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.

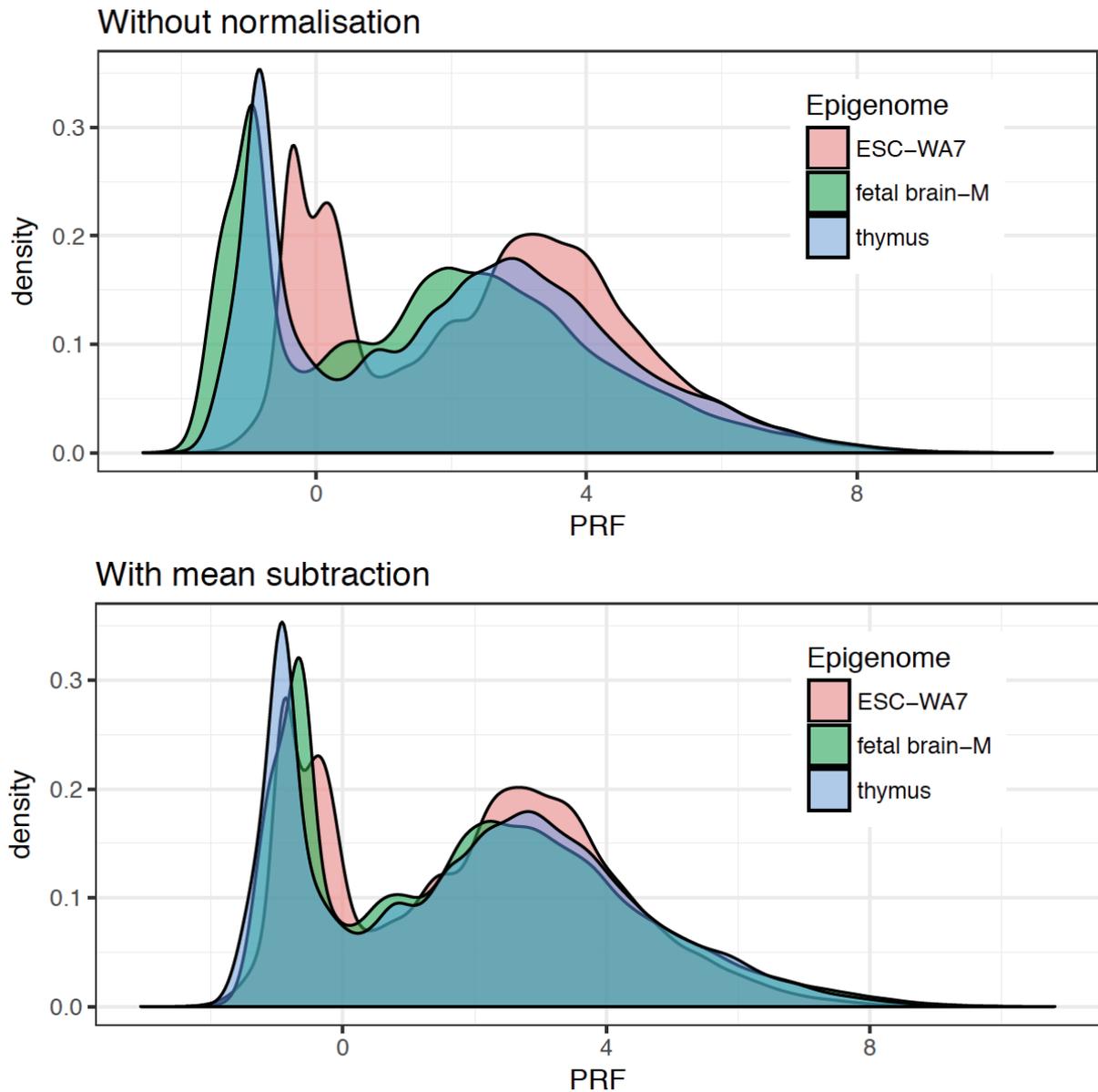## 4.2.1 Average PRF scores differ across epigenomes

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.
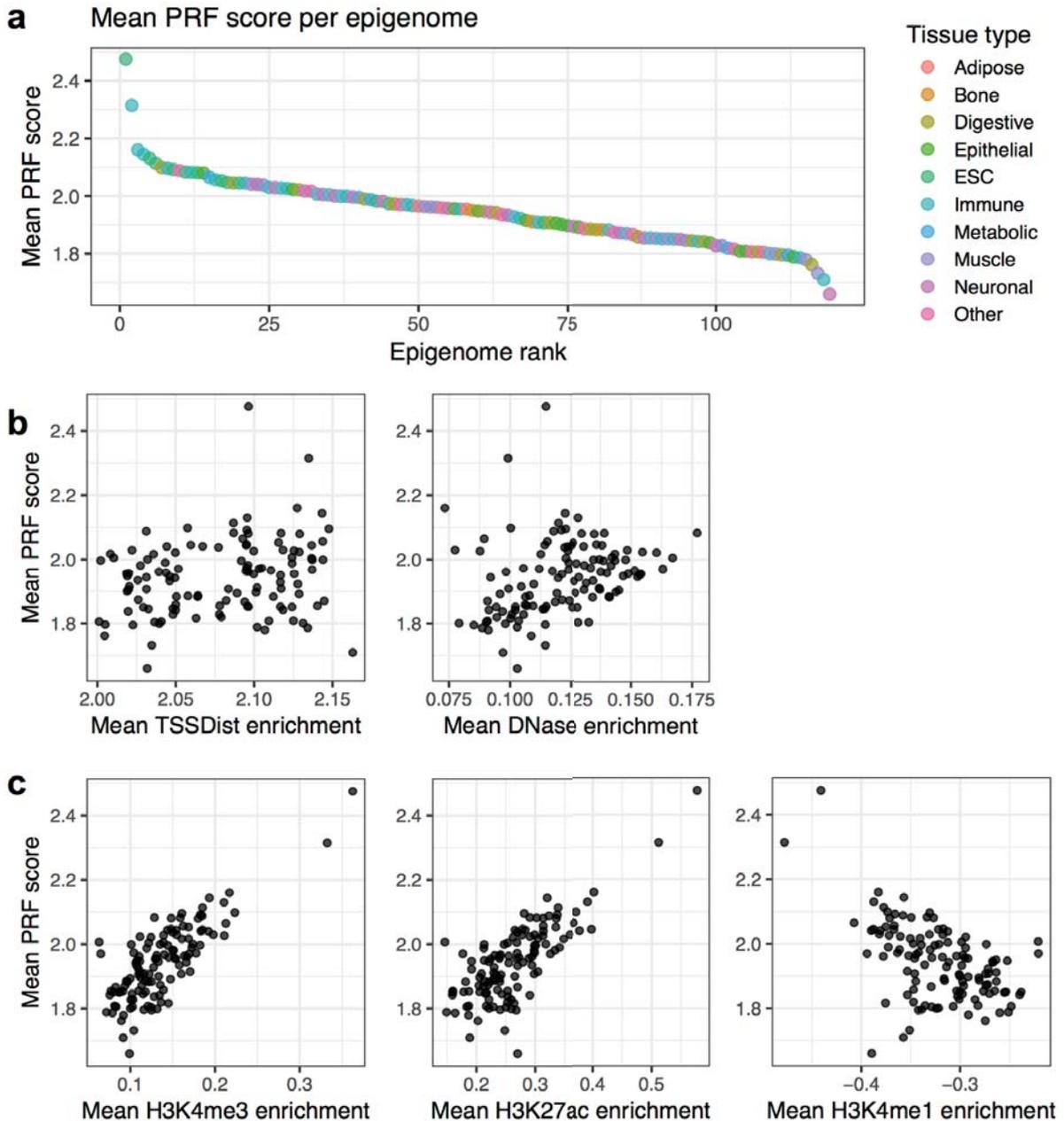
Figure 1: Global differences between epigenomes. (**a**) Mean PRF score for each epigenome across ~288,000 evenly spaced genomic positions. (**b**) Mean TSSDist and DNase hypersensitivity enrichments are only weakly correlated with epigenome mean PRF scores. (**c**) Moderate to strong correlations were found between mean PRF scores and enrichments due to quantitative histone modification annotations H3K4me3 (Pearson $r^2=0.62$, $p<2\times10^{-16}$), H3K27ac ($r^2=0.58$, $p<2\times10^{-16}$), H3K4me1 ($r^2=0.25$, $p=3\times10^{-9}$) and H3K36me3 (not shown; $r^2=0.41$, $p=4\times10^{-15}$).

