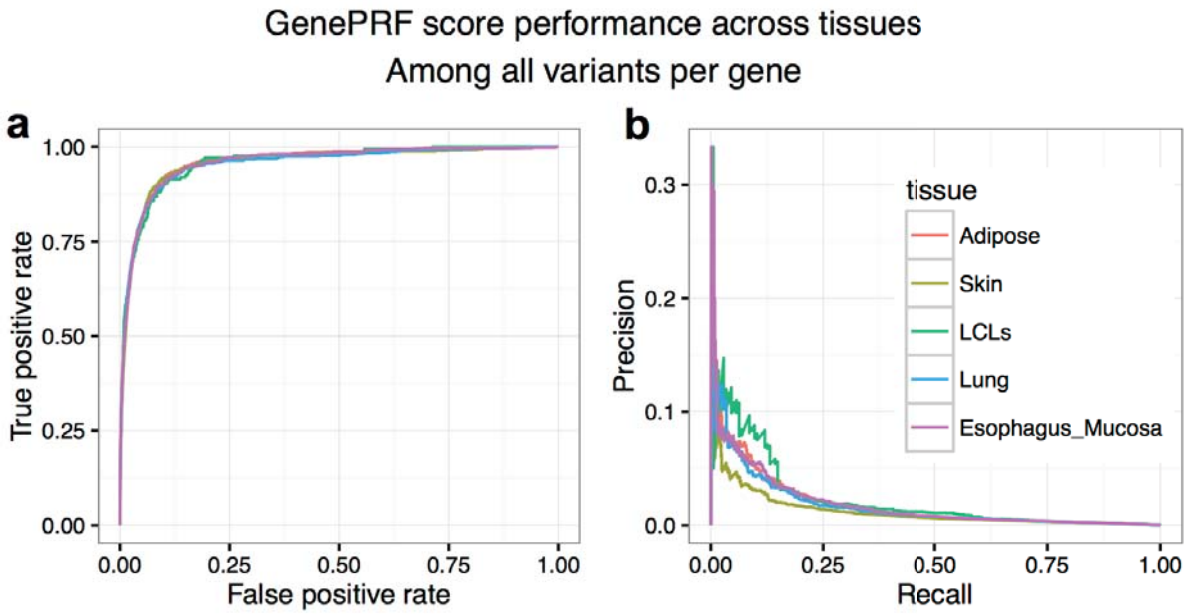


Figure 20: (**a**) ROC curves, and (**b**) precision-recall curves for genePRF scores in identifying lead variants in GTEx tissues from among all variants within 1 Mb. PRF scores were computed for each GTEx tissue using the most relevant epigenome. Performance on different GTEx tissues is indicated by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