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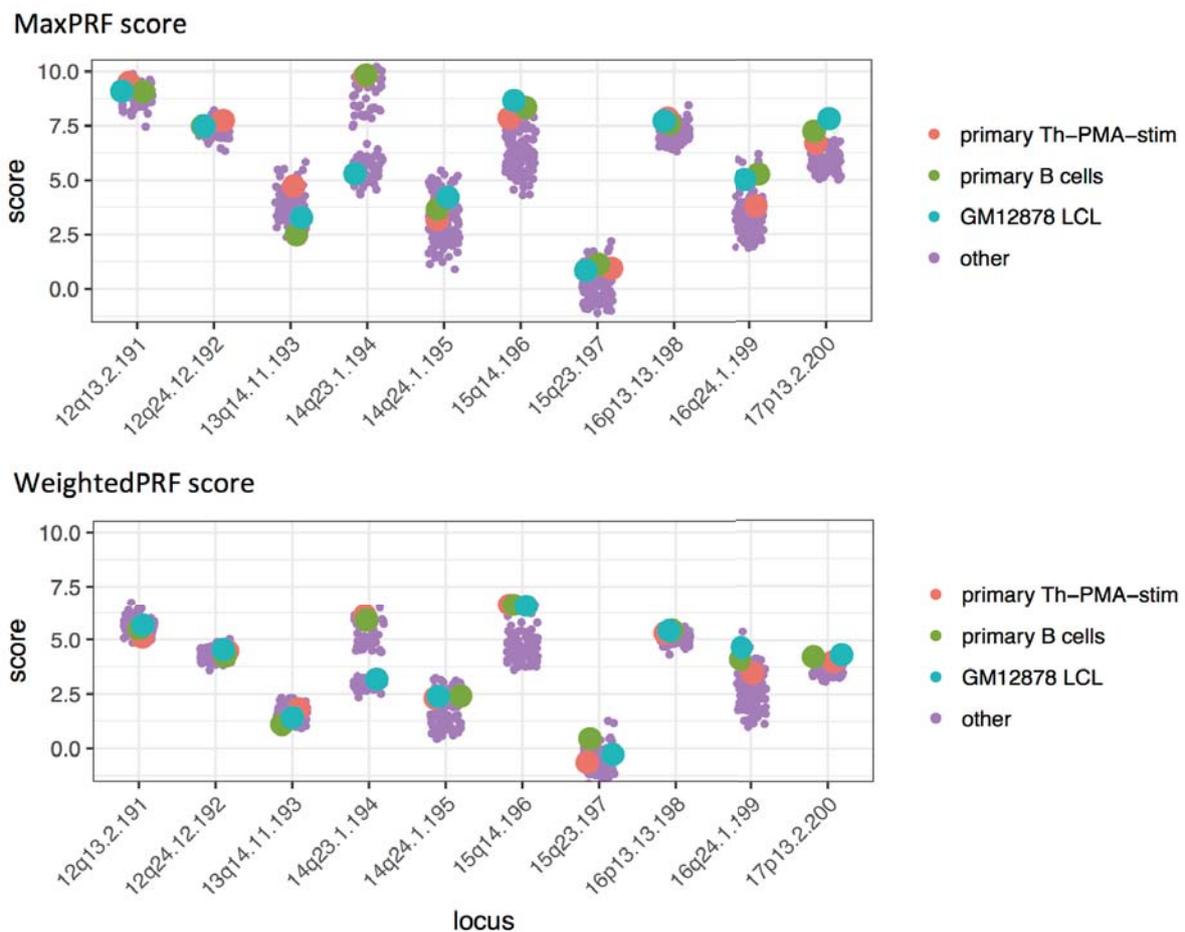
Figure 3: PRF scores across epigenomes for ten loci of an RA GWAS (Okada et al. 2014). Loci are labelled on the x axis. All 119 epigenomes are shown as points for each locus, but three epigenomes are highlighted: PMA-stimulated T-helper cells, primary B cells from peripheral blood, and GM12878 LCLs.

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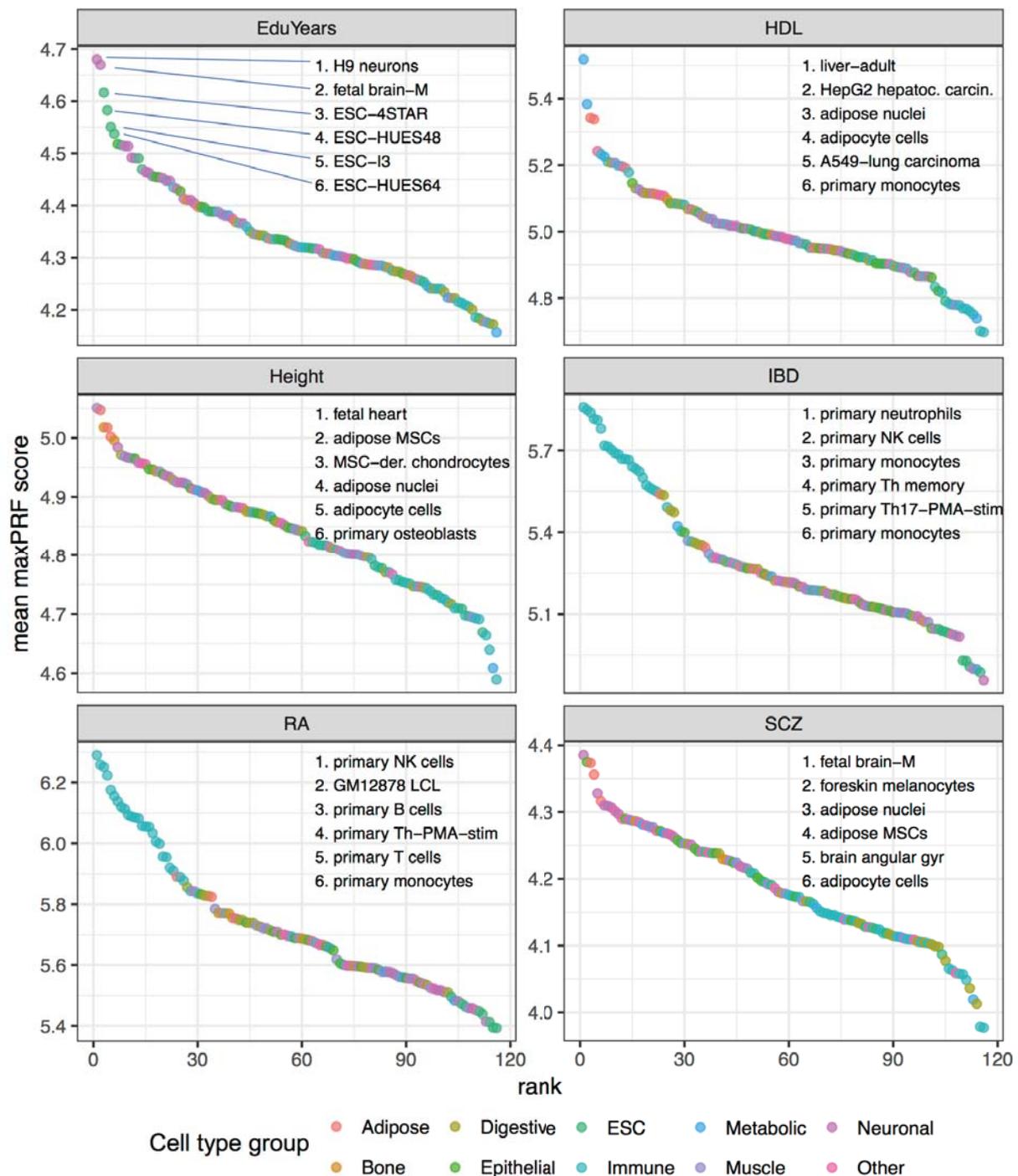
Figure 4: Ranking of epigenomes using the mean normalised maxPRF score across associated loci for six GWAS. The top six epigenomes are labeled for each GWAS, and epigenomes are ordered based on their mean scores. Points are coloured by epigenome tissue type.

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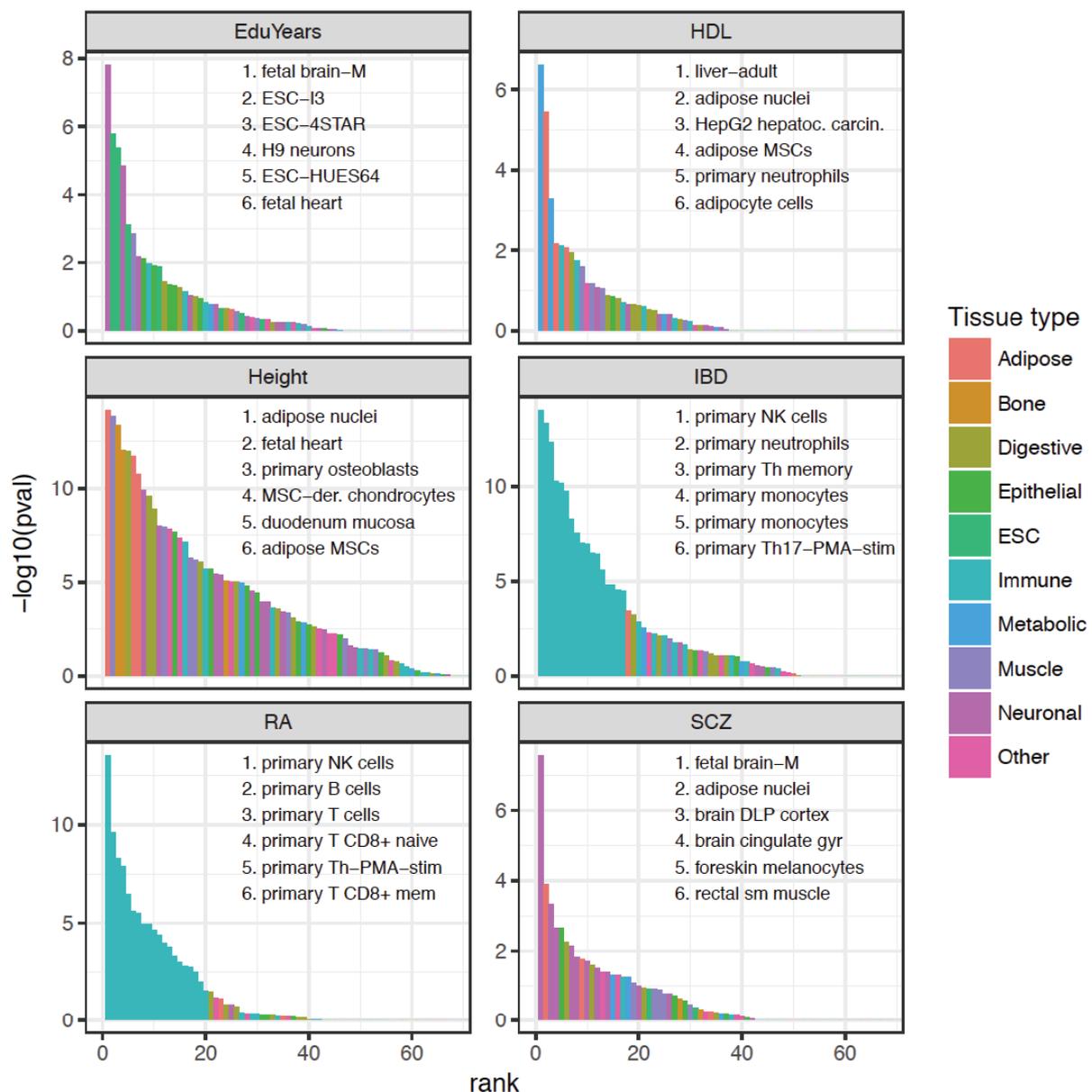
Figure 5: Enriched epigenomes identified by robust rank aggregation with maxPRF score for six GWAS.

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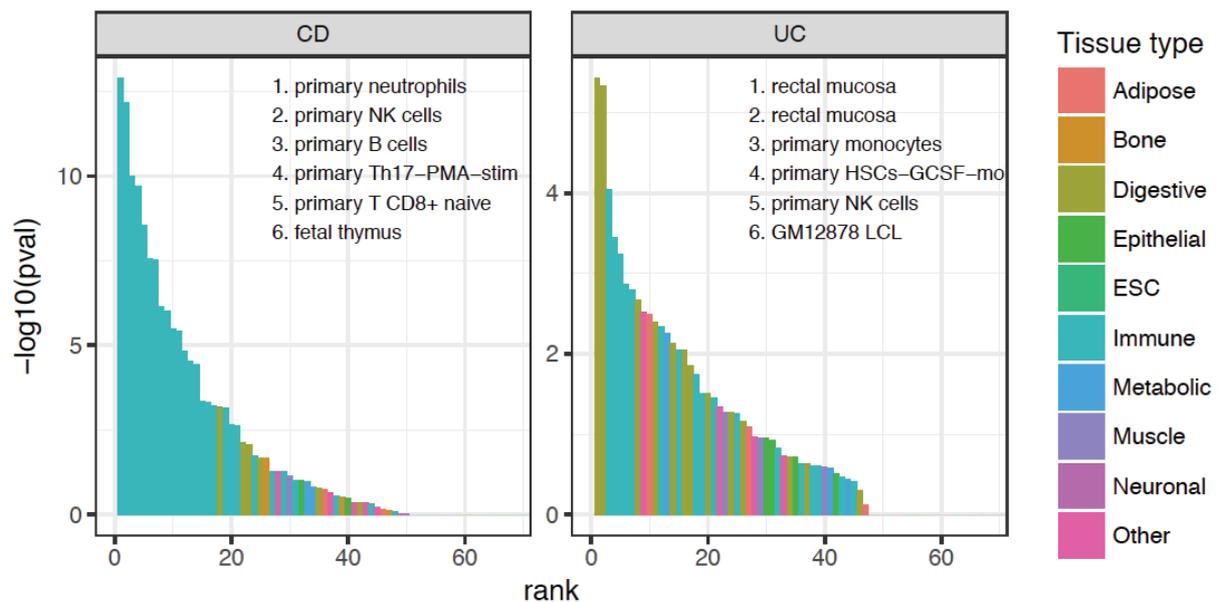
Figure 6: Enriched epigenomes in GWAS of ulcerative colitis and Crohn's disease.

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.
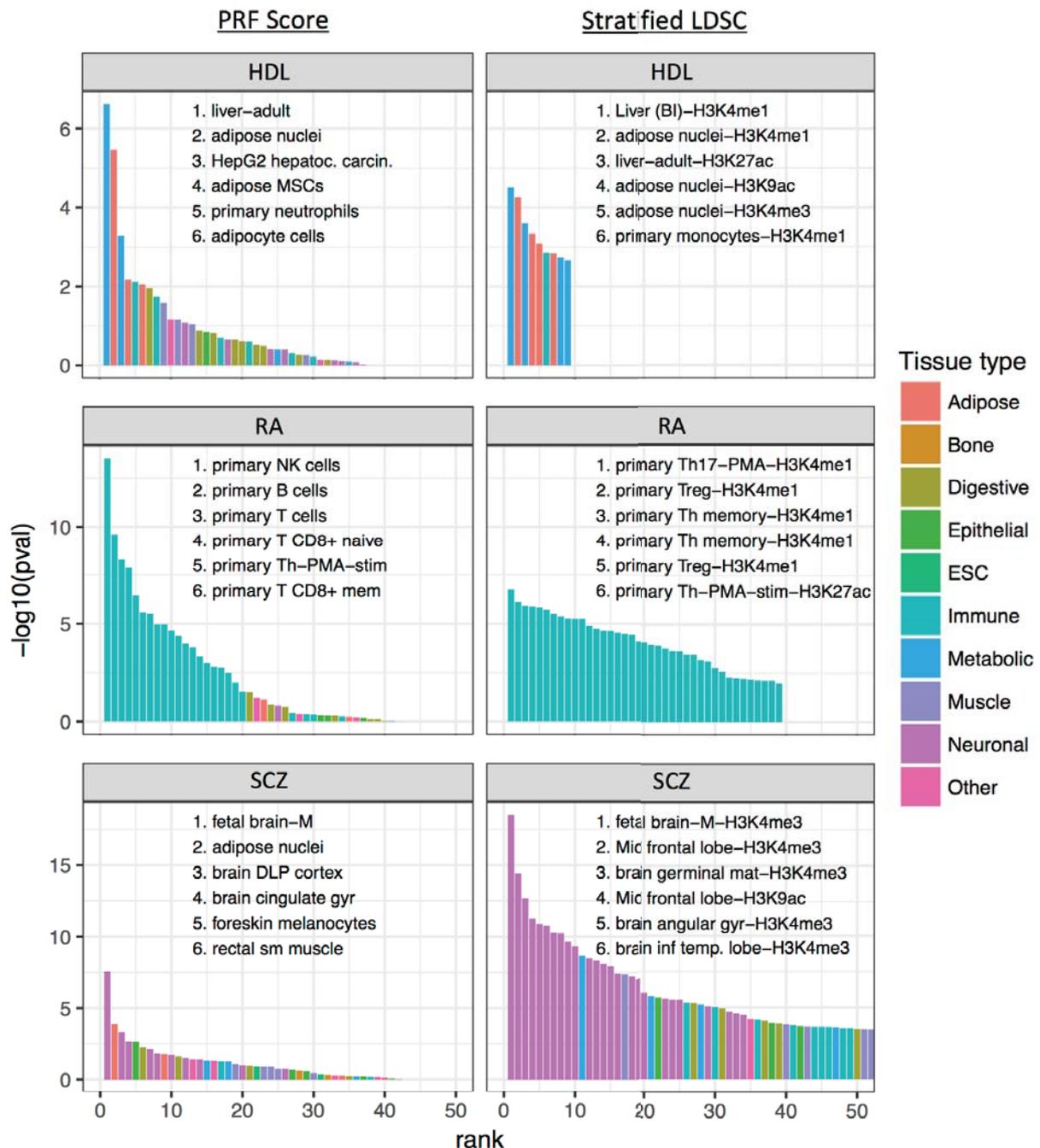
Figure 7: Comparison of the cell type enrichments discovered using PRF scores or stratified LD score regression (LDSC), across three GWAS traits: HDL cholesterol, rheumatoid arthritis, and schizophrenia. The log p value axes differ between traits, but are the same in each case for PRF scores and stratified 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; $p=2\times10^{-21}$, 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.