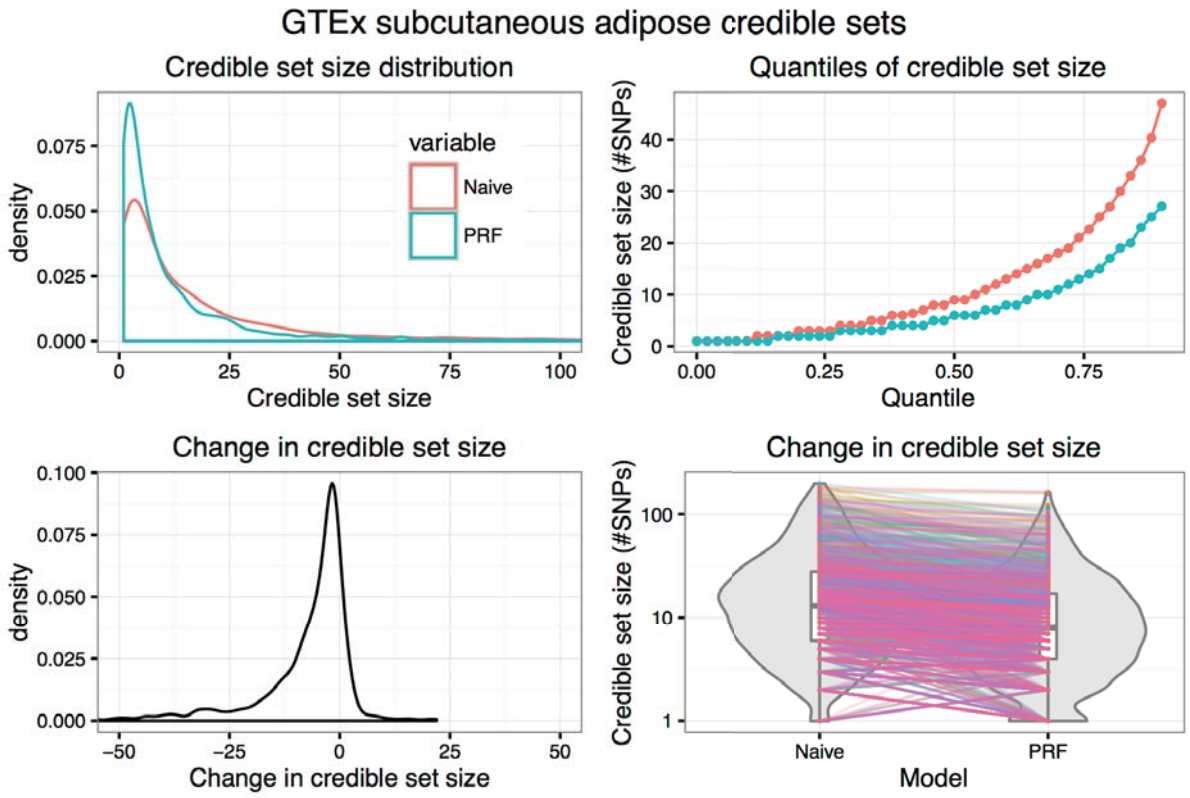


Figure 21: Sizes of 95% credible sets determined using either statistical information only, or combining statistical information with PRF scores. Each of the plots presents a different view showing that credible set sizes are reduced on average when PRF scores are incorporated.

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).