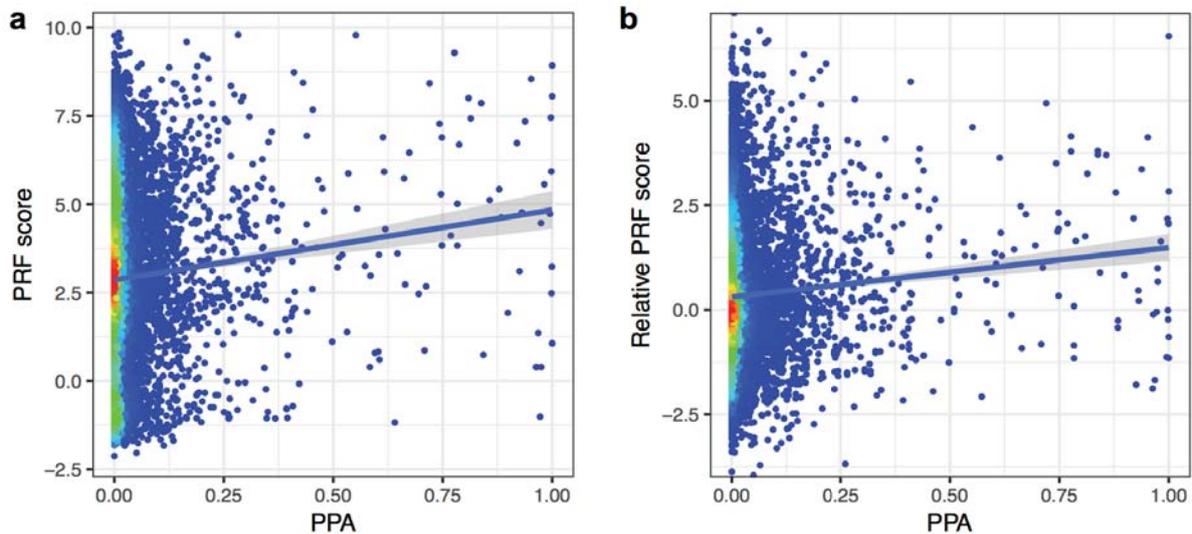

Figure 7: Scatter plots with linear fits of the relationship between PRF scores and the posterior probability of association, across all loci of five GWAS. The shaded region represents the 95% confidence interval for the fit. For plotting purposes, points with PPA < 0.1 were downsampled. (**a**) PRF 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.
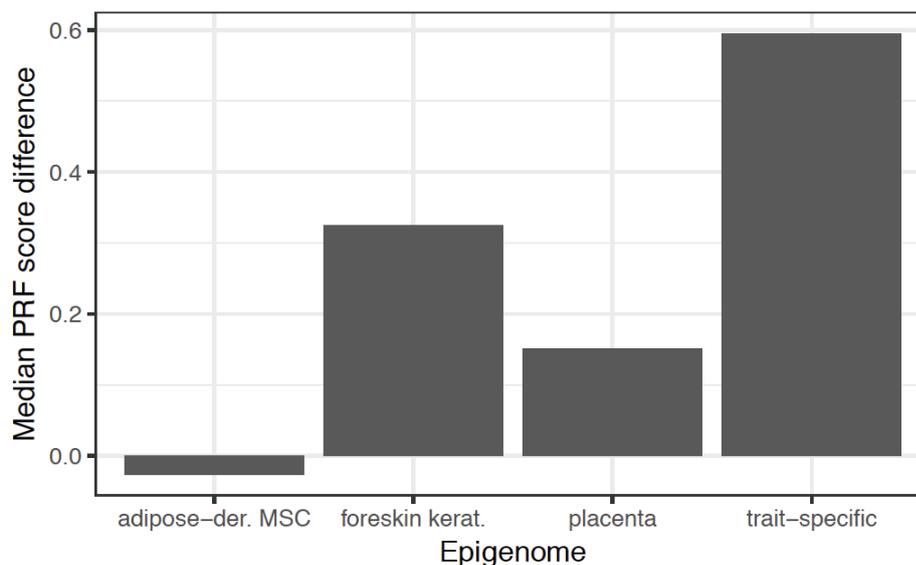
Figure 8: Bar plot of the median difference between PRF scores for variants with statistical PPA > 0.5 and the top 20 variants by statistical association at each locus, computed across five GWAS either for trait-relevant epigenomes (noted in Table 1) or for 3 trait-irrelevant epigenomes.

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).
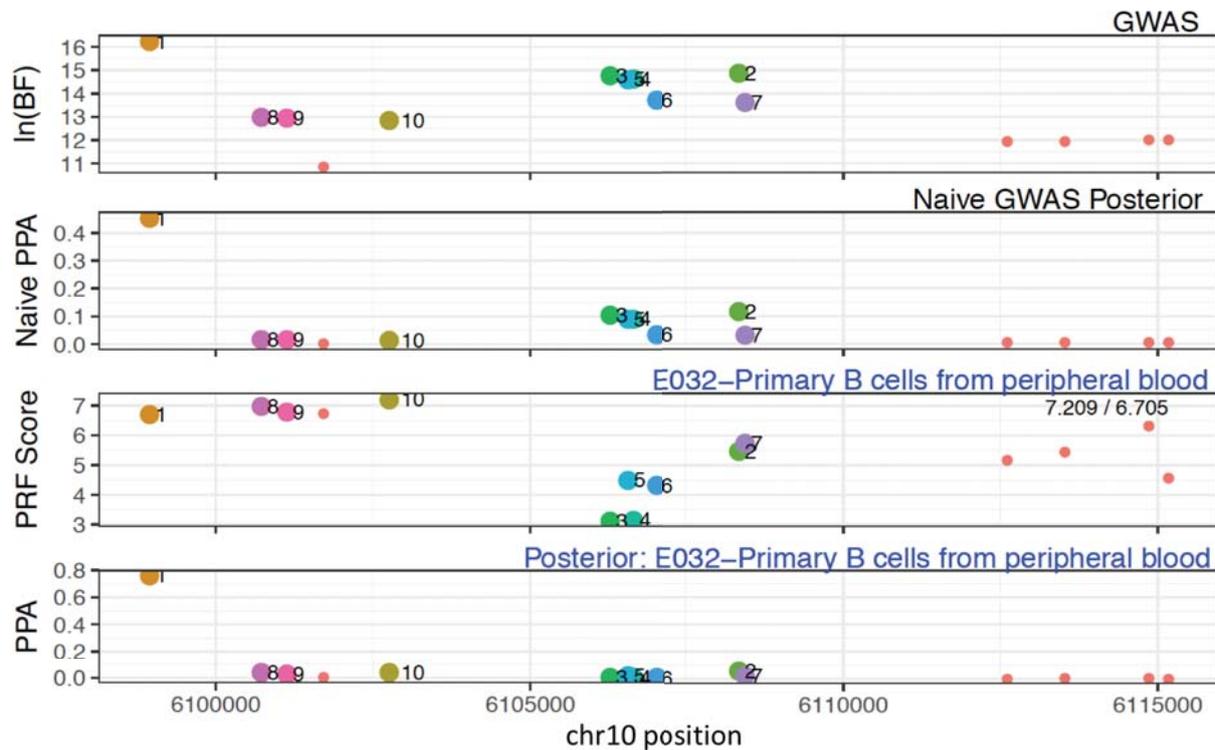
Figure 9: PRF score fine-mapping plot of a rheumatoid arthritis GWAS association at the *IL2RA* locus using the epigenome E032-Primary B cells from peripheral blood. Variants with naive PPA > 0.001 are shown; variants are numbered according to their statistical association, with number 1 being the most 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.
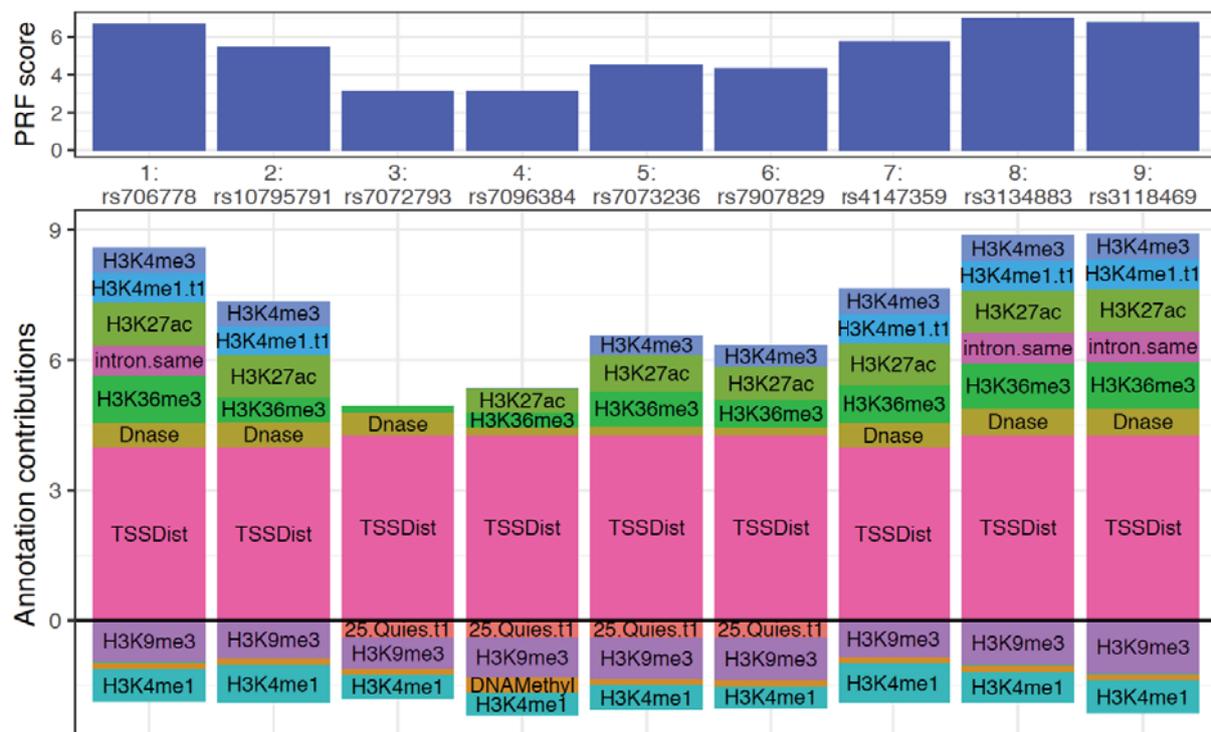
Figure 10: PRF score bar plot showing annotation contributions to the score for each variant in the epigenome E032-Primary B cells from peripheral blood. Annotation labels above zero contribute positively to the PRF score, whereas those below zero subtract from the PRF score. Many of these enrichments 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.
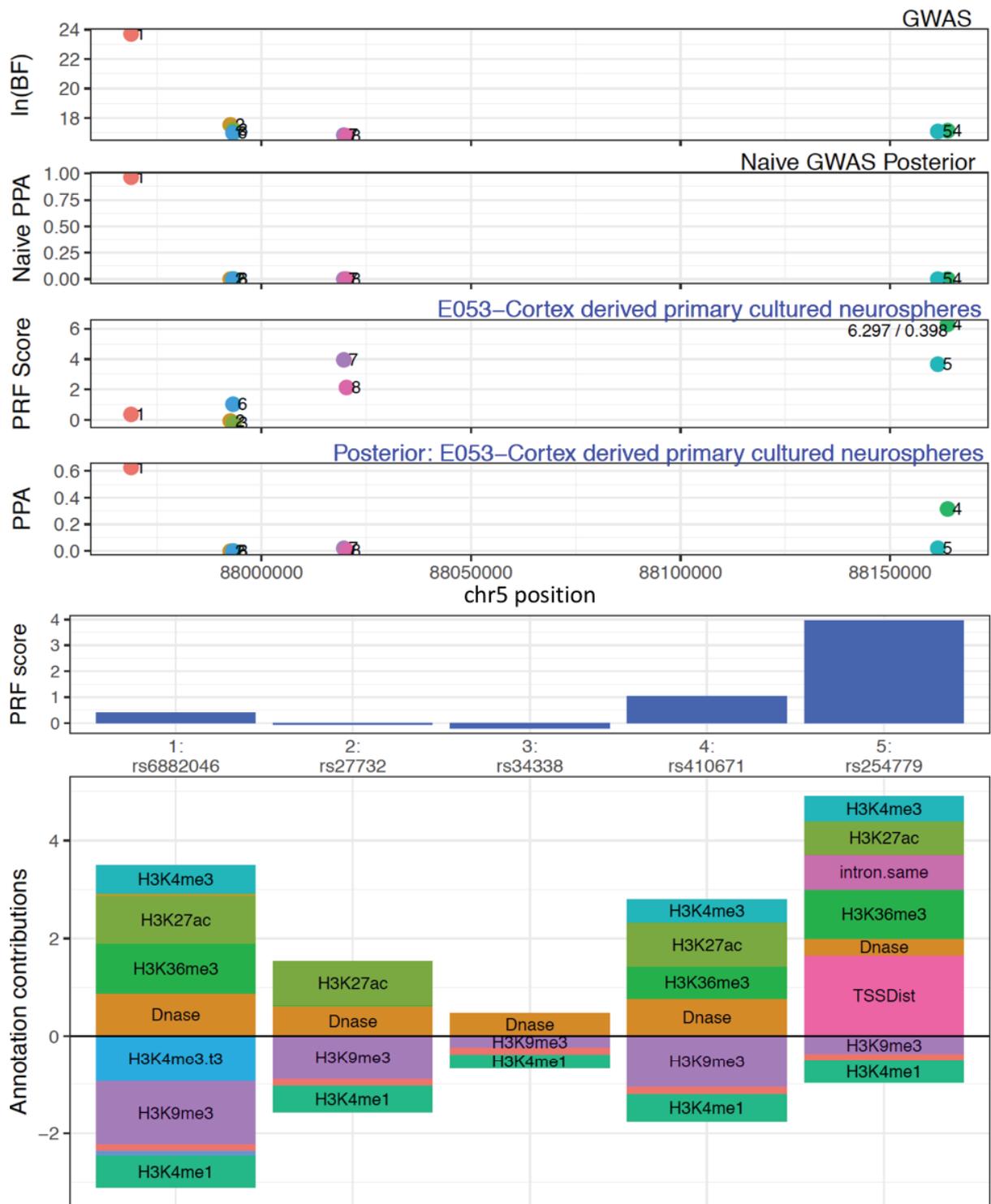
Figure 12: PRF score fine-mapping of an educational attainment association at the *MEF2C* locus using the epigenome E053−Cortex derived primary cultured neurospheres.

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*.
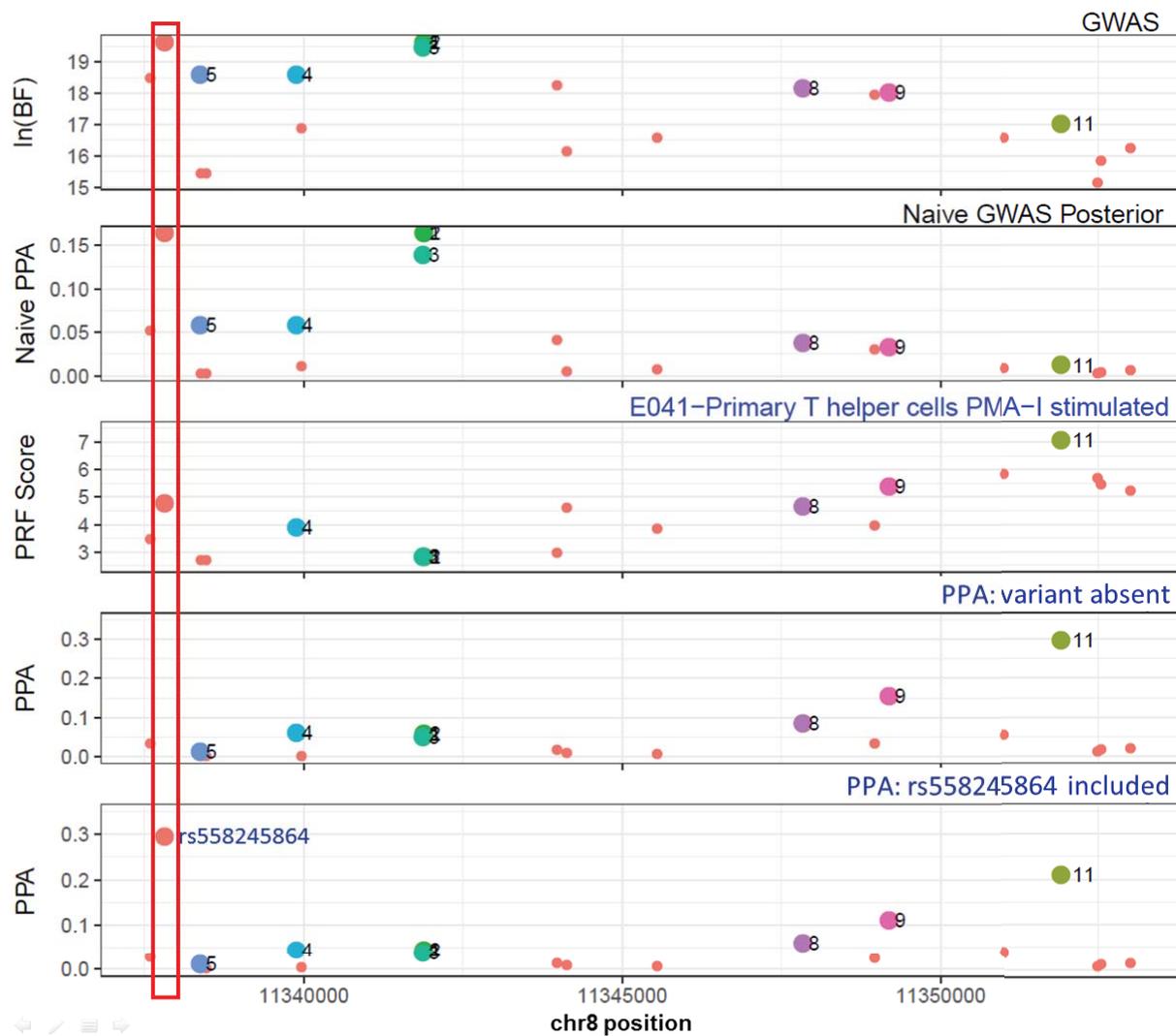
Figure 13: PRF score fine-mapping of a rheumatoid arthritis association at the *BLK* locus using the epigenome E041-PMA-I-stimulated T helper cells. The position of rs558245864 is outlined in red. The bottom panel shows functional PPAs computed when rs558245864 is included, whereas the second-to-bottom 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 1×10^{-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).