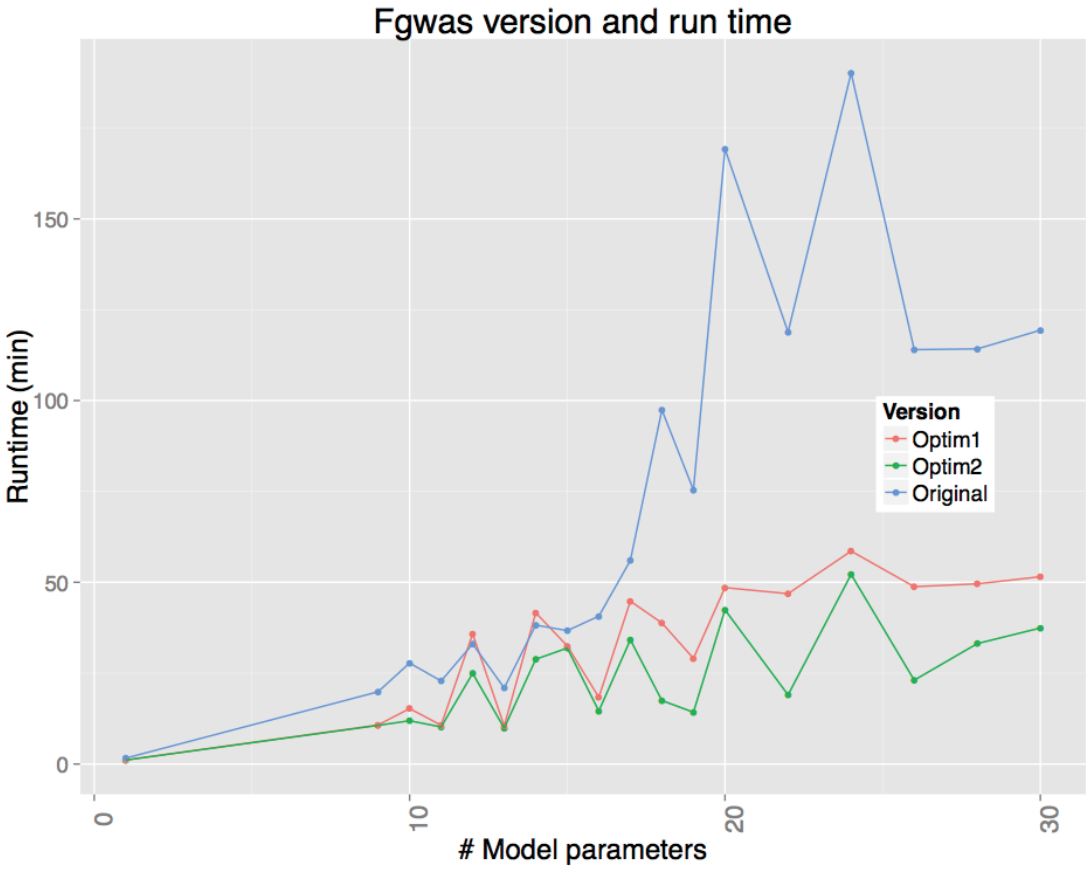


Figure M1: Fgwas running time for models with different numbers of parameters after addition of a new stopping criterion (Optim1) or code optimisations based on profiling as well as the stopping criterion (Optim2).

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.

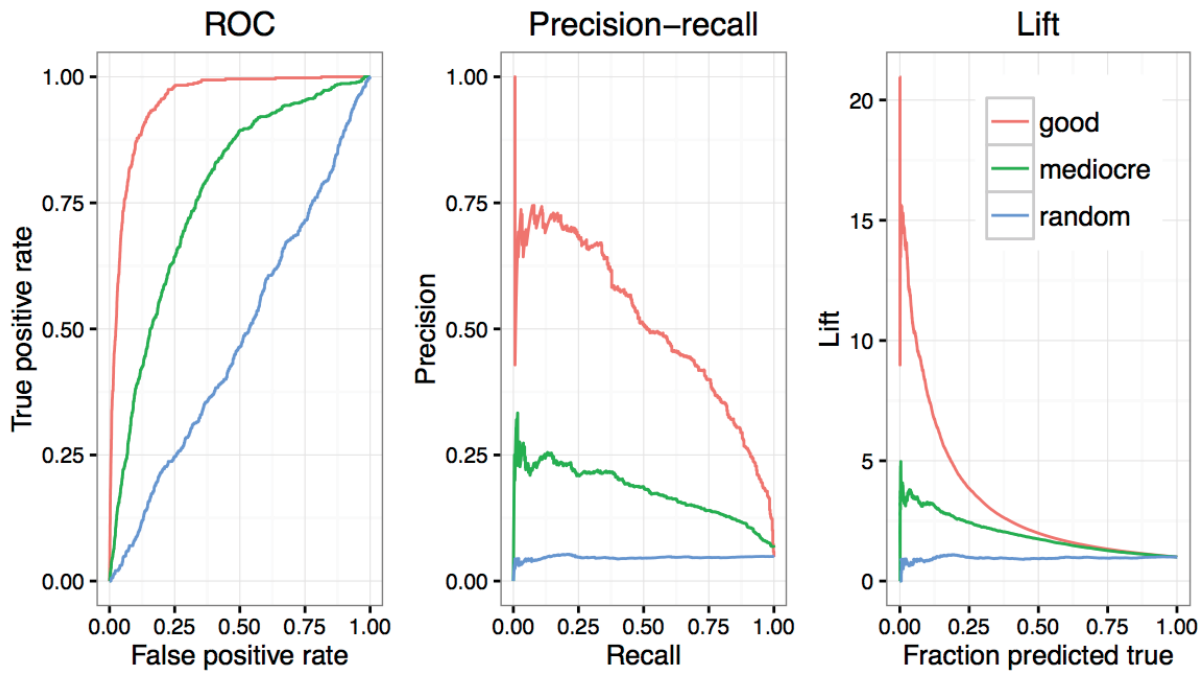


Figure A1: Performance metrics for three classifiers based on simulated data, which are labeled as good (red line), mediocre (green line), or random (blue line). The random classifier has random prediction with respect to the true classification. Shown are (**a**) ROC curves, (**b**) precision-recall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