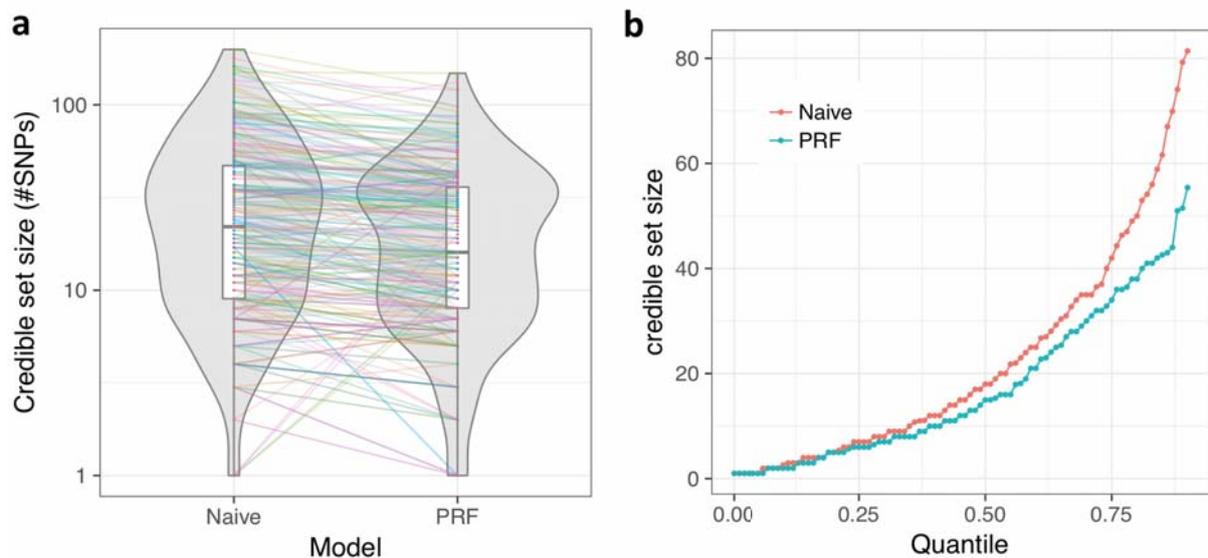
Figure 14: Credible set sizes are reduced when applying PRF score fine-mapping. (**a**) Violin plot of 95% credible set sizes across all loci either using statistical fine-mapping ("naive") or PRF score fine-mapping. Lines show the change in credible set size for each associated locus, with different 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).
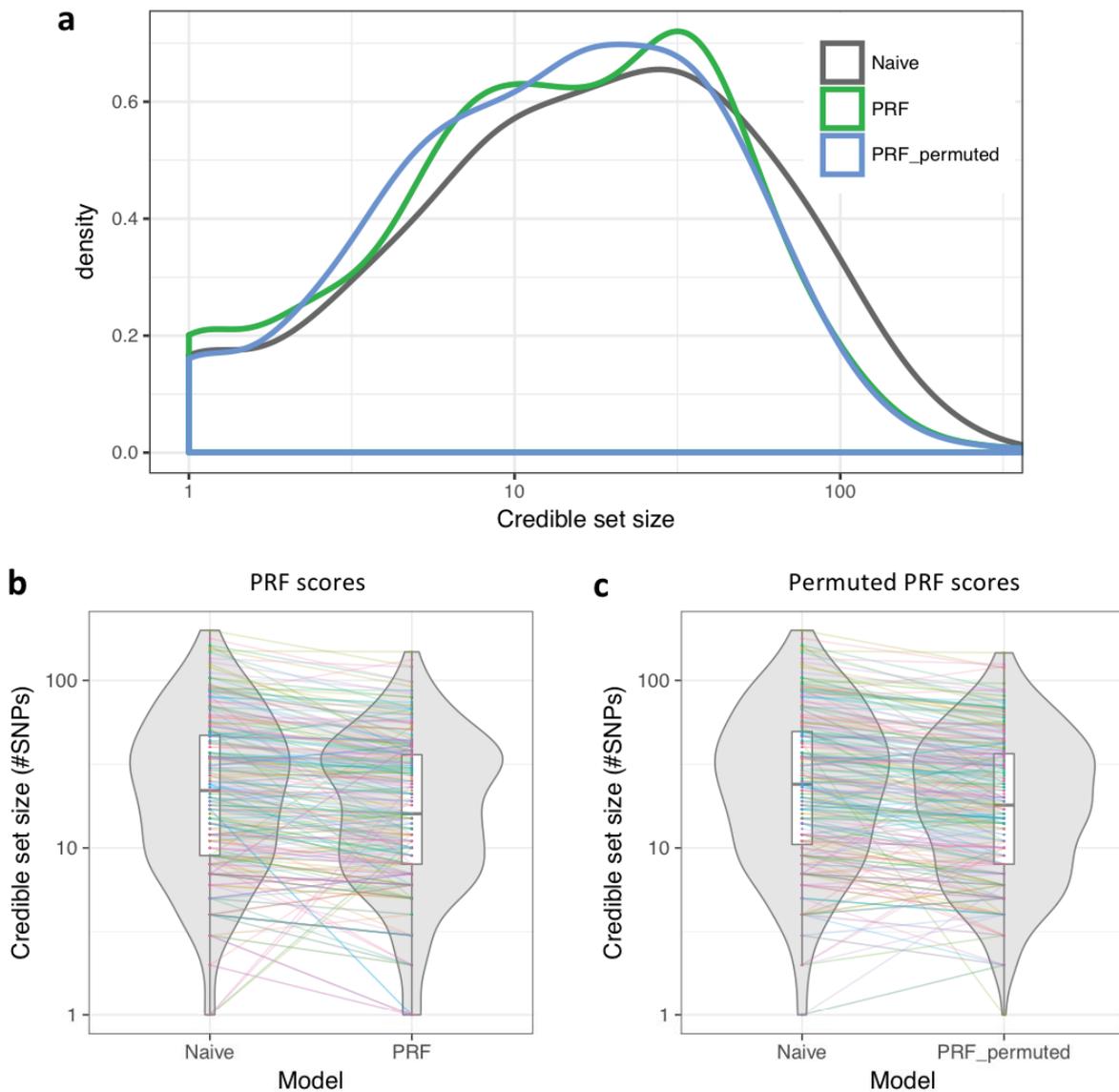
Figure 15: (**a**) Density plot of credible set sizes from fine-mapping 317 trait-associated loci using either true PRF scores or PRF scores permuted among variants with naive PPA > 0.001 at each locus, compared with credible sets from naive statistical fine-mapping. Using both true scores and permuted 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.
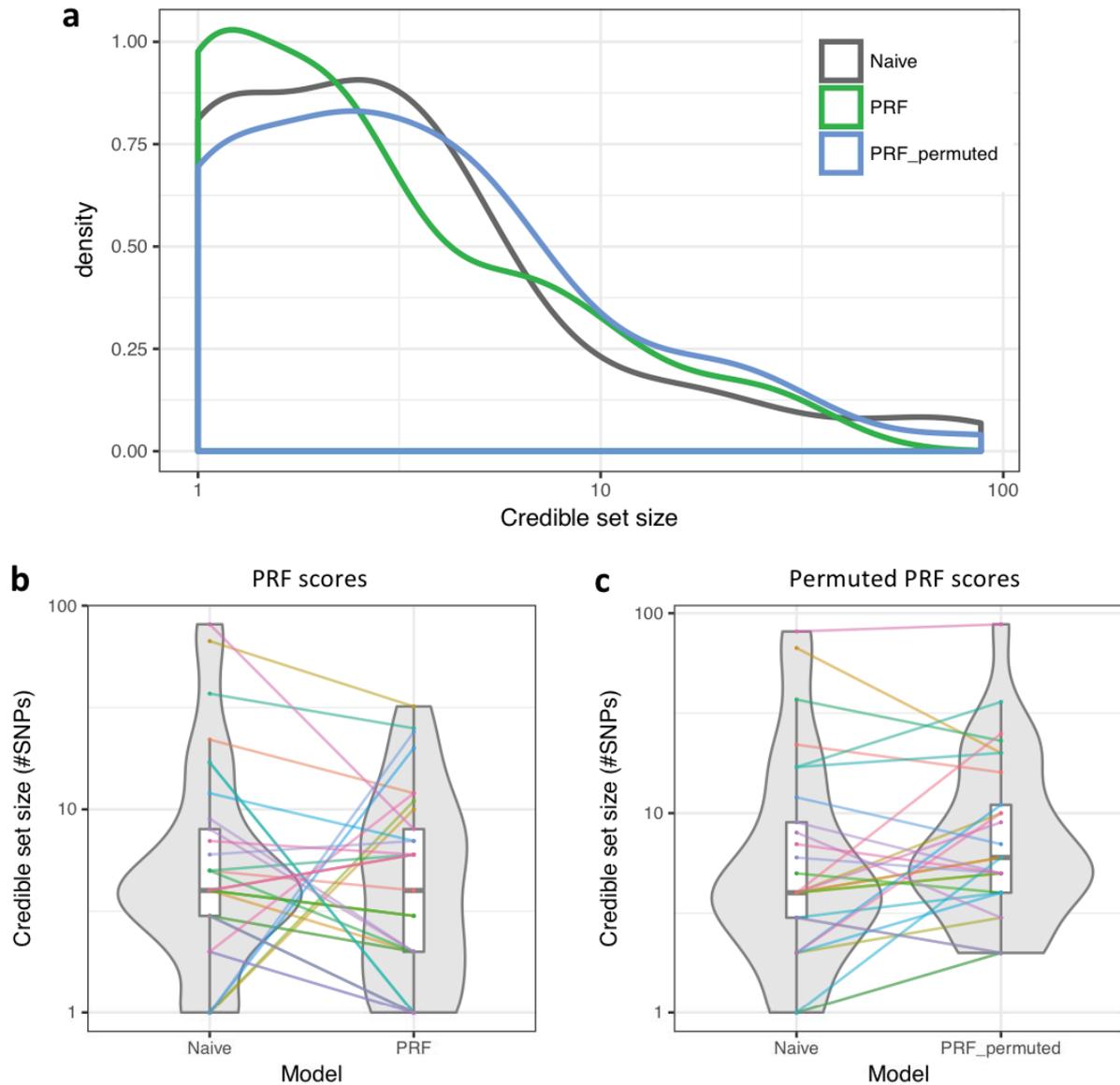

Figure 16: Fine-mapping of 62 loci with a lead variant having naïve PPA > 0.5, using either true or permuted PRF scores. (**a**) Density plot of credible set sizes from fine-mapping. True PRF scores give a slightly higher density of small credible sets, while permuted PRF scores have a slightly higher 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.
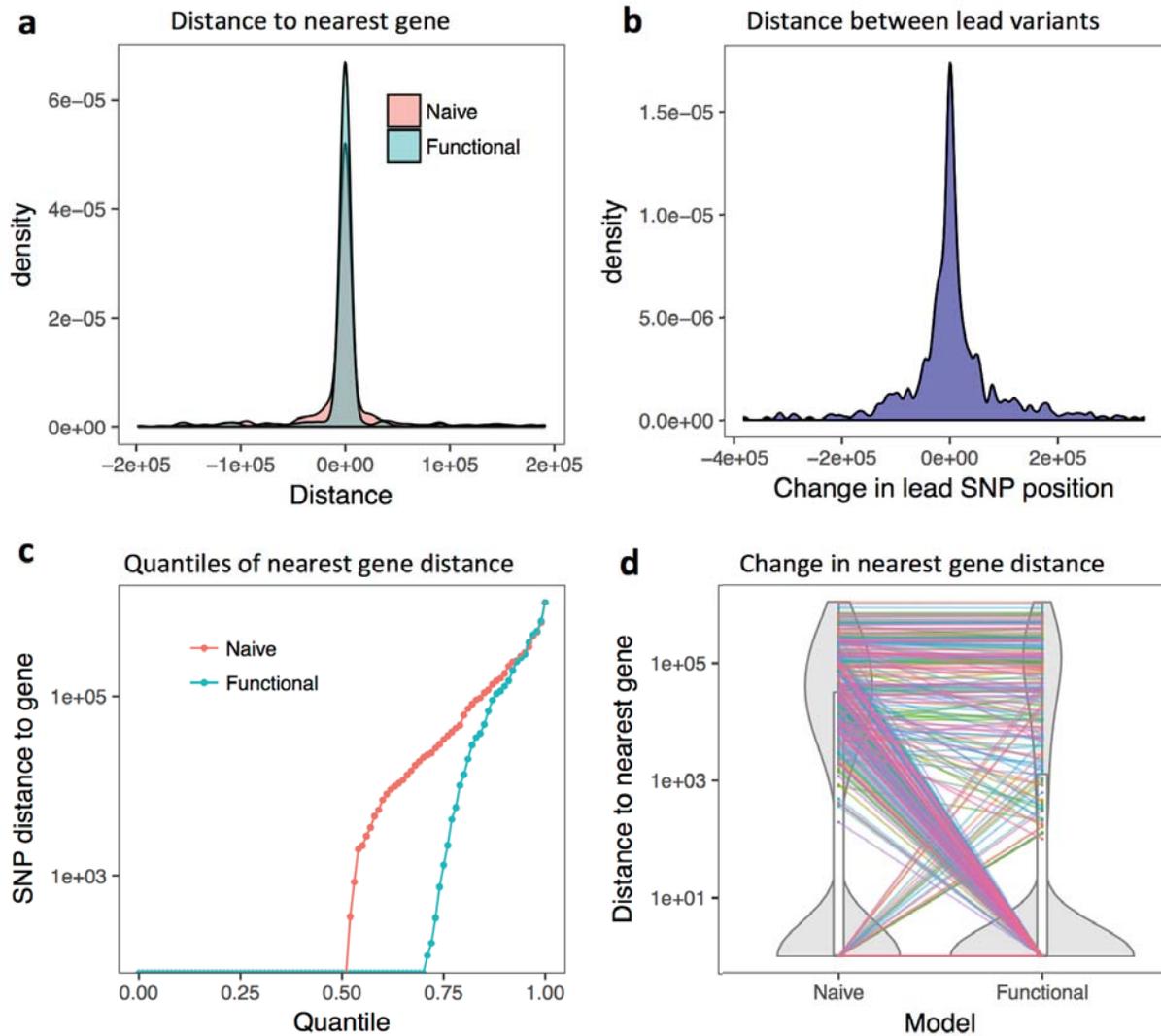
Figure 17: How implicated genes are changed by PRF score fine-mapping for 1,002 GWAS associations across six traits. (**a**) Density plot of distance to the nearest gene for lead variants either pre- or post-fine-mapping. (**b**) Density plot of the distance from the functional lead variant to the naive 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).
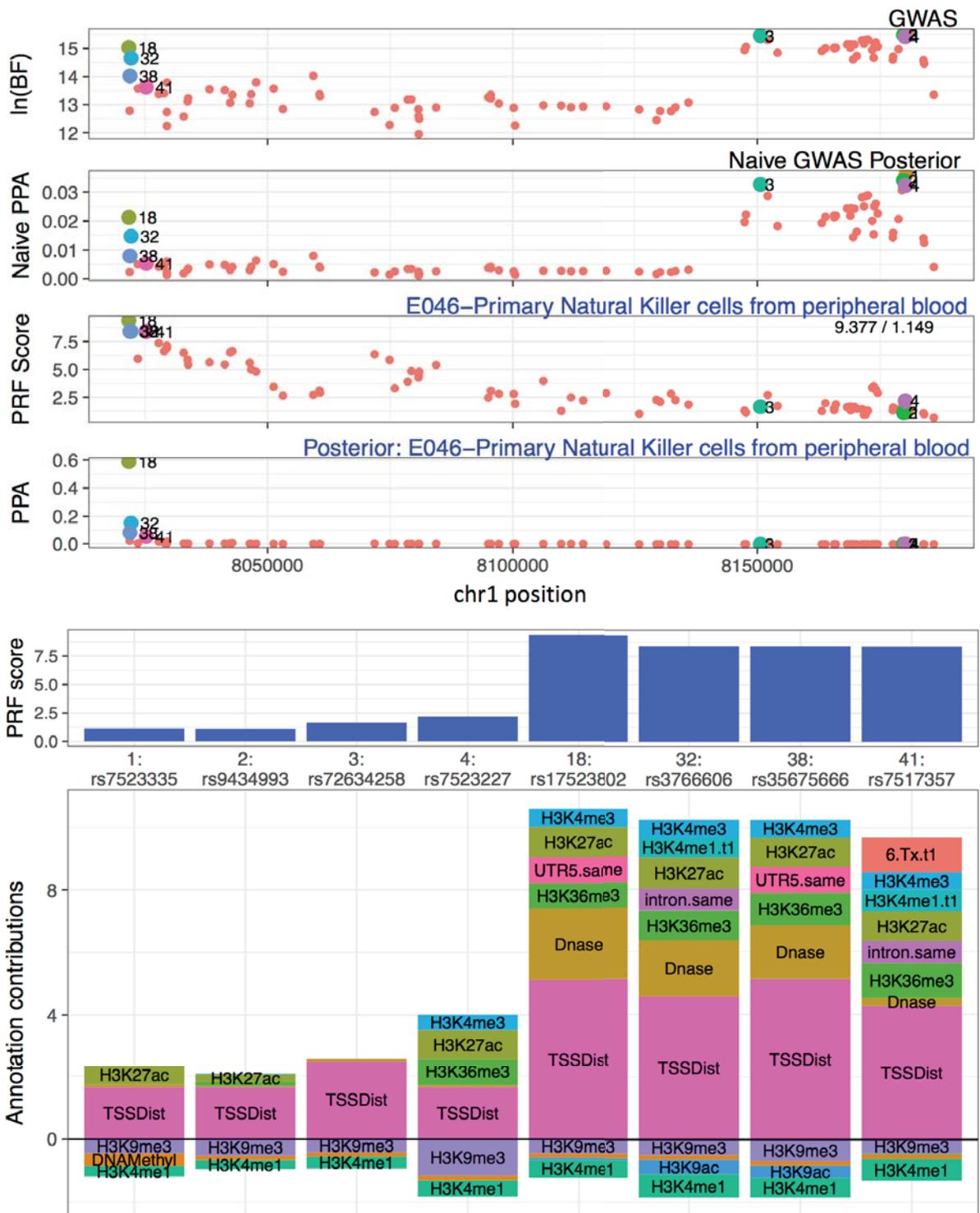
Figure 18: Fine-mapping an IBD association at the *PARK7-ERFFI1* locus localises a candidate causal variant at the promoter of *PARK7*.

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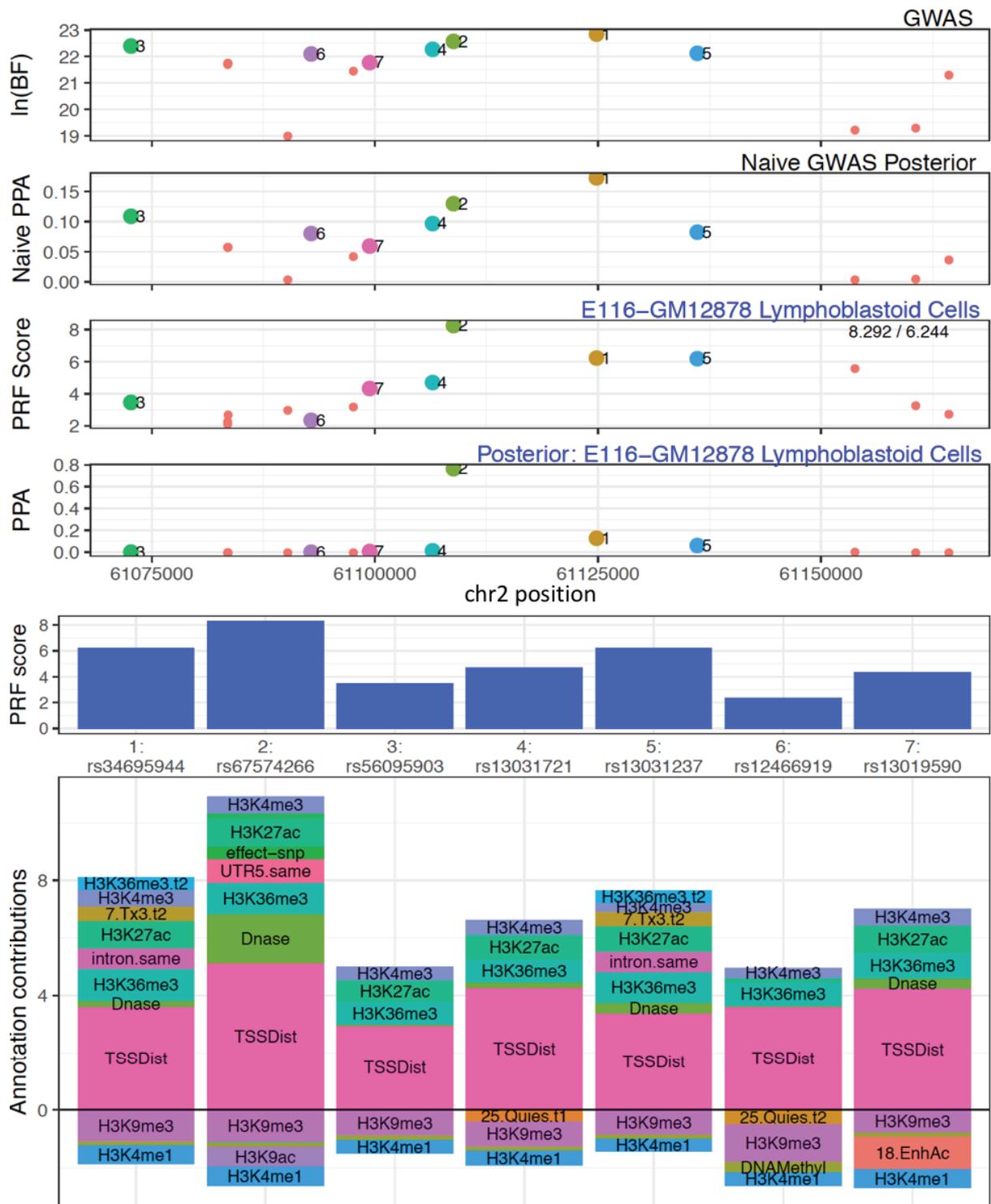
Figure A1: PRF score fine-mapping plots at the RA-associated *REL* locus.

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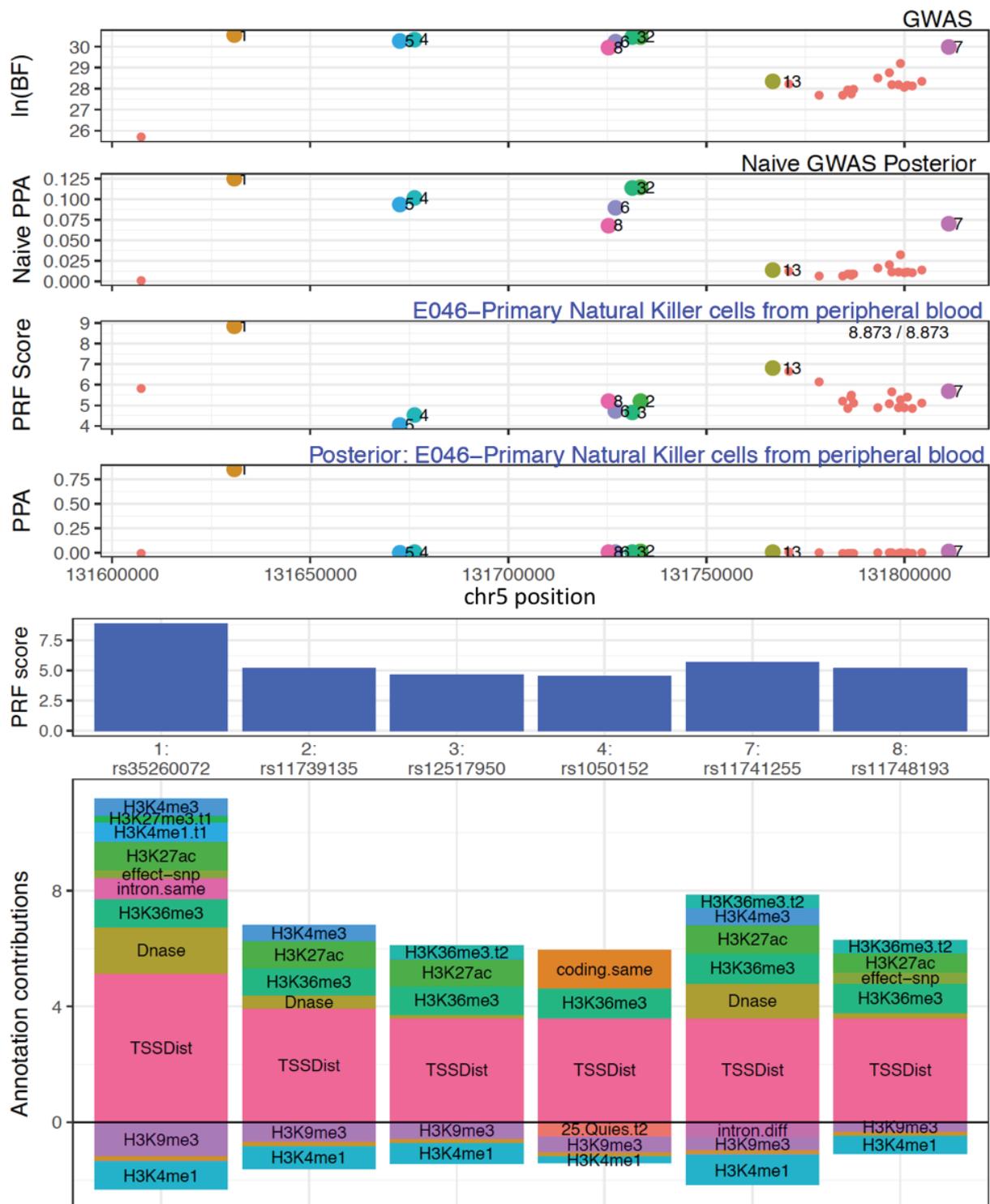
Figure A3: PRF score fine-mapping plot at the IBD-associated *SLC22A4* locus